Bindu Verma

KeshuShuklaDtuThesisPlagCheck_updated

bindu

Document Details

Submission ID
trn:oid:::27535:140806921

Submission Date
May 28, 2026, 6:57 PM GMT+5:30

Download Date
May 28, 2026, 7:12 PM GMT+5:30

File Name
KeshuShuklaDtuThesisPlagCheck_updated.pdf

File Size
1.8 MB

66 Pages

22,034 Words

119,689 Characters

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 8 words)

Match Groups

70 Not Cited or Quoted 4%
Matches with neither in-text citation nor quotation marks

4 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
1% Publications
3% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **70** Not Cited or Quoted 4%
  Matches with neither in-text citation nor quotation marks
- **4** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
1% Publications
3% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	
Delhi Technological University on 2026-05-15		2%
2	Internet	
github.com		<1%
3	Internet	
arxiv.org		<1%
4	Internet	
www.coursehero.com		<1%
5	Student papers	
Central University Of Jharkhand on 2026-04-22		<1%
6	Internet	
proceedings.iclr.cc		<1%
7	Student papers	
National Taiwan University on 2023-06-16		<1%
8	Internet	
vdoc.pub		<1%
9	Publication	
Han, Fangfang, Guopeng Zhang, Huafeng Wang, Bowen Song, Hongbing Lu, Dazh...		<1%
10	Publication	
Shan Lu, Guixia Kang, Qiqu Zhu, Ping Zhang. "A Orthogonal Superimposed Pilot f...		<1%

11 Student papers
Imperial College of Science, Technology and Medicine on 2019-09-12 <1%

12 Internet
ebin.pub <1%

13 Student papers
University of Petroleum and Energy Studies on 2026-04-26 <1%

14 Student papers
University College London on 2018-09-07 <1%

15 Student papers
University of Adelaide on 2026-03-16 <1%

16 Internet
1library.net <1%

17 Student papers
University of Newcastle on 2021-11-20 <1%

18 Student papers
University of Sydney on 2021-06-23 <1%

19 Student papers
LNM Institute of Information Technology on 2026-04-21 <1%

20 Student papers
University of Liverpool on 2025-05-19 <1%

21 Internet
docslib.org <1%

22 Publication
Carlos, L.. "Intermediate distributions and primary yields of phenolic products in ... <1%

23 Student papers
Jacobs University, Bremen on 2015-10-08 <1%

24 Publication
Yue, Zhixiong. "Learning Architecture for Multiple Tasks in Transfer Learning", U... <1%

25 Internet
deepai.org <1%

26 Internet
wrap.warwick.ac.uk <1%

27 Internet
www.researchgate.net <1%

28 Student papers
Aalto Yliopisto on 2026-05-18 <1%

29 Publication
Jin Liu, Jianxin Wang, Yi Pan. "AI in MRI-based Brain Disease Prediction", CRC Pres... <1%

30 Student papers
Universiteit van Amsterdam on 2018-09-04 <1%

31 Internet
dr.ntu.edu.sg <1%

32 Internet
etheses.whiterose.ac.uk <1%

33 Internet
residential.skanska.cz <1%

34 Internet
stacks.stanford.edu <1%

35 Internet
www.mdpi.com <1%

31

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Take four radiologists reading the same CT slice of a right lower lobe pulmonary nodule. Three draw tight contours around the solid core. The fourth traces a wider boundary to include the surrounding ground-glass haze. Overlaid on the same image, the four delineations do not cluster around a common answer with minor scatter — they represent four distinct clinical judgements about where the lesion ends, and the imaging data does not settle the question.

At a lung nodule's margin, tissue density falls off gradually rather than sharply, ground-glass opacity has no edge a ruler could trace, and the convention for where the nodule stops differs between practitioners and between institutions. These are not correctable errors. Joskowicz et al. [6] reviewed CT annotation studies across anatomical structures and found rater disagreement large enough to change downstream clinical management, and for lung nodules in particular the disagreement is systematic enough that label variability overtakes image noise and class imbalance as the primary source of training uncertainty.

A segmentation model trained on one radiologist's masks learns that radiologist's boundary convention and nothing else. A model trained on the pixel-wise average of four masks chases a boundary that no radiologist actually drew — when two annotators went tight and two went wide, the average is a blurred intermediate that calibrates the model to a compromise opinion nobody holds. Both choices discard what a clinician in a second-opinion setting most needs: not a single boundary location, but a representation of how much expert judgment divides over this case and in which direction each expert resolved the ambiguity.

What the setting actually requires is a model whose output at test time is a distribution over plausible segmentations, calibrated to what the radiologist population produces. On a nodule with a clean, isolated, solid boundary where all four annotators converge, the model's samples should cluster. On a nodule with ill-defined margins, a part-solid texture, and four substantially different contours from four qualified radiologists, the samples should span that range. A single-output model has one forward

pass and one mask per input: the architecture gives the clinician reviewing the case a prediction, not a characterisation of how much the radiologist population divides on that particular boundary.

1.1.1 The LIDC-IDRI Dataset and the GED Metric

The primary dataset is LIDC-IDRI [1]: 1,018 CT scans, each with four independent radiologist annotations of pulmonary nodules, yielding 1,609 nodule patches after preprocessing (described in full in §2.6). Armato et al. [2] documented that in this dataset, which annotator's mask is treated as ground truth shapes model calibration more than image noise or class imbalance does — boundary extent differences between radiologists were that large. The difference between single-output models and probabilistic ones shows up in measured GED and $\text{Dice}_{\text{match}}$, and the clinical stakes of miscalibrated uncertainty in a nodule screening setting are well-documented.

The Generalised Energy Distance [11] captures both failure modes in a single scalar. When a model collapses to the same prediction on every forward pass, its within-sample distance term goes to zero regardless of how accurate that one prediction is — GED penalises this heavily. A model that spreads predictions widely but misses every annotator's mask also scores poorly. Throughout this thesis, GED is the primary metric and lower values are better. $\text{Dice}_{\text{match}}$ is the secondary metric: it uses optimal Hungarian assignment to match model samples to individual rater annotations and measures geometric closeness to each annotator separately, so diversity and per-rater accuracy can be tracked independently.

1.1.2 Probabilistic Segmentation Models

Probabilistic segmentation models address the distribution-output requirement by replacing the single-mask prediction with a stochastic latent variable that encodes annotation-level uncertainty. The Probabilistic U-Net (ProbUNet) [11], which combined a conditional variational autoencoder (VAE) [9] with the U-Net backbone [14], was the first model to demonstrate genuinely diverse segmentation hypotheses on LIDC-IDRI. Subsequent work extended this line in different directions: PHiSeg [3], which addressed coarse-versus-fine scale disagreement by maintaining separate latent variables at each spatial resolution, and Stochastic Segmentation Networks [12], which bypassed the latent variable altogether and parameterised uncertainty as a distribution over the full output mask. Each addressed different aspect of how annotation uncertainty should be represented. The qualitative result across all three is the same: diverse, plausible samples can be drawn at test time by sampling from a learned prior over the latent space.

In the multi-rater setting, however, training a single posterior encoder on all annotators' masks simultaneously introduces a problem that persists regardless of what

architectural choices are made above the posterior level. The problem originates in the backward pass: conflicting rater gradients accumulate at the shared latent code and the encoder's weights shift toward a compromise that satisfies none of the individual rater signals cleanly.

1.2 The Gradient Conflict Problem

In the shared posterior setup, a single encoder takes the image and all four annotators' masks as a concatenated 5-channel tensor and outputs a distribution over a shared latent code $\mathbf{z}$. A decoder maps samples from $\mathbf{z}$ to segmentation masks. At test time only the prior runs — the posterior encoder plays no role in inference — so the diversity of prior samples at test time is what GED ends up measuring.

Training this encoder requires a reconstruction loss for each annotator (a term that pushes the decoder's output toward each radiologist's mask). The gradient of each such term flows back through the decoder and accumulates at the shared latent code $\mathbf{z}$. When four radiologists annotated the same nodule with four different boundary decisions, their four reconstruction gradients push $\mathbf{z}$ in four different directions simultaneously. The shared encoder's weights are updated according to the sum of these conflicting signals.

Take radiologist 1, who drew a tight contour around the solid core, and radiologist 4, who extended the boundary outward into the surrounding opacity. The gradient from radiologist 1's reconstruction loss pushes $\mathbf{z}$ toward a code from which a tight prediction decodes. The gradient from radiologist 4's loss pushes $\mathbf{z}$ the other way. Both gradients hit the shared encoder at same backward pass, their effects partially cancel, and the encoder's weights shift toward a compromise that fits neither boundary decision. When two annotators label completely non-overlapping regions, the cancellation is exact and the net gradient at $\mathbf{z}$ is zero.

This problem is structurally equivalent to gradient conflict in multi-task learning. Yu et al. [19] showed this conflict can be measured through the cosine similarity between per-task gradients at the shared encoder, which quantifies how much those tasks compete rather than reinforce each other, and the resulting per-task performance loss tracks that cosine similarity. A negative cosine similarity means two tasks' gradients actively pull encoder weights in opposite directions, and the net update that the encoder receives is a blurred compromise that serves neither task well. Training a shared posterior on N annotators creates the same condition, with each annotator's reconstruction loss acting as a separate task competing over the shared encoder, and the net gradient encoding none of the competing boundary decisions cleanly.

1.2.1 D-Persona and the Limits of Stage 2 Personalisation

34

D-Persona [17] is the current state of the art in multi-rater segmentation and the primary baseline for this thesis. Its response to the shared posterior limitation is a two-stage training pipeline. Stage 1 trains the shared posterior encoder with all annotators' masks concatenated, producing a shared latent code $\mathbf{z}$. Stage 2 introduces per-rater projection heads that steer this shared code toward individual annotator styles at inference time. On LIDC-IDRI, the full two-stage pipeline achieves GED 0.1507 and $\text{Dice}_{\text{match}}$ 0.8909, a substantial improvement over ProbUNet (GED 0.2234, $\text{Dice}_{\text{match}}$ 0.8836).

Stage 2 was built on the assumption that a shared latent code contains enough per-rater structure for a small projection head to extract and amplify. A head applied to $\mathbf{z}$ can redistribute whatever information is already there — it can rotate dimensions, amplify or suppress components, and bias the output toward given annotator's boundary style. The problem is that Stage 1's gradient averaging removes per-rater structure from $\mathbf{z}$ before Stage 2 ever sees it. Each annotator's gradient signal gets partially cancelled by the others during Stage 1's backward pass, and Stage 2's heads have nothing rater-specific to work with.

The ablation study in Chapter 4 makes this concrete. Adding Stage 2 projection heads on top of a Stage 1 trained with per-rater posteriors (which already provides clean, rater-specific gradients) raises GED by 27%, from 0.1444 to 0.1836. When Stage 2 is applied on top of a Stage 1 that already provides clean, rater-specific gradients, it introduces shared-posterior dynamics on a model that does not need them. The projection heads find no rater-specific structure to amplify and instead collapse the diversity Stage 1 created.

1.2.2 The Architectural Capacity Hypothesis

Before arriving at the per-rater formulation, a simpler hypothesis should be ruled out: if the shared posterior's bottleneck is representational capacity rather than training objective, a more expressive encoder should help regardless of how the objective is structured. Transformer-based architectures, specifically the Mix Transformer (MiT-B2) from Xie et al. [18], represent exactly this class of increased capacity. MiT-B2's attention mechanism is not limited by local kernel size; it can, in principle, relate any two spatial locations regardless of distance. If annotation-ambiguous regions produce interacting gradient signals that a convolutional encoder cannot disentangle because of its limited receptive field, MiT-B2 should show improvement over ResNet34 in shared posterior setting. Substituting MiT-B2 into the shared posterior encoder tests this: if the problem is architectural expressiveness, a stronger encoder should close the GED gap.

The experiment is described in full in §4.2.3 (Table 4.4, Row 2). MiT-B2 yields GED 0.1531 versus the ResNet34 baseline's 0.1507, with no change in $\text{Dice}_{\text{match}}$. A more expressive encoder learns a more expressive average of conflicting

rater gradients; the conflict itself is unaffected by how large the encoder is. Increasing encoder expressiveness changes nothing about gradient conflict, because the conflict is a property of the training objective (specifically which gradient signals reach the encoder) and not of the encoder's capacity.

1.2.3 Per-Rater Posterior Encoders: The Proposed Solution

Replacing the shared posterior with N independent encoders, each dedicated to one annotator, removes the conflict where it originates. Each encoder receives a 2-channel input: the image and that annotator's mask only. Its gradient signal reflects only that annotator's boundary decisions, with no contribution from any other annotator's loss. The shared decoder and the shared prior receive N rater-specific signals rather than one conflict-averaged signal.

The training objective decomposes into N per-rater ELBO terms, one per annotator. Since each ELBO term depends only on its own encoder's output, $\partial \mathscr{L}_i / \partial \mathbf{z}_j = 0$ for $i \neq j$ — the gradient from encoder i carries no information from encoder j, and no regularisation is needed to enforce this. It follows from the objective structure alone.

At test time, the N posterior encoders are dropped. Inference is identical to the D-Persona baseline: sample $\mathbf{z}$ from the shared prior, decode. The training cost is approximately $4.5\times$ higher; inference cost is unchanged. Every performance difference in Chapter 4 originates from training-time gradient quality, since nothing else differs between the two models.

On LIDC-IDRI, this Stage 1-only per-rater model achieves GED 0.1444 ± 0.0141 (-4.2%) and $\text{Dice}_{\text{match}}$ 0.9112 ± 0.0061 ($+2.28\%$) compared to the full D-Persona two-stage pipeline, with per-rater Dice improving for each individual expert. Cross-dataset validation on NPC-170 (nasopharyngeal carcinoma MRI, 4 annotators, 3-seed evaluation) produces a GED gap of 0.0011, which falls within the seed variance of ± 0.0085, indicating neither improvement nor degradation.

1.3 The Annotation Sparsity Extension

LIDC-IDRI guarantees four annotations per case by design. Most clinical imaging archives do not. A retrospective dataset assembled from routine reads may have three annotators on some scans, one on others, and the coverage is uneven in ways that reflect staffing history rather than data quality decisions. The full-annotation results matter, but the more operationally relevant test is how each model degrades when annotation coverage is patchy.

1.3.1 Sparse Annotation in the Shared Posterior

In the shared posterior model, an absent annotator's mask channel is set to zero. The encoder still receives a full 5-channel input; the absent annotator's channel contains only zeros. Two things follow from this. First, the zero-mask channel produces a reconstruction gradient that pushes the latent code toward predicting an empty segmentation for the absent annotator (a gradient that competes with the present annotators' reconstruction gradients). Second, as more annotators become absent, the proportion of zero-mask gradient signals grows. When only one of four annotators is present, the shared encoder receives three zero-mask gradients and one real annotation gradient. The latent code is simultaneously pulled toward the present annotator's boundary decision and toward predicting nothing for three absent annotators. With three absent annotators versus one present, the zero-mask gradients outnumber the real annotation gradient three to one, and the net update at $\mathbf{z}$ is pulled predominantly toward predicting empty masks.

1.3.2 Sparse Annotation in Per-Rater Posteriors

In the per-rater model, absent annotators are handled through omission: if annotators 1, 3, and 4 are absent for a given image, their three encoders simply do not execute. The gradient update for that training step comes only from encoder 2, carrying only encoder 2's rater-specific signal. Nothing in the architecture forces a zero-channel placeholder into the gradient computation; the absent raters leave no trace in the parameter update at all. Chapter 4's sparsity results follow directly from that difference in how absence is represented.

1.3.3 Experimental Evidence

At one annotator per training image, the per-rater model achieves $+21.4\%$ GED over the shared baseline. At two annotators, $+17.8\%$. At three, $+11.5\%$. The gap widens at every step down in coverage, and across all 12 per-fold comparisons (three sparsity levels times four folds), the per-rater model wins every single one. Under a null hypothesis of no systematic difference, $(1/2)^{12} \approx 0.024\%$ is the probability of that sweep.

A gradient alignment diagnostic provides a mechanistic picture of what is happening inside the shared baseline as annotation coverage decreases. The mean pairwise cosine similarity of per-rater reconstruction gradients at the shared latent code rises from 0.167 at full annotation to 0.976 at one annotator present (a near-complete collapse to a single gradient direction, as three zero-mask signals drive all gradients toward the same degenerate prediction of an empty mask). The within-fold spread of this alignment measure collapses approximately 19-fold (standard deviation from 0.439

at full annotation to 0.023 at one annotator present). At maximum sparsity, the collapse is nearly universal across test cases, not limited to specific high-disagreement nodules. The per-rater model's gradient alignment stays at 0.000 at all sparsity levels, since the architecture has no shared $\mathbf{z}$ to measure alignment on.

At full annotation the per-rater GED advantage is 0.5%, inside the noise. Fold 1 makes the sparsity-specificity of the advantage concrete: at full annotation the shared baseline actually beats the per-rater model in that fold (GED 0.1552 vs. 0.1658, baseline wins by 6.8%), yet at one annotator per training image the same fold reverses completely (0.2323 vs. 0.1806, per-rater wins by 22.3%). A model that was worse at full annotation in Fold 1 outperforms the shared baseline by 22.3% in the same fold once annotation drops to a single rater. The reversal is tied to the sparsity condition, not to any overall quality difference between the two models: the shared baseline's zero-channel gradients dominate the latent code under single-annotator training in ways that do not occur at full annotation.

1.4 Thesis Contributions

Four contributions follow from this work.

1. **Per-rater posterior encoders that isolate each annotator's gradient at Stage 1.** Giving each annotator a dedicated posterior encoder, trained only on that annotator's mask, eliminates gradient conflict at the source. The isolation property ($\partial \mathscr{L}_i / \partial \mathbf{z}_j = 0$ for $i \neq j$) falls out of the per-rater ELBO decomposition without any additional regularisation. On LIDC-IDRI (1,609 nodule patches, 4-fold cross-validation), this Stage 1-only design achieves GED 0.1444 ± 0.0141 and $\text{Dice}_{\text{match}}$ 0.9112 ± 0.0061, a 4.2% GED reduction and 2.28% $\text{Dice}_{\text{match}}$ gain over the full D-Persona two-stage pipeline; every individual annotator's per-rater Dice improves. On NPC-170 (nasopharyngeal carcinoma MRI), the GED gap is 0.0011, within seed variance, confirming the method transfers without LIDC-specific tuning.

2. **A seven-row ablation that isolates the posterior encoder as the load-bearing component.** Four design alternatives were tested before the per-rater formulation: a transformer backbone (MiT-B2), an orthogonality regularisation loss, a discretised prior bank ($k = 100$), and a dual diversity loss. None improves GED consistently, because none changes what gradients the encoder receives during training. Adding Stage 2 style vectors on top of per-rater Stage 1 (Row 7) makes things worse — GED rises from 0.1444 to 0.1836 — confirming that Stage 2 is only useful when Stage 1 has failed to encode rater signals, and actively harmful in the opposite case.

3. **Sparsity experiments spanning three annotation coverage levels, with a gradient collapse diagnostic.** Training both models at $\text{np} \in \{1, 2, 3\}$ annotators per image shows the per-rater advantage widening as coverage drops: $+11.5\%$

at three annotators, +17.8% at two, +21.4% at one, with the per-rater model winning all 12 per-fold comparisons. The collapse mechanism in the shared baseline is measurable: mean pairwise gradient cosine similarity at the shared latent code rises from 0.167 at full annotation to 0.976 at one annotator, with within-fold standard deviation shrinking 19-fold (0.439 to 0.023) — gradient conflict goes from case-specific to effectively universal.

4. **Pearson correlation analysis linking nine LIDC-IDRI nodule attributes to inter-rater mask variance.** Run on 1,603 cases with complete attribute records, the analysis finds nodule margin clarity ($r = 0.318$, $p < 0.001$, confirmed in all four folds with per-fold r ranging from 0.238 to 0.400) as the strongest predictor of boundary disagreement. Malignancy is negatively correlated ($r = -0.202$): a nodule perceived as highly suspicious need not have an ambiguous contour, and the two dimensions do not necessary travel together. Uncertainty-aware methods are most relevant for ill-defined, lobulated, part-solid nodules rather than for the cases rated most dangerous.

1.5 Thesis Organisation

The literature review in Chapter 2 traces the gradient-conflict problem through probabilistic segmentation and multi-annotator learning, identifying where each strand leaves the shared-encoder bottleneck intact. Chapter 3 formalises the per-rater ELBO, proves gradient isolation, and sets up both the sparse annotation protocol and the two diagnostic analyses. Experimental results, the ablation, and the sparsity study are in Chapter 4, together with a discussion of scope and limitations. Conclusions, open questions, and deployment implications for settings where radiologist coverage is thin are in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

The work reviewed here spans deterministic segmentation, probabilistic models for annotation diversity, multi-annotator label fusion, and gradient conflict in shared-encoder training. Each strand advanced part of the multi-rater segmentation problem. The shared posterior encoder's training dynamics — how conflicting rater gradients accumulate at $\mathbf{z}$ and what that does to the latent code — remained unaddressed across all of them. Chapters 3 and 4 take up that gap directly.

2.1 Medical Image Segmentation and the Uncertainty Problem

Medical image segmentation assigns each pixel to an anatomical structure or pathological region, and that assignment feeds directly into clinical decisions. Ronneberger et al.'s U-Net [14] made dense pixel prediction practical on limited annotated data by pairing a contracting encoder with skip connections that carried high-resolution spatial detail into the expanding decoder. Within two years of the 2015 paper it had become the default architecture across modalities and anatomies — a rapid adoption driven by a simple fit between the architecture and the setting it was designed for, where each training image came with one reference mask.

That assumption breaks down in practice. A radiologist delineating a lung nodule, a cardiologist tracing a ventricular wall, or a radiation oncologist contouring a tumour volume all face the same problem: tissue boundaries in medical images are gradual rather than sharp, imaging noise obscures the transition, and where exactly the structure ends depends on conventions that differ between institutions and between individual practitioners. Multiple trained experts examining the same scan can produce genuinely different boundaries, all of them defensible.

2.1.1 Two Kinds of Uncertainty

Kendall and Gal [7] separated uncertainty into two categories: the kind that shrinks as more training data accumulates (epistemic, about model parameters), and the kind that

persists regardless of how much data is collected (aleatoric, in the observation itself).

Expert disagreement at nodule margins falls in the second category. The four radiologists in LIDC-IDRI are responding to a tissue boundary that the CT image does not resolve to a single location — their disagreement is a property of the image, and collecting more annotated scans would produce more instances of that same disagreement rather than resolving it. A model minimising expected loss against one annotator's masks is calibrated to that annotator's convention and nothing else. What clinical deployment actually requires is a distribution over plausible boundaries, matched to the spread the expert population produces on that specific image.

A single predicted boundary for an ambiguous structure presents one outcome as certain when the clinical reality is a range of defensible contours. The distribution over boundaries carries clinical information too: which boundaries are plausible and how much the experts diverge. A nodule where all four radiologists agree has a fundamentally different uncertainty profile from one where they draw four substantially different contours. A model that collapses both to identical single predictions treats the two situations as equivalent, even though the annotation spread is the clinically meaningful difference between them.

2.1.2 The Scale of Inter-Rater Variability

Joskowicz et al. [6] reviewed CT annotation studies across anatomical structures and found rater disagreement large enough, in many cases, to change clinical management decisions rather than scatter narrowly around a stable consensus. For lung nodules, where ground-glass opacity and tissue density gradients complicate boundary definition, the disagreement grows substantial enough that the annotation label itself becomes the dominant source of training uncertainty — larger in practice than image noise or class imbalance.

The LIDC-IDRI dataset [1], described in detail in §2.6, makes this concrete. Four radiologists independently annotated 1,018 CT scans, and the annotation disagreement is large enough that treating any single radiologist mask as ground truth introduces systematic bias: the resulting model is calibrated to one radiologist's style and uncalibrated to all others (this is established experimentally in §1.2).

2.1.3 Why Deterministic Models Are Insufficient

U-Net and its variants produce single point estimate for each image, one mask per forward pass calibrated to one assumed correct answer. This works well for the problems they were designed to solve, cases with a single clear annotation. For ambiguous anatomy with multiple annotators, this design is structurally incomplete.

A good U-Net can approximate the mean radiologist decision well enough on average. The problem is that it has no mechanism to represent the variability around that mean. It cannot tell a clinician: “for this nodule, two of the four radiologists drew a tight contour around the core and two included the surrounding ground-glass region. My prediction represents one of these choices, not the consensus.” That kind of output — quantified boundary uncertainty, attributed to specific annotator decisions — is what a point-estimate architecture is structurally incapable of generating, regardless of how well the single prediction tracks the mean annotator.

The model the setting requires outputs a distribution over masks, generating multiple distinct hypotheses each reflecting a plausible radiologist decision. Kohl et al., Baumgartner et al., and Monteiro et al. each built toward this goal from different angles.

2.2 Probabilistic Segmentation Models

Moving from deterministic to probabilistic segmentation meant changing what the model’s output represents. Rather than one mask per image, the goal is a learned distribution $p(\hat{y} \mid \mathbf{x})$ from which diverse samples can be drawn at test time. Three architectures developed in this direction, each solving a different part of the problem.

2.2.1 Probabilistic U-Net

Kohl et al. [11] combined a conditional variational autoencoder [9] with the U-Net backbone to create the Probabilistic U-Net (ProbUNet). Its key structural property is that the training pathway and the inference pathway are different. During training, a posterior encoder receives the image together with the annotation, using that annotation signal to supervise the latent variable $\mathbf{z}$. At test time, the posterior encoder is not used at all; $\mathbf{z}$ is sampled from a prior that is conditioned on the image alone. The KL term in the training objective pulls this prior toward the annotation distribution during training, so that prior samples at test time span the range of plausible boundary decisions. This was the first model to produce genuinely varied predictions for the same input. Four samples from ProbUNet span qualitatively different boundary decisions on the same lung nodule; the deterministic U-Net rules this out entirely, since its single forward pass produces one fixed prediction per input with no mechanism to represent boundary ambiguity.

However, ProbUNet has a specific limitation that becomes apparent when multiple annotators are involved. The original model was designed and evaluated with one annotation per training image (randomly selected from among the available annotators). Applying it to a multi-rater dataset requires a decision about what the posterior encoder should receive. In D-Persona’s implementation [17], the posterior is adapted to receive all four rater masks simultaneously. As discussed below, this

adaptation introduces the bottleneck that this thesis targets.

A separate issue is that ProbUNet's 6-dimensional latent space is a single global code for the entire image patch. It captures patch-level uncertainty (whether the overall annotation is large or small, tight or loose) but has limited capacity to represent spatially structured disagreement where radiologists agree on some regions and disagree on others. This limitation motivated the hierarchical extension discussed next.

2.2.2 PHiSeg: Hierarchical Probabilistic Segmentation

Hierarchical latent variable models for segmentation were proposed concurrently by Kohl et al. [10] (Hierarchical ProbUNet) and Baumgartner et al. [3] (PHiSeg). Both start from the observation that annotation uncertainty is not scale-invariant: two radiologists may agree on which region contains a nodule but disagree on where exactly its boundary falls. A single global latent variable cannot separately represent these two distinct levels of disagreement: it conflates a coarse placement decision with a fine boundary decision into one low-dimensional code. PHiSeg's response is to maintain a separate latent variable at each spatial resolution of the encoder, so that disagreements at different scales can be independently captured and sampled.

PHiSeg demonstrated improved sample diversity and calibration on LIDC-IDRI compared to the single-latent ProbUNet, and its samples can differ at fine spatial scales while remaining consistent at coarse scales — a structure that matches expert disagreement better than a single global code.

The gradient conflict problem, however, operates at every level of the hierarchy. Each hierarchical posterior still receives all annotators' masks together, and each scale's shared latent code accumulates the same sum of competing per-rater gradients. Adding more levels multiplies the representational capacity but leaves the training dynamics at each level unchanged. A more expressive hierarchy with same shared-posterior training objective inherits the same bottleneck.

2.2.3 Stochastic Segmentation Networks

Monteiro et al. [12] took a structurally different approach. Rather than compressing annotation uncertainty into a low-dimensional latent variable, they parameterised a distribution directly over the full output mask. This lets SSN represent spatially-structured uncertainty: a nodule core where radiologists agree behaves differently in the output distribution than a margin region where they diverge. A global 6-dimensional latent code cannot represent this distinction, because it collapses the entire image into one low-dimensional uncertainty estimate. SSN makes this spatial parameterisation computationally tractable through a structured covariance approximation. There is no posterior/prior split: the model has one pathway for both training and inference.

SSN remains a single-distribution model, however. It learns one distribution $p(\hat{y} \mid \mathbf{x})$ over all annotations pooled together, without distinguishing individual annotator styles. Sampling from SSN produces masks that span the annotation distribution in aggregate, but individual samples cannot be attributed to specific annotators. In a setting where the clinical goal is to produce four predictions each reflecting one radiologist's style, SSN provides a distribution over the pooled annotation space with no mechanism to steer individual samples toward a specific radiologist's boundary convention.

2.2.4 The Common Bottleneck

ProbUNet, PHiSeg, and SSN each solve part of the problem, producing diverse predictions that outperform deterministic baselines on LIDC-IDRI, yet all three were designed assuming a single training pathway processes all available annotations together. The question of whether mixing annotators' gradients inside a shared encoder is the right design choice was not part of how any of them were built.

Each model, when applied to multi-rater data, encodes annotation information through a single pathway (a latent code, a hierarchical latent code, or a direct output distribution) that must simultaneously accommodate every annotator's mask. The training signal that reaches this shared representation is an aggregate over all annotators. When annotators disagree, these aggregated signals partially cancel each other. The shared representation is pulled toward a weighted compromise across four conflicting annotation signals, one that fits no individual annotator's style well. Methods that explicitly model individual annotators — STAPLE, confusion matrices, calibrated consensus — take a different approach, though as the following review shows, the shared-encoder gradient problem persists in all of them.

2.3 Multi-Annotator Learning

A separate literature addressed multi-annotator learning from a different angle: modelling the annotation process explicitly, asking who annotated what, how reliable each annotator is, and how to aggregate or disentangle their contributions. Zhang et al. [21] provide a recent taxonomy. Five approaches from this line are directly relevant to the per-rater posterior contribution of this thesis.

2.3.1 STAPLE: Simultaneous Truth and Performance Level Estimation

Warfield et al. [16] formalised multi-annotator fusion as an expectation-maximisation problem over two coupled unknowns: what the true segmentation is, and how reliable each annotator is relative to that truth. The E-step estimates per-annotator reliability given the current consensus; the M-step updates the consensus given those reliability

estimates. Iteration continues until both quantities stabilise. The final output is a single weighted consensus mask, with annotators whose labels track the group receiving more influence over the result than those who deviate.

STAPLE is still used in clinical settings where a consensus reference standard is needed for downstream training. For the purposes of this thesis, a single consensus label answers a different question from the one being asked: the goal here is a distribution over plausible annotations, and a consensus collapses exactly the disagreement that distribution is meant to preserve. It models annotator reliability as a property of the annotator relative to an assumed ground truth, not as a source of clinically meaningful variation. For lung nodule segmentation, where expert disagreement reflects genuine boundary ambiguity rather than annotator error, assuming a single hidden true contour that all four radiologists are noisily trying to identify is not appropriate.

2.3.2 Annotator Confusion Models

Rodrigues et al. [13] and Tanno et al. [15] developed more flexible models of annotator behaviour. Rodrigues et al. placed Gaussian process classifiers over each annotator's labels with per-annotator latent functions, marginalising over annotator-specific noise to estimate a true label. Tanno et al. [15] took a discriminative approach: each annotator is characterised by a per-class confusion matrix encoding how systematically they deviate from the true label. This gives each annotator a learned behavioural fingerprint (rather than a single reliability scalar), so the model can distinguish an annotator who consistently over-segments from one whose labels are correct on average but noisy at boundaries. Both models learn annotator-specific behaviour and produce consensus estimates that account for those learned properties.

Both models move beyond STAPLE's single reliability scalar. Annotator-specific parameterisation lets the model weight each annotator's contribution in proportion to how much their labels agree with the others, so a radiologist who consistently over-segments gets lower weight than one who tracks the group more closely. Tanno et al.'s confusion matrix formulation has been applied to medical image segmentation, where the per-annotator reliability estimate produces more accurate consensus labels than treating all annotators equally.

The limitation is the same one that affects STAPLE: the output is a consensus label, not a distribution over plausible annotations. The confusion matrix framework also assumes there exists a true class label that annotators are confusedly trying to identify. This assumption is defensible for tasks like pathology slide classification where a ground truth (the tissue pathology) exists in principle, even if it is difficult to observe. For lung nodule boundary delineation, the assumption breaks down because the tissue transition at the nodule edge is gradual enough that four qualified radiologists can draw four genuinely different contours, each defensible, with no hidden ground truth contour that any of them is approximating.

2.3.3 Multi-Rater Calibration

Rather than averaging annotator decisions uniformly, Ji et al. [5] weight each annotator's contribution proportionally to how much that annotator's label matches the broader population at each pixel, producing a consensus more representative of the annotation distribution than a simple mean.

For many clinical applications that calibrated consensus is useful — a model that recognises when four radiologists converge versus when they split two-against-two carries more information than one that always produces the same confidence. Producing four distinct predictions each reflecting one radiologist's boundary style, spanning the full range of clinically defensible decisions, is a different goal and one that weighted consensus methods leave unaddressed.

2.3.4 Disentangling Human Error from Ground Truth

Zhang et al. [20] introduced a framework that separated annotation variance into two structurally distinct components: variability driven by the image itself (present regardless of which annotator is involved) and variability driven by individual annotator tendencies (idiosyncratic errors or biases specific to each person). Because these two sources have different structures (one is a property of the image, the other is a property of the annotator), pooling them into a single noise term wastes model capacity and produces poorly calibrated outputs. Treating them separately allows the model to assign image-level uncertainty to cases where any expert would struggle, and annotator-level uncertainty to cases where one annotator systematically diverges from the others.

This decomposition is the closest conceptually to what per-rater posterior encoders achieve. By modelling each annotator's contribution separately and identifying the annotator-specific component, Zhang et al. move toward individual-level annotation modelling rather than population-level averaging. The limitation is that the framework produces a discriminative model (a classifier with per-annotator noise components) rather than a generative model that can sample from the distribution of plausible annotations. GED evaluation requires drawing multiple diverse prior samples at test time, and a discriminative classifier with no prior has no sampling mechanism to support this.

2.3.5 D-Persona: Per-Rater Personalisation of Shared Posteriors

D-Persona [17] introduced a two-stage training pipeline that combined the latent-space diversity of ProbUNet with per-rater personalisation — the first method to target individual radiologist style from a probabilistic segmentation base.

In Stage 1, a shared posterior encoder receives all four annotators' masks simultaneously as a single multi-channel input and produces one shared latent code. This adapts ProbUNet to the multi-rater setting by widening the posterior from one annotation to four. An auxiliary range loss pushes the prior to cover the full breadth of annotator decisions rather than collapsing toward the mean. Feeding all four masks into one encoder produces gradient conflict at the shared latent code.

Stage 2 adds per-rater projection heads on top of the frozen Stage 1 prior. Stage 1 parameters are not updated. Each head takes the shared $\mathbf{z}$ and applies a small rater-specific transformation, steering samples toward that annotator's boundary style at inference time. Stage 2 can steer samples toward a rater's style only if that style is already present somewhere in the shared latent code. Whether that condition holds depends on what Stage 1's training objective actually encodes, and the gradient analysis in §3.2.3 examines this.

On LIDC-IDRI, D-Persona substantially outperforms ProbUNet on both GED and $\text{Dice}_{\text{match}}$; full numerical comparisons are in Table 4.2.

12 D-Persona represents the state of the art in per-rater prediction, but its Stage 1 training introduces a structural bottleneck that Stage 2 cannot compensate for. Because the shared encoder processes all four annotators' masks simultaneously, the gradient update at the shared latent code is an average of four per-rater terms. On nodules where all four radiologists draw similar boundaries, these terms reinforce each other and the shared code receives a clean signal. At ill-defined nodule margins — the region that matters clinically — individual rater gradients point in different directions and partially cancel, pulling the latent code toward a weighted average that fits none of the four boundary decisions cleanly; what the shared $\mathbf{z}$ encodes is shaped by all four radiologists at once rather than by any single annotator's boundary convention. The full derivation of this gradient conflict is in §3.2.3.

Stage 2 personalisation then has to extract per-rater information from a $\mathbf{z}$ that does not contain it. Whatever the projection heads h_i do to the shared code, they are working with the gradient-averaged residual from Stage 1 — whatever was cancelled out during Stage 1's backward pass is gone before h_i ever runs. Stage 1's gradient averaging is the bottleneck, and Stage 2 operates too late in the pipeline to address it.

The ablation in §4.2.3 confirms this empirically: adding Stage 2 style vectors after a Stage 1 trained with per-rater posteriors worsens GED by 27% (from 0.1444 to 0.1836). Stage 2 projection heads find something to steer only when Stage 1 has left per-rater structure underspecified in $\mathbf{z}$. After a per-rater Stage 1, that structure is already encoded cleanly, and the heads collapse it.

Giving each annotator a dedicated posterior encoder — trained on that annotator's mask alone — removes the gradient averaging from Stage 1's training objective, the point where the information loss actually occurs.

2.4 Variational Inference and the ELBO

The per-rater posterior formulation rests on standard variational inference. Two properties of the ELBO are relevant: that the per-rater objective is a valid variational lower bound, and that the $\frac{1}{N}$ normalisation keeps KL regularisation pressure comparable to the baseline — without it, four KL terms together would over-regularise and collapse the latent diversity GED requires. §2.5 then connects these mathematical foundations to the gradient conflict problem.

2.4.1 Variational Autoencoders and the Evidence Lower Bound

Kingma and Welling [9] introduced the variational autoencoder (VAE) as a framework for learning latent-variable generative models. The difficulty is tractability: the true posterior $p(\mathbf{z} \mid x)$ requires integrating over all possible $\mathbf{z}$, which has no closed form for continuous latent variables. The VAE resolves this through amortised inference: a learned encoder $q_\phi(\mathbf{z} \mid x)$ stands in for the intractable posterior, and the training objective is constructed so that pushing this approximation closer to the true posterior simultaneously improves the generative model.

The ELBO is derived by applying Jensen's inequality to the log-marginal likelihood:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(\mathbf{z}|x)}\big[\log p_\theta(x \mid \mathbf{z})\big] - \mathrm{KL}\big(q_\phi(\mathbf{z}|x) \,\|\, p(\mathbf{z})\big). \tag{2.1}$$

The reconstruction term measures how well the decoder recovers the original observation from a latent sample. The KL term keeps the approximate posterior anchored near the prior: an unconstrained posterior would collapse each observation to a near-deterministic code, fitting training data well but producing incoherent outputs at test time when the prior is sampled instead. Maximising the bound jointly over encoder and decoder parameters drives both toward a useful equilibrium.

In ProbUNet and D-Persona, the observation is the segmentation mask y, the latent variable is $\mathbf{z}$, and both the prior and posterior are conditioned on the image $\mathbf{x}$:

$$\log p(y \mid \mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)}\big[\log p(y \mid \mathbf{z},\mathbf{x})\big] - \mathrm{KL}\big(q(\mathbf{z}|\mathbf{x},y) \,\|\, p(\mathbf{z}|\mathbf{x})\big). \tag{2.2}$$

2.4.2 The AxisAlignedConvGaussian Encoder

Both the prior $p(\mathbf{z} \mid \mathbf{x})$ and the posterior $q(\mathbf{z} \mid \mathbf{x}, y)$ in ProbUNet are implemented as AxisAlignedConvGaussian networks [11]: convolutional encoders whose output is a mean μ and a log-variance $\log \sigma^2$, one scalar per latent dimension. The diagonal covariance assumption is an efficiency choice: a full $D \times D$ covariance would be

impractical to parameterise and invert, while a diagonal Gaussian with flexible means and variances retains enough expressiveness for the 6-dimensional latent space used here.

Backpropagating through a stochastic sample would ordinarily block gradient flow, since drawing from a distribution is not a differentiable operation. Kingma and Welling's reparameterisation [9] resolves this: by expressing the sample as $\mathbf{z} = \mu + \sigma \odot \varepsilon$, the stochastic draw is factored into a deterministic function of encoder outputs and a noise variable $\varepsilon \sim \mathcal{N}(0, I)$ that carries no parameter dependence. Gradients reach μ and σ through the deterministic path; ε is never differentiated.

2.4.3 Additivity of the ELBO for Independent Observations

The mathematical foundation for the per-rater ELBO is a standard property of the ELBO under independent observations. If N observations $\{y_i\}_{i=1}^{N}$ are conditionally independent given $\mathbf{z}$ and $\mathbf{x}$, then the log-joint likelihood decomposes:

$$\log p(y_1, \ldots, y_N \mid \mathbf{x}) = \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}). \tag{2.3}$$

Applying the ELBO bound to each term independently and summing:

$$\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}) \geq \sum_{i=1}^{N} \Big[\mathbb{E}_{q_i(\mathbf{z}|\mathbf{x},y_i)} \big[\log p(y_i \mid \mathbf{z}, \mathbf{x}) \big] - \mathrm{KL}\big(q_i \,\|\, p(\mathbf{z}|\mathbf{x})\big) \Big]. \tag{2.4}$$

The right-hand side of Equation 2.4 is the sum of N valid lower bounds, each for a different posterior q_i and observation y_i; since linearity preserves the bound, their sum is also a valid lower bound. Normalising by N and adding the auxiliary range loss $\mathcal{L}_{\text{bound}}$ gives the per-rater ELBO objective of Equation 3.5 (Chapter 3). No approximation beyond the standard variational one is needed; the per-rater ELBO follows from ELBO additivity without any additional modelling assumption.

The $\frac{1}{N}$ normalisation has a practical consequence that follows directly from the term count. Without it, N KL terms each weighted 1 exert $4\times$ the regularisation pressure of the baseline's single KL term, which would over-regularise the posteriors toward the prior and collapse the latent space diversity GED requires. Dividing by N keeps the total KL pressure equal to the baseline — each rater's posterior pulls the shared prior with the same weight as in the original formulation.

2.5 Gradient Conflict in Multi-Task Learning

Each annotator's reconstruction loss can be treated as a separate task competing over the shared encoder — the same structural situation that motivated gradient conflict

research in multi-task learning. Multi-task learning's findings about shared-encoder gradient interference translate directly to the shared posterior setting and provide both the diagnostic tools and the architectural vocabulary for addressing it.

2.5.1 Gradient Surgery and the Interference Problem

Yu et al. [19] showed that a shared encoder trained on tasks with opposing gradients receives a net update that is less informative for each individual task than a task-specific update would be. The per-task performance loss tracks the cosine similarity between task gradients. The more conflicted the gradients, the worse each task learns. Their practical contribution was a gradient projection fix: modifying each task's gradient to remove the component that conflicts with the other tasks before the shared encoder update is applied. What matters for our purposes is diagnostic framework: gradient conflict in shared encoders is measurable and its magnitude predicts how much each task suffers.

2.5.2 Multi-Rater Segmentation as Multi-Task Learning

In the D-Persona Stage 1 setting, the shared posterior encoder processes all four annotators' labels simultaneously, which maps directly onto the multi-task learning structure that motivated Gradient Surgery. The reconstruction loss for the i-th annotator is one task, and the four tasks share the single posterior encoder as their joint feature extractor. When two annotators disagree on a boundary, the reconstruction gradients from their respective tasks point in opposite directions in the latent space, creating exactly the conflict that Gradient Surgery was designed to mitigate.

Per-rater posterior encoders are the multi-task analogue of task-specific feature extractors: each annotator's task gets a dedicated encoder, and gradient conflict is eliminated because no two tasks share the same encoder parameters. The decoder (FComb) remains shared, but at the decoder level the gradients are N separate signals arriving from independent encoders, not N conflicting signals mixed inside a shared encoder.

Per-rater posteriors apply this principle to probabilistic multi-rater segmentation, where the variational framework and ELBO additivity provide the mathematical justification for treating each annotator as an independent task.

The connection also makes the solution's failure mode predictable. In multi-task learning, task-specific encoders are most valuable when tasks conflict. When tasks align, a shared encoder can be more data-efficient because all tasks reinforce the same representation. By analogy, per-rater posteriors should provide the largest advantage when annotators disagree most: under annotation sparsity (where absent annotators' zero-channel gradients introduce a particularly severe form of conflict) and for nodules

with ill-defined margins (where annotation disagreement is structurally highest). Chapter 4 reports both effects, with the sparsity advantage growing monotonically from $+11.5\%$ at three annotators to $+21.4\%$ at one.

2.6 The LIDC-IDRI Dataset and Evaluation Metrics

2.6.1 LIDC-IDRI

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [1] is the standard benchmark for multi-rater lung nodule segmentation. It contains 1,018 low-dose CT scans from seven academic medical centres, each read by four thoracic radiologists under a two-phase protocol: an initial blinded read where each radiologist marked lesions independently, followed by a revision pass in which the four readers could see each other's markings before locking their final contours. Because every radiologist had access to the others' boundaries during the revision phase, contours that still differ after revision reflect deliberate interpretive choices rather than isolated reading errors.

Armato et al. [2] quantified the surviving disagreement and found it large enough, and consistent enough across the full scan collection, to rule out fatigue or ambiguous instructions as explanations. What persists is the tissue itself: density gradients at a pulmonary nodule's periphery do not resolve to a single edge in CT, and four readers trained on the same guidelines can still place four defensible contours at different locations on the same case.

Following D-Persona's preprocessing protocol [17], cases with fewer than four complete rater annotations are excluded, and the remaining cases are cropped to 128×128 patches centred on each annotated nodule, yielding 1,609 patches. Each patch has four binary segmentation masks, one per radiologist, and nine per-radiologist clinical attribute ratings (malignancy, texture, spiculation, lobulation, margin, sphericity, calcification, internalStructure, subtlety) on an integer scale of 1–5. These attribute ratings are used in the annotation disagreement characterisation analysis (§4.5).

The dataset is split into four cross-validation folds following D-Persona's exact partitions, with test fold sizes of 450, 375, 412, and 372 patches respectively. All methods are trained and evaluated on identical splits; all comparisons are therefore direct.

2.6.2 NPC-170

The NPC-170 dataset [17] provides a cross-dataset validation context. It contains 170 nasopharyngeal carcinoma MRI cases, each annotated by four independent annotators

on 3-channel MRI (T1, T1CE, T2 sequences). The training set comprises 2,405 slices; validation and test sets each contain 20 cases. NPC-170 differs from LIDC-IDRI in modality (MRI vs CT), anatomy (head and neck vs thorax), and imaging protocol (multi-channel vs single-channel). Using it as a secondary validation tests whether the per-rater design transfers across these differences.

2.6.3 Generalised Energy Distance

The Generalised Energy Distance (GED) [11] is the primary evaluation metric throughout this thesis. Its value as a metric comes from what it jointly rewards: a model must produce predictions that are both geometrically close to individual annotator masks and spread across the range those annotators cover. Collapsing to one prediction, no matter how accurate, drives the within-sample diversity term to zero and GED climbs; spreading samples widely without tracking any rater's boundary is equally penalised. GED reaches zero only when the model's sampling distribution and the annotation distribution are identical. The formal definition is in §3.1.

2.7 Summary of Research Gaps

Each of the four literature strands above advanced multi-rater segmentation in one direction while leaving a specific gap open — gap that this thesis addresses.

Every probabilistic multi-rater segmentation model reviewed in §2.2 and §2.3 trains with a shared posterior that encodes all annotators together. The reconstruction gradient at the shared latent code is a sum of per-rater terms; when annotators disagree, these partially cancel and $\mathbf{z}$ drifts toward a compromise that fits no individual style. Giving each annotator a dedicated posterior encoder — one that receives only that annotator's mask and therefore carries only that annotator's gradient — is a design the literature reviewed here leaves unexplored.

Building on this gap, D-Persona's Stage 2 projection heads attempt to personalise a shared $\mathbf{z}$ by steering it toward individual rater styles at inference time. Projection heads can only redistribute structure already in the latent code. Rater-specific information that Stage 1's gradient averaging removed is gone before Stage 2 runs. The ablation in §4.2.3 confirms the consequence: adding Stage 2 on top of a per-rater Stage 1 — where the latent code already has clean rater-specific gradients — actively worsens GED.

A third gap concerns annotation coverage. Real clinical datasets rarely have every image annotated by all radiologists, yet every probabilistic multi-rater method in this literature is evaluated under complete coverage. How absent annotators degrade a shared posterior — specifically, zero-channel inputs generating empty-mask gradients that contaminate $\mathbf{z}$ — has received no systematic treatment.

Finally, LIDC-IDRI provides nine clinical attribute ratings per nodule per radiologist, but no systematic analysis has asked which attributes drive inter-rater boundary disagreement. Such an analysis would show where uncertainty-aware methods matter most and where a deterministic model is adequate. §3.7 takes this up, with results in §4.5.

CHAPTER 3

METHODOLOGY

Understanding why per-rater posteriors work requires first seeing what goes wrong inside the shared baseline during training. The gradient conflict analysis in §3.2 establishes the baseline bottleneck; §3.4 then derives the per-rater ELBO, proves the gradient isolation property, and specifies the architecture. Sparse annotation training (§3.5) and the two diagnostic analyses — gradient alignment (§3.6) and nodule attribute characterisation (§3.7) — follow from the same structural difference between the two designs.

3.1 Problem Formulation

Let $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ denote an input medical image, and let $\{y_i\}_{i=1}^{N}$ be the binary segmentation masks provided by N independent annotators, each $y_i \in \{0,1\}^{H\times W}$. For LIDC-IDRI, $H = W = 128$, $C = 1$ (grayscale CT), and $N = 4$. For NPC-170, $C = 3$ (T1, T1CE, T2 MRI channels) with the same N.

The goal is to learn a model that, given $\mathbf{x}$ at test time, can generate a set of diverse segmentation predictions that reflect the true spread of annotation disagreement: not one averaged prediction, and not N identical samples. Formally, we want a learned distribution $p(\hat{y} \mid \mathbf{x})$ from which samples span the range of plausible annotator decisions.

3.1.1 The Generalised Energy Distance

We adopt the Generalised Energy Distance (GED) [11] as the primary metric because it formalises this joint requirement in a single score. For a set of S model predictions $\{\hat{y}_s\}$ and a set of N ground-truth annotations $\{y_i\}$:

$$\text{GED} = 2\,\mathbb{E}\big[d(\hat{y},y)\big] - \mathbb{E}\big[d(\hat{y},\hat{y}')\big] - \mathbb{E}\big[d(y,y')\big] \tag{3.1}$$

where $d(\cdot,\cdot)$ is a ground distance (here, $1-\text{IoU}$), and the expectations are over independently drawn samples from the model distribution and the annotation set respectively.

GED equals zero only when the model distribution and the annotation distribution are identical. The second term penalises a model that collapses to a single prediction: if all $\hat{y}_s$ are identical, $\mathbb{E}[d(\hat{y},\hat{y}')] = 0$, and GED reduces to twice the average prediction error. A model can achieve low prediction error on each individual rater mask and still score poorly on GED if its predictions are insufficiently diverse.

Diversity and accuracy are ordinarily at odds, since a model that improves GED by producing more diverse predictions would normally sacrifice per-rater accuracy ($\text{Dice}_{\text{match}}$). Per-rater posteriors improve both simultaneously, as the results in Chapter 4 show — an outcome that points to better gradient signal quality during training rather than a diversity-accuracy trade-off, since the latter would move the two metrics in opposite directions.

3.1.2 Probabilistic Formulation

We follow the variational framework of the Probabilistic U-Net [11]. A latent variable $\mathbf{z} \in \mathbb{R}^D$ ($D = 6$ throughout) captures annotation uncertainty. An image-conditioned prior $p(\mathbf{z} \mid \mathbf{x})$ defines the sampling distribution at test time: drawing $\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})$ and passing it through a decoder $f_{\text{comb}}(\mathbf{z}, \mathbf{x})$ produces one predicted mask. Diversity over multiple draws then reflects uncertainty captured in $p(\mathbf{z} \mid \mathbf{x})$. Training requires a posterior $q(\mathbf{z} \mid \mathbf{x}, \text{annotation})$ that pushes $p(\mathbf{z} \mid \mathbf{x})$ toward the annotation distribution. The design of this posterior is the crux of both the baseline and the proposed method. Table 3.1 collects the symbols and variables that appear from this point on.

Table 3.1 Notation index for Chapters 3 and 4.

Symbol	Meaning
$\mathbf{x}$	Input image
y_i	Segmentation mask from annotator i
Y	All annotation masks $\{y_1, \ldots, y_N\}$
N	Number of annotators (4 throughout)
$\mathbf{z}$, $\mathbf{z}_i$	Shared / per-rater latent variable
D	Latent dimension ($D = 6$)
$q(\mathbf{z} \mid \mathbf{x}, Y)$	Shared posterior (baseline)
$q_i(\mathbf{z} \mid \mathbf{x}, y_i)$	Per-rater posterior for annotator i (proposed)
$p(\mathbf{z} \mid \mathbf{x})$	Image-conditioned prior
f_{comb}	Segmentation decoder (FComb)
β	KL weighting coefficient ($\beta = 0.5$)
$\mathcal{L}_{\text{bound}}$	Auxiliary range loss (D-Persona)
np	Number of annotators present during training

3.2 Baseline: D-Persona with a Shared Posterior

D-Persona [17] is the baseline. Among published methods for per-rater personalised probabilistic segmentation it achieves the best reported GED and $\text{Dice}_{\text{match}}$ on LIDC-IDRI, and its two-stage architecture contains the gradient conflict that the per-rater posterior is designed to remove.

3.2.1 Architecture

D-Persona builds on the Probabilistic U-Net [11]. The backbone is a ResNet34 encoder [4] that extracts image features. The prior network is an AxisAlignedConvGaussian that takes the image $\mathbf{x}$ alone and outputs a diagonal Gaussian $p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mu_p, \text{diag}(\sigma_p^2))$. The posterior network has the same architecture but takes a concatenation of the image and *all four* rater masks as input (a 5-channel tensor (x, y_1, y_2, y_3, y_4)). It outputs a shared posterior $q(\mathbf{z} \mid \mathbf{x}, Y) = \mathcal{N}(\mu_q, \text{diag}(\sigma_q^2))$.

A lightweight FComb decoder [11] takes two inputs: the U-Net's spatial feature maps (passed through skip connections) and a single sample drawn from the latent distribution. The spatial features supply local boundary detail; the latent sample globally biases which boundary decision the decoder resolves to. This is Stage 1. Stage 2 adds four per-rater projection heads $\{h_i\}_{i=1}^{N}$, each a small MLP that takes the shared $\mathbf{z}$ and produces a rater-specific style offset, which is then added to $\mathbf{z}$ before decoding. Stage 1 parameters are frozen before Stage 2 begins; the projection heads operate on whatever $\mathbf{z}$ Stage 1 produced, with no path back to the Stage 1 training objective.

3.2.2 The Stage 1 Training Objective

Stage 1 is trained with the following evidence lower bound (ELBO) [9]:

$$\begin{aligned} \mathcal{L}_{\text{base}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x},Y)}\big[\log p(Y \mid \mathbf{z}, \mathbf{x})\big] \\ - \text{KL}\big(q(\mathbf{z}|\mathbf{x},Y) \,\|\, p(\mathbf{z}|\mathbf{x})\big) + \beta\, \mathcal{L}_{\text{bound}}. \end{aligned} \tag{3.2}$$

The first term is the reconstruction likelihood: the posterior sample $\mathbf{z}$ must decode to all four rater masks well. The second term is the KL divergence that pulls the posterior toward the prior. This keeps the latent space coherent for test-time sampling. The third term, $\mathcal{L}_{\text{bound}}$, is an auxiliary diversity floor loss from D-Persona. It operates by drawing samples from the prior at training time and applying a hinge penalty whenever the spread of the decoded outputs falls below the observed spread of the four rater masks. Unlike the reconstruction term, it acts on prior samples rather than posterior samples, shaping the model's test-time output distribution. We retain this term unchanged with $\beta = 0.5$.

3.2.3 The Gradient Conflict Problem

The reconstruction likelihood in Equation 3.2 requires the single posterior sample $\mathbf{z}$ to simultaneously explain all four rater masks. Backpropagating through the first term gives the gradient at $\mathbf{z}$:

$$\frac{\partial \mathcal{L}_{\text{base}}}{\partial \mathbf{z}} = \sum_{i=1}^{N} \frac{\partial \log p(y_i \mid \mathbf{z}, \mathbf{x})}{\partial \mathbf{z}}. \tag{3.3}$$

This gradient is a sum of N per-rater terms. When all four annotators agree, these terms reinforce each other. But when they disagree on a boundary, as they routinely do at ill-defined nodule margins in LIDC-IDRI, the per-rater gradients point in different directions. For a pixel that rater i labels as lesion and rater j labels as background, the gradient $\partial \log p(y_i \mid \mathbf{z}, \mathbf{x})/\partial \mathbf{z}$ pushes $\mathbf{z}$ toward predicting 1 for that pixel, while $\partial \log p(y_j \mid \mathbf{z}, \mathbf{x})/\partial \mathbf{z}$ pushes it toward 0. In the extreme case of complete disagreement ($y_i = 1 - y_j$ everywhere), the two terms cancel exactly and the net gradient at $\mathbf{z}$ is zero.

In practice the cancellation is partial, not total, and it is case-dependent. At full annotation, we measure the mean pairwise cosine similarity between per-rater reconstruction gradients at $\mathbf{z}$: it is 0.167, distributed across test cases with a within-fold standard deviation of 0.439 (meaning the conflict is much worse for some cases than others). At full annotation the shared posterior still receives a net signal — the conflict does not zero out the gradient entirely — but the latent space degrades: rather than encoding the distribution of annotation styles, the shared $\mathbf{z}$ is pulled toward a compromise that fits no individual rater well.

3.2.4 Why Stage 2 Cannot Fix This

Stage 2 projection heads h_i steer samples from the shared $\mathbf{z}$ toward individual annotator styles by applying a small rater-specific offset before decoding. For this to work, Stage 1 must encode per-rater structure into $\mathbf{z}$ in the first place. Whatever per-rater information gradient averaging removed during Stage 1's backward pass is absent from $\mathbf{z}$ before h_i ever executes; the projection heads operate on the gradient-averaged residual, not on the original rater signals.

The ablation study in §4.2.3 makes this failure mode concrete. When Stage 2 is applied on top of a Stage 1 trained with per-rater posteriors — where $\mathbf{z}$ already carries clean, rater-specific gradient signals — GED rises from 0.1444 to 0.1836, a 27% degradation. Stage 2 disrupts the diversity that Stage 1 already encoded, because it was designed to compensate for a deficit that no longer exists.

3.3 Transformer Encoder Investigation

The gradient conflict analysis in §3.2.3 establishes that the shared posterior's training signal is a sum of conflicting per-rater gradients. Before concluding that the fix must operate at the training objective level, a simpler alternative must be ruled out: that the bottleneck is *representational capacity*. ResNet34 was designed for image classification and its convolutional structure constrains how far apart two spatial locations can interact within one forward pass. If the gradient blurring is in fact a capacity problem (a representation too limited to encode four raters' styles simultaneously), then a more expressive backbone should alleviate it regardless of how the training objective is structured. The experiment tests whether an architectural fix suffices or whether the training objective itself must change.

We test this hypothesis by replacing the ResNet34 backbone in the shared posterior encoder with MiT-B2 [18], a Mix Transformer with a substantially larger parameter count and attention-based feature extraction. The encoder architecture, input channels, and training objective are otherwise unchanged.

The result, reported in §4.2.3 (Table 4.4, Row 2), is unambiguous. MiT-B2 achieves GED 0.1531, marginally *worse* than the ResNet34 baseline at 0.1507, with no change in $\text{Dice}_{\text{match}}$. Attention mechanisms and long-range spatial reasoning do not alleviate the gradient conflict problem. A more expressive encoder learns a more expressive weighted average of conflicting rater signals; the conflict itself is a property of the training objective, not of the encoder's receptive field. Both encoders receive the same 5-channel shared-posterior input and the same sum of per-rater reconstruction gradients — the training signal reaching each backbone is structurally the same, whatever the backbone's capacity for representing it.

3.4 Per-Rater Posterior Encoders

Instead of one shared encoder receiving all four masks simultaneously, each annotator gets a dedicated posterior encoder conditioned only on that annotator's mask.

3.4.1 Design

We replace the single shared posterior with N independent encoders, each an AxisAlignedConvGaussian identical in architecture to the original:

$$q_i(\mathbf{z} \mid \mathbf{x}, y_i) = \mathcal{N}\big(\mu_i, \text{diag}(\sigma_i^2)\big), \quad i = 1, \ldots, N. \tag{3.4}$$

Each encoder takes a 2-channel input: the image $\mathbf{x}$ and rater i's mask y_i. The input width drops from 5 channels (baseline) to 2 channels per encoder. No weights are shared

between encoders. The prior $p(\mathbf{z} \mid \mathbf{x})$ and the FComb decoder f_{comb} are identical to the baseline: the only change is in how the posterior signals are collected during training.

At test time, there are no posterior encoders in the loop. Predictions are drawn by sampling $\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})$ and passing through $f_{\text{comb}}(\mathbf{z}, \mathbf{x})$, exactly as in the baseline. The inference procedure is unchanged, and inference cost is identical.

What separates the two architectures is where in the computation graph the four raters' gradients first meet. In the baseline they meet inside the shared encoder, before any rater-specific information can be preserved; in the proposed design they meet only at the decoder, after each encoder has independently updated its representation from single rater's mask. Figure 3.1 shows both paths side by side.

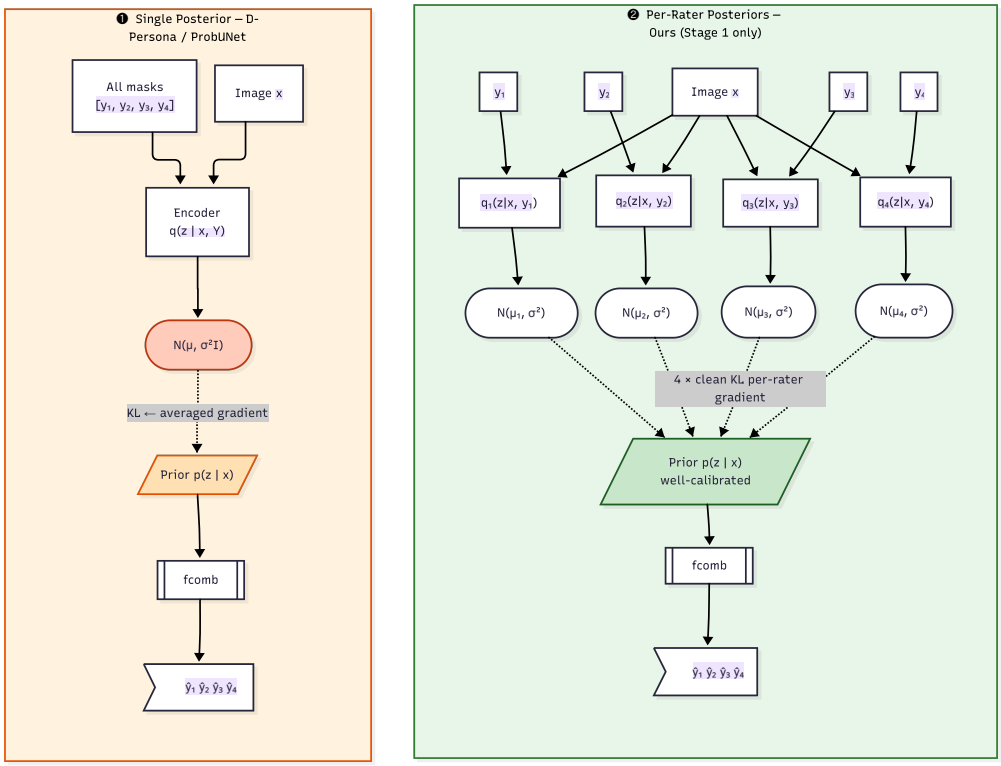

Figure 3.1 Architecture: shared posterior baseline (left) vs. per-rater design (right).

3.4.2 Training Objective

With N independent posteriors, the joint training objective decomposes into N separate ELBO terms:

$$\mathcal{L}_{\text{ours}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{q_i}\big[\log p(y_i \mid \mathbf{z}, \mathbf{x})\big] - \frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}\big(q_i \,\|\, p(\mathbf{z}|\mathbf{x})\big) + \beta\,\mathcal{L}_{\text{bound}}. \tag{3.5}$$

Each summand is a standard ELBO for rater i. Because a sum of valid lower bounds is itself a valid lower bound (by linearity), $\mathcal{L}_{\text{ours}}$ is a proper variational lower bound on $\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x})$.

The $\frac{1}{N}$ normalisation matters practically. Without it, the N KL terms together exert N times the regularisation pressure of the baseline's single KL term, which would over-regularise the posteriors toward the prior and collapse the latent space diversity that GED requires. Dividing by N keeps the total KL pressure equal to the baseline — each rater's posterior pulls the shared prior with the same weight as the single posterior in the original formulation, and no single annotator's distribution dominates.

3.4.3 Gradient Isolation

With N independent encoders, per-rater gradients cannot reach each other's parameters.

Proposition. *For the per-rater ELBO $\mathcal{L}_{\text{ours}}$, the gradient of the i-th reconstruction term with respect to the j-th latent variable is zero for all $i \neq j$: $\partial \mathcal{L}_i / \partial \mathbf{z}_j = 0$.*

Proof. Let $\mathcal{L}_i = \mathbb{E}_{q_i(\mathbf{z}_i|\mathbf{x},y_i)}[\log p(y_i \mid \mathbf{z}_i, \mathbf{x})]$. The expectation is taken over $\mathbf{z}_i \sim q_i$, which depends only on $(\mathbf{x}, y_i)$. Since q_i does not depend on q_j for $j \neq i$, and $\mathbf{z}_i$ is drawn independently from $\mathbf{z}_j$, the term $\mathcal{L}_i$ is a function only of the parameters of q_i. By the chain rule, $\partial \mathcal{L}_i / \partial \mathbf{z}_j = 0$ for all $j \neq i$. □

The reconstruction gradient that reaches encoder q_i reflects rater i's mask exclusively, a structural consequence of the objective decomposition with no regularisation term required to enforce it. What rater j labelled has no effect on how encoder q_i is trained. The FComb decoder still receives the gradients from all N reconstruction terms, but they arrive as N separate, rater-specific signals — each encoder has already processed its own rater's mask independently before anything reaches the decoder. The decoder learns to reconcile them through usual gradient accumulation that any multi-task network experiences, but each signal is clean and rater-specific when it arrives.

To see why this matters, contrast with Equation 3.3. In the baseline, the N rater gradients are summed before reaching any part of the encoder; in the proposed objective, they are summed only at the decoder, after each encoder has been updated with its own rater's signal. What changes between the two designs is what information the encoder is trained to represent, not how the decoder processes it afterward.

3.4.4 Architectural Details and Implementation

The implementation adds $N - 1 = 3$ additional AxisAlignedConvGaussian encoder networks to the D-Persona Stage 1 architecture. Each has same convolutional structure as the original posterior encoder and is initialised independently. No weight tying, no shared convolutional layer. The prior network and FComb decoder are taken from the D-Persona codebase without modification. Stage 2 projection heads are not used.

Training follows the D-Persona hyperparameters exactly to ensure fair comparison: Adam optimiser [8] with learning rate 10^{-4}, cosine annealing schedule, batch size 12, latent dimension $D = 6$, and 100 training epochs per fold. Four-fold cross-validation is used on LIDC-IDRI, with fold test sizes of 450, 375, 412, and 372 nodule patches respectively.

The additional encoders increase training time significantly. Stage 1 with per-rater posteriors takes approximately 47 hours across all four folds on Apple MPS hardware (4.5 times the baseline Stage 1 training time of approximately 10.5 hours, and 2.5 times the full D-Persona two-stage pipeline of approximately 18.8 hours). We acknowledge this as a genuine practical barrier, particularly for institutions without GPU access. Inference cost, however, is identical to the baseline: one sample from $p(\mathbf{z} \mid \mathbf{x})$ and one decoder forward pass.

3.5 Sparse Annotation Training

The experiments in Chapter 4 include a systematic study of how both models behave when annotation coverage is incomplete.

3.5.1 Motivation

In clinical datasets, it is common for different radiologists to annotate different subsets of cases. The LIDC-IDRI dataset used for the main experiments enforces complete four-rater annotation, but this is atypical of real deployment conditions. A model that relies on complete annotation coverage would be fragile in practice. Most large clinical collections do not have every image annotated by all radiologists. The sparse annotation experiments ask whether per-rater posteriors handle this condition better than the shared

baseline.

The sparsity level is controlled by a parameter $\text{np} \in \{1,2,3\}$ specifying how many annotators provide masks for each training image at each training step. Full annotation corresponds to $\text{np} = 4$ and is used as the reference point. In each training step, $N - \text{np}$ annotators are chosen at random and treated as absent.

3.5.2 The `drop_raters()` Mechanism and Its Asymmetric Effect

Absent annotators are handled asymmetrically by the two architectures, and this asymmetry drives the sparsity results.

In the shared baseline, an absent annotator's mask channel is set to zero. At $\text{np} = 2$ with annotators a and b present, the 5-channel input is $(x, y_a, y_b, 0, 0)$. The encoder processes all five channels — including the two zeros — and the reconstruction gradient from the zero channels pushes $\mathbf{z}$ toward predicting empty segmentations for the absent raters. At $\text{np} = 1$, three of the four gradient contributions point toward empty-mask predictions and only one carries real annotation signal. The latent code is simultaneously pulled toward the present annotator's boundary decision and toward predicting nothing on three channels that carry no real information.

In the per-rater model, if rater j is absent for a given training step, encoder q_j does not run. There is no zero-channel input and no loss term for that rater. The gradient for that step comes only from the np present encoders, each providing a clean signal from its own rater's mask, and absent encoders contribute nothing: no zero-mask contamination and no competing gradient. This follows structurally from the per-encoder architecture.

3.5.3 Training Configuration for Sparse Experiments

Sparse model are trained for 300 epochs rather than the 100 used at full annotation. Fewer annotators per step means fewer gradient updates per epoch that carry meaningful segmentation signal; 300 epochs compensate for this reduced information density. All other hyperparameters (learning rate, cosine schedule, batch size, latent dimension) remain identical to the full-annotation setting.

Evaluation is always performed with all four annotators' masks, regardless of the training sparsity level. This ensures that sparse-training GED values are measured on the same distribution as full-annotation GED values and can be compared on the same scale. Results at $\text{np} = 4$ with 300-epoch sparse training are excluded from the main comparisons: the shared baseline degrades at 300 epochs with full annotation (winning only one of four folds), which is a training-artefact confound introduced by the longer schedule, not a genuine phenomenon of the per-rater design. We compare

the 100-epoch full-annotation models as the reference and the 300-epoch sparse models at np $\in \{1,2,3\}$ as the sparsity conditions.

3.6 Gradient Alignment Measurement

The gradient conflict hypothesis predicts that as annotators are removed during training, the gradients at the shared $\mathbf{z}$ become more aligned. The mechanism is zero-channel dominance: absent annotators each contribute an empty-mask gradient, and as their count grows these degenerate signals increasingly pull $\mathbf{z}$ in the same direction. To measure whether this collapse is detectable, we compute a scalar diagnostic index for each fold.

3.6.1 Measurement Procedure

For each test image $\mathbf{x}$ with four rater masks $\{y_i\}_{i=1}^{4}$, we perform the following:

1. Forward pass through the trained baseline to obtain the shared posterior mean $\bar{\mathbf{z}} = \mu_q(\mathbf{x}, Y)$.
2. For each rater i, compute the per-rater reconstruction gradient: $\mathbf{g}_i = \partial \log p(y_i \mid \bar{\mathbf{z}}, \mathbf{x}) / \partial \bar{\mathbf{z}}$.
3. Compute pairwise cosine similarity across all $\binom{4}{2} = 6$ pairs: $\cos(\mathbf{g}_i, \mathbf{g}_j) = (\mathbf{g}_i \cdot \mathbf{g}_j)/(\|\mathbf{g}_i\|\|\mathbf{g}_j\|)$.
4. Average across pairs and across all test cases in the fold (100 cases sampled per fold).

The resulting scalar summarises how aligned the per-rater reconstruction gradients are at the shared posterior mean. A value near 0 indicates independent gradients pointing in different directions in the 6-dimensional latent space. A value near 1 indicates all gradients have collapsed to approximately the same direction. The latent code receives the same update signal regardless of which annotator it tries to represent.

3.6.2 Why Per-Rater Alignment Is Zero by Construction

There is no shared posterior mean in the per-rater model. Each encoder has its own $\mathbf{z}_i$, and by Proposition 3.4.3, $\partial \mathcal{L}_i / \partial \mathbf{z}_j = 0$ for $i \neq j$. The concept of cross-rater gradient alignment at a shared latent point does not exist in the per-rater architecture. The measured alignment for per-rater posteriors is 0.000 at all sparsity levels, by construction rather than by measurement.

3.6.3 What the Diagnostic Measures, and What It Does Not

The alignment index is a diagnostic: it measures whether degeneration has occurred, not why. The mechanism is zero-mask channel dominance in the shared encoder (§3.5.2). As fewer annotators are present, zero channels constitute a larger fraction of the 5-channel input, and their gradients increasingly dominate the update at $\mathbf{z}$. When $\text{np} = 1$, the single present annotator provides one gradient, and the three absent annotators each contribute a gradient directed toward predicting an empty mask. Three of the four gradient directions point the same way: all pushing the latent code toward predicting empty masks for absent annotators. The one real gradient from the present rater is outnumbered three to one. The pairwise cosine similarity across all six pairs therefore rises sharply, reaching 0.976 at $\text{np} = 1$.

The alignment rise and the GED gap both grow monotonically as np decreases. This co-movement is not coincidental; both are downstream of same mechanism. The alignment is a symptom of gradient collapse, and the GED degradation is a consequence of the same collapse degrading latent space quality — two observed effects of zero-mask channel contamination rather than a direct causal relationship between the two metrics. Stating that "gradient collapse causes higher GED" would therefore overstate what the diagnostic can show.

3.7 Attribute Characterisation of Annotation Disagreement

The final methodological component is an analysis that is independent of our model entirely. LIDC-IDRI provides not only segmentation masks but also nine per-rater per-nodule attribute ratings: malignancy, texture, spiculation, lobulation, margin, sphericity, calcification, internalStructure, and subtlety, each on an integer scale. We use these ratings to ask which nodule attributes most strongly predict how much the four radiologists will disagree about a nodule's boundary.

The analysis uses only the LIDC-IDRI ground-truth annotations; neither model is involved. Clinically, this identifies the nodule types where annotation uncertainty is structurally highest and uncertainty-aware methods are most relevant.

3.7.1 Setup

For each nodule case k, we compute the inter-rater attribute standard deviation across the four radiologists' ratings, $\sigma_{\text{attr}}^{(k)}$, and the inter-rater mask variance $\sigma_{\text{mask}}^{(k)}$ (variance in binary segmentation decisions across the four masks). We then compute the Pearson correlation coefficient r between these two quantities across all cases with complete attribute records.

Out of the 1,609 LIDC-IDRI nodule patches used in the main experiments, 1,603 have complete attribute data for all nine attributes and all four annotators. The remaining 6 cases are excluded from this analysis. The correlation is computed independently within each of the four cross-validation folds to verify that the result is not a data-split artefact.

3.7.2 What Was Attempted and Dropped

Two further analyses were attempted during development and both had to be abandoned.

Analysis A (GED by nodule margin quartile) stratified the test-set GED comparison by the inter-rater margin disagreement quartile, aiming to show that per-rater posteriors improve most on the most ambiguous nodules. This analysis was dropped for two reasons. First, the third quartile (Q3, moderately ambiguous nodules) showed a -5.0% reversal where the baseline outperformed per-rater, directly contradicting the intended narrative. Second, the GED formula in the stratification script used mean squared error rather than the IoU-based formulation of Equation 3.1. The computed values (0.001–0.004) are incomparable with the main GED results (0.14–0.22). The analysis is not included anywhere in this thesis.

Analysis C attempted to demonstrate that the GED improvement from per-rater posteriors is concentrated on clinically ambiguous nodules, specifically that the per-rater advantage correlates with per-attribute inter-rater disagreement. The Pearson r between per-rater GED improvement and inter-rater attribute disagreement was: subtlety $r = -0.042$ ($p = 0.093$), malignancy $r = -0.023$ ($p = 0.356$), spiculation $r = 0.035$ ($p = 0.157$). None of these is significant at $p < 0.05$. The claim has no statistical support with the available data and is reported in the Limitations section (§4.7). It is not treated as a finding.

3.8 Summary of Design Decisions

Table 3.2 captures every training-time difference between the two architectures. Both models use an identical inference procedure — sampling from $p(\mathbf{z} \mid \mathbf{x})$ and decoding — so any performance difference in Chapter 4 traces back to the training-time change rather than to differences in sampling or decoding strategy.

Table 3.2 Design decisions: shared baseline vs. per-rater.

Design aspect	**Shared baseline**	**Per-rater (proposed)**
Posterior encoders	1 shared encoder	$N = 4$ independent encoders
Encoder input	$N + 1 = 5$ channels	2 channels (image + 1 mask)
Training objective	Single joint ELBO	Sum of N per-rater ELBOs
Stage 2	Used	Not used
Absent rater handling	Zero-channel input	Encoder not executed
Gradient at posterior	Averaged over N raters	Pure single-rater signal
Inference	Sample from $p(\mathbf{z}\|\mathbf{x})$	Identical
Training time	$\sim$10.5 hours (4 folds)	$\sim$47 hours (4 folds)

The proposed method makes one change to the training procedure, giving each annotator a dedicated posterior encoder. Everything else (prior, decoder, loss weighting, hyperparameters, inference) is held constant. This isolation allows any performance difference in Chapter 4 to be attributed to the per-rater encoder design rather than to differences in optimiser, backbone, or loss function.

CHAPTER 4

RESULTS AND DISCUSSION

Full-annotation comparisons and the ablation are in §4.2; the sparsity results, which show the larger performance gap, are in §4.3. §4.5 reports the attribute correlation analysis, which uses only LIDC-IDRI ground-truth masks and attribute ratings and is independent of both segmentation models.

4.1 Experimental Setup

4.1.1 Datasets

LIDC-IDRI [1]: Dataset details and preprocessing are described in §2.6. Briefly: 1,609 nodule patches across four cross-validation folds (test sizes 450, 375, 412, 372), each patch carrying four independent radiologist masks and nine per-rater clinical attribute ratings. All experiments are trained and evaluated independently on each fold; results are 4-fold means $\pm$ standard deviation unless stated otherwise.

NPC-170 [17]: 170 nasopharyngeal carcinoma MRI cases, each annotated by four annotators. The input is 3-channel (T1, T1CE, T2 MRI). Training uses 2,405 slices, with 20 validation and 20 test cases on a single train/test split. Three random seeds are used to account for split variance. The NPC-170 experiment is a cross-dataset check on whether the per-rater approach transfers to a different modality and anatomy.

4.1.2 Evaluation Metrics

Generalised Energy Distance (GED $\downarrow$) [11]: defined formally in Equation 3.1. Rewards predictions that are both geometrically close to individual rater masks and spread across the range those raters cover; lower is better. The formal properties and training implications are discussed in §3.1.

Dice$_{\text{match}}$ ($\uparrow$) [17]: Hungarian-matched assignment of model predictions to rater an-

notations, averaged over the four raters. Measures geometric closeness to individual annotator masks.

Dice$_{\text{soft}}$ (↑): Soft Dice between the mean model prediction and the mean rater annotation. Measures average accuracy without accounting for per-rater variation.

4.1.3 Baselines

ProbUNet [11]: trained on the pixel-wise mean of all four rater annotations as a single target. This sets the lower bound: collapsing four annotations into one mean target discards all per-rater structure, and any multi-rater method that fails to beat it substantially has gained nothing from using multiple annotations.

D-Persona Stage 1+2 [17]: the full two-stage pipeline: shared posterior Stage 1 followed by per-rater projection head Stage 2. This is the primary baseline.

Proposed (per-rater, Stage 1 only): four independent posterior encoders, per-rater ELBO, no Stage 2. Stage 1 only.

For the sparsity experiments, the comparison collapses to shared posterior (Stage 1 baseline trained with `drop_raters()`) versus per-rater Stage 1. Stage 2 is not applicable in sparse settings where annotators are routinely absent during training.

4.1.4 Implementation Details

All hyperparameters match D-Persona's published settings to ensure fair comparison: ResNet34 backbone [4], Adam optimiser [8] with learning rate 10^{-4}, cosine annealing, batch size 12, latent dimension $D = 6$. Full-annotation models are trained for 100 epochs per fold. Sparse-annotation models are trained for 300 epochs to compensate for the reduced information density at each step. All experiments run on Apple MPS hardware; the full configuration is given in Table 4.1.

Table 4.1 Hyperparameter configuration for all experiments.

Setting	Value
Backbone	ResNet34
Optimiser	Adam, lr $= 10^{-4}$
LR schedule	Cosine annealing
Latent dimension D	6
Batch size	12
β (KL weight)	0.5
Epochs (full annotation)	100 per fold
Epochs (sparse annotation)	300 per fold
Number of annotators N	4
Training hardware	Apple MPS

4.2 Full-Annotation Performance

4.2.1 Main Comparison on LIDC-IDRI

Table 4.2 lists GED, $\text{Dice}_{\text{match}}$, and $\text{Dice}_{\text{soft}}$ for all three methods. GED and $\text{Dice}_{\text{match}}$ both improve in the per-rater model while $\text{Dice}_{\text{soft}}$ holds at exactly 0.9015, unchanged from D-Persona's value. Diversity and per-rater accuracy improve together, with no loss in average prediction quality.

Table 4.2 LIDC-IDRI segmentation results, 4-fold CV.

Method	GED ↓	$\text{Dice}_{\text{match}}$ ↑	$\text{Dice}_{\text{soft}}$ ↑
ProbUNet [11]	$0.2234_{\pm.0211}$	$0.8836_{\pm.0111}$	$0.8827_{\pm.0135}$
D-Persona (S1+S2) [17]	$0.1507_{\pm.0088}$	$0.8909_{\pm.0037}$	$0.9015_{\pm.0039}$
Ours (per-rater, S1)	$\mathbf{0.1444}_{\pm.0141}$	$\mathbf{0.9112}_{\pm.0061}$	$0.9015_{\pm.0066}$

GED falls from 0.1507 ± 0.0088 (D-Persona) to 0.1444 ± 0.0141 (per-rater), a 4.2% reduction. $\text{Dice}_{\text{match}}$ rises from 0.8909 to 0.9112, a $+2.28\%$ gain. Both metrics move in the right direction simultaneously. Models that improve diversity by spreading predictions more widely usually pay for it with worse per-rater geometric accuracy; per-rater posteriors move both metrics in the same direction, which points to improved gradient signal quality rather than a redistribution of the diversity-accuracy budget.

$\text{Dice}_{\text{soft}}$ sits at exactly 0.9015 for both D-Persona and the per-rater model, meaning average prediction quality against the mean annotation is unchanged. The gain is in per-rater calibration: the model approximates individual annotator boundary decisions better, without any shift in centrist prediction.

The per-rater Stage 1 alone outperforms the full D-Persona Stage 1+2

pipeline. Stage 2 adds roughly eight hours of training across four folds and makes GED worse.

ProbUNet (0.2234 GED, 0.8836 $\text{Dice}_{\text{match}}$, 0.8827 $\text{Dice}_{\text{soft}}$) is the lower bound — training on the mean of four masks and generating near-identical samples. The gap between ProbUNet and both D-Persona variants confirms that multi-rater probabilistic training adds real value.

The per-rater model's cross-fold $\text{Dice}_{\text{match}}$ variance (± 0.0061) is wider than D-Persona's (± 0.0037). Some folds contain more nodules with ill-defined margins where gradient conflict is severe and per-rater encoders help more; other folds are more homogeneous. The wider variance tracks this fold-level heterogeneity rather than any instability in the training.

4.2.2 Per-Expert Dice Breakdown

Table 4.3 decomposes the $\text{Dice}_{\text{match}}$ improvement by individual annotator, averaged across all four folds.

Table 4.3 Per-expert $\text{Dice}_{\text{match}}$ breakdown on LIDC-IDRI, 4-fold average.

Method	Expert 1	Expert 2	Expert 3	Expert 4
D-Persona (S1+S2)	0.8816	0.8947	0.8986	0.8888
Ours (per-rater, S1)	0.9062	0.9147	0.9182	0.9058
Improvement	+0.0246	+0.0200	+0.0196	+0.0170

The gain ranges from $+0.0170$ (Expert 4) to $+0.0246$ (Expert 1) across all four annotators; no single rater drives the aggregate while others regress. Each radiologist's predictions are better approximated under per-rater training than under the shared baseline.

Expert 1 gains the most ($+0.0246$). Their boundary style is better captured when their gradients reach the encoder without being pooled with three other raters. Expert 4 shows the smallest gain at $+0.0170$ but still improves, and no single annotator drives the aggregate at the expense of the others.

4.2.3 Ablation Study

Table 4.4 evaluates the D-Persona baseline against six independent design modifications. Each row is a separate standalone experiment, not a cumulative stack. All ablation results use 4-fold cross-validation on LIDC-IDRI.

Table 4.4 Ablation study on LIDC-IDRI, 4-fold CV.

Method	**GED** ↓	**Dice$_{\text{match}}$** ↑
(1) D-Persona baseline (S1+S2)	$0.1507_{\pm.0088}$	$0.8909_{\pm.0037}$
(2) + MiT-B2 backbone	$0.1531_{\pm.0127}$	$0.8909_{\pm.0068}$
(3) + Orthogonality loss	$0.1519_{\pm.0070}$	$0.8890_{\pm.0033}$
(4) + Prior bank ($k = 100$)	$0.2212_{\pm.0055}$	$0.8816_{\pm.0016}$
(5) + Dual diversity loss	$0.1509_{\pm.0070}$	$0.8889_{\pm.0032}$
(6) **Per-rater posteriors (S1)**	$\mathbf{0.1444}_{\pm.0141}$	$\mathbf{0.9112}_{\pm.0061}$
(7) Per-rater + Style vectors (S1+S2)	$0.1836_{\pm.0108}$	$0.8876_{\pm.0049}$

Rows 2 through 5 represent four independent attempts to improve over the baseline before arriving at the per-rater formulation.

Substituting MiT-B2 for ResNet34 (Row 2) tests the architectural capacity hypothesis directly. The attention mechanism in MiT-B2 can relate any two spatial locations regardless of distance — considerably more expressive than a convolutional backbone. GED comes out at 0.1531 versus the baseline's 0.1507, with Dice$_{\text{match}}$ unchanged. A more expressive transformer learns a more expressive average of conflicting rater gradients; the conflict itself is a property of which signals reach the encoder, not how much capacity the encoder has.

Row 3, Orthogonality loss: An explicit loss term penalising cosine similarity between per-rater posteriors (intended to push the shared encoder toward more orthogonal rater representations) marginally worsens both metrics (0.1519 GED, 0.8890 Dice$_{\text{match}}$). The gradient averaging inside the shared encoder happens before the orthogonality loss ever sees the encoder's output — the loss is applied to a representation already shaped by conflicting gradients, so the per-rater structure it is trying to enforce was never encoded in the first place.

Row 4, Prior bank ($k = 100$): Replacing the continuous prior with $k = 100$ discrete k-means centroids produces the worst result in the table. GED reaches 0.2212, worse than ProbUNet. GED rewards diversity from stochastic prior sampling; replacing that with deterministic prototype retrieval eliminates the sample variance the metric is measuring.

Row 5 adds a dual diversity loss penalising pairwise Dice similarity between prior samples, reaching GED 0.1509 — essentially tying the baseline. Diversity pressure applied to the prior cannot create rater-specific structure that Stage 1's gradient averaging never encoded in the first place.

Across all four alternatives, none changes how gradients reach the posterior encoder. That is the consistent reason for failure. Per-rater posteriors (Row 6) are the only modification that improves both GED and Dice$_{\text{match}}$ simultaneously: 0.1444 GED and 0.9112 Dice$_{\text{match}}$ with Stage 1 only.

Row 7 resolves the question of whether Stage 2 can be layered on top of a per-rater Stage 1. GED rises from 0.1444 to 0.1836 — a 27% increase. Stage 2 was designed to personalise a shared latent code that lacks rater-specific structure. Applied after per-rater Stage 1, where each annotator already has a distinct gradient pathway, it introduces shared-posterior dynamics on top of a model that does not need them and collapses the diversity that Stage 1 created.

4.2.4 Cross-Dataset Validation on NPC-170

On NPC-170 the GED gap is 0.0011, which falls within the seed variance of ±0.0085. Table 4.5 reports the full results.

Table 4.5 Cross-dataset validation on NPC-170, 3-seed mean ± std.

Method	**GED** ↓	**Dice$_{\text{match}}$** ↑
Baseline (shared posterior, S1)	0.1824	0.8249
Ours (per-rater, S1)	$0.1813_{\pm.0085}$	$0.8225_{\pm.0008}$

On NPC-170, per-rater posteriors and the shared baseline perform at essentially the same level: GED 0.1813 versus 0.1824, a gap of 0.0011 against a seed variance of ±0.0085.

Two possible explanations are worth considering but cannot be confirmed without further experiments. First, NPC-170 has a single train/test split rather than four-fold cross-validation; the lack of independent validation folds means the baseline GED (0.1824) is itself a point estimate with unknown variance. Any small gap is therefore uninterpretable. Second, NPC-170 has 2,405 training slices across 170 cases (a smaller dataset relative to the complexity of 3-channel MRI input); the per-rater encoders may need more training cases to distinguish four annotators' styles on 3-channel MRI.

What the NPC-170 result does establish is that per-rater posteriors do not degrade on different modality and anatomy. The method transfers without special adaptation to 3-channel MRI and a different disease site, and does not require LIDC-specific tuning.

4.2.5 Qualitative Analysis

The diversity gap is most visible at maximum rater disagreement; two fold 3 cases with the widest annotation spread are shown in Figure 4.1 (four prior samples each). ProbUNet's four samples cluster near-identically in both cases. The model has learned a single average representation and sampling from the prior produces negligible variation (despite 100 prior samples being available), so the within-model distance term in GED

approaches zero and ProbUNet is penalised heavily regardless of how accurate each individual sample is.

Our model's four samples span the range visible across the four rater masks. Where radiologists differ on whether the ground-glass halo around the nodule core should be included, some of our samples include it and some do not. The samples match the annotation spread. The prior is calibrated to produce this range by per-rater training signals.

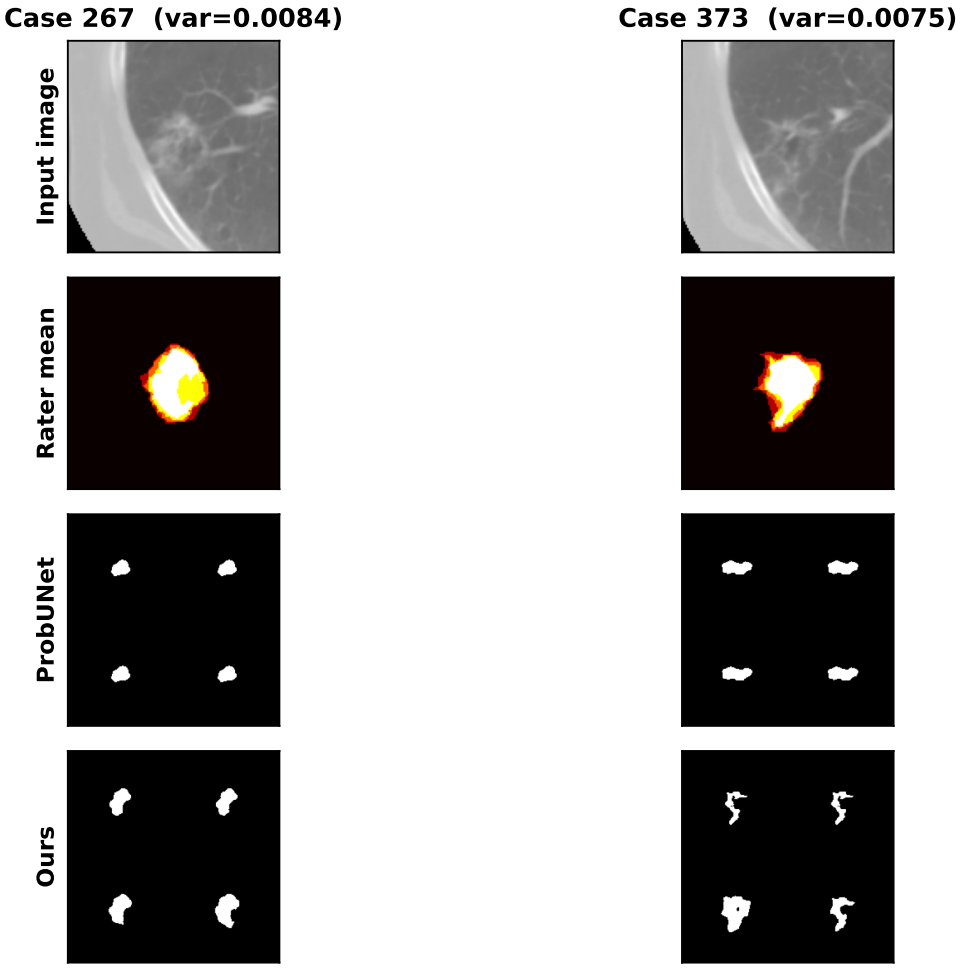

Figure 4.1 Qualitative comparison, high-disagreement LIDC-IDRI fold 3 cases.

Figure 4.2 shows all four posterior clouds and the prior mean overlapping in the t-SNE projection of 412 fold 3 test cases — no distinct per-rater clusters form. Each encoder pushes the prior in a rater-specific direction within a shared latent region rather than separating the modes. Stage 2 personalisation depends on per-rater modes being geometrically separated enough to steer from; that separation is absent here, which explains the Row 7 GED degradation.

The FComb decoder recovers rater-specific boundary decisions from directional differences between encoder signals, without the modes being globally separated

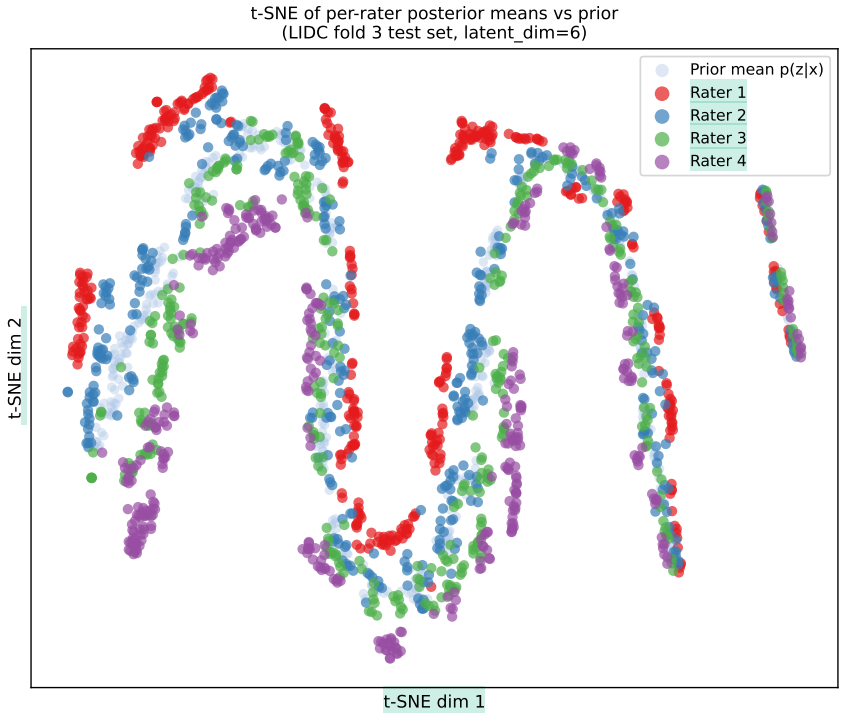

Figure 4.2 t-SNE of posterior and prior means, fold 3.

in the 6-dimensional space. Collapsing 6 dimensions to 2 for t-SNE discards these directional differences while preserving global structure, so the overlap in Figure 4.2 is consistent with — not a contradiction of — the GED gain. The improvement comes from what the training gradients encode, not from where the posterior means land.

4.3 Annotation Sparsity Robustness

At full annotation, per-rater posteriors improve GED and $\text{Dice}_{\text{match}}$ by 4.2% and 2.28% respectively — real but modest. Both models were then retrained at np $\in \{1, 2, 3\}$ annotators per training step, always evaluated on the full four-rater annotation set, to test how each degrades when coverage is incomplete.

4.3.1 GED Under Sparse Annotation

As coverage drops from four raters to one, the shared baseline loses more than half its GED performance while the per-rater model degrades far less severely; Table 4.6 quantifies this divergence (reference row: 100-epoch models; sparse rows: 300-epoch models). Two features of this table require explicit note. First, the baseline here is D-Persona Stage 1 only (GED 0.1436), not the Stage 1+2 pipeline of Table 4.2 (GED 0.1507): Stage 2 is omitted because it cannot be applied when annotators are routinely

absent during training. Second, the per-rater full-annotation GED (0.1429) differs slightly from the main-table value (0.1444) because these models were trained under the sparse experimental protocol and the main experiments were independent training runs; both are 100-epoch per-rater Stage 1 models, and the difference is within the cross-run variance expected from random initialisation.

Table 4.6 Per-rater vs. baseline GED across annotation coverage levels.

Annotators present	Baseline GED	Per-rater GED	Improvement
Full (np=4, reference)	$0.1436_{\pm.0076}$	$0.1429_{\pm.0143}$	+0.5% (n.s.)
np=3 (3 annotators)	$0.1810_{\pm.0175}$	$0.1601_{\pm.0090}$	**+11.5%**
np=2 (2 annotators)	$0.2039_{\pm.0239}$	$0.1677_{\pm.0144}$	**+17.8%**
np=1 (1 annotator)	$0.2220_{\pm.0093}$	$0.1745_{\pm.0119}$	**+21.4%**

The gap is monotonic: +11.5%, +17.8%, +21.4% as annotators drop from three to two to one, with no reversal at any level. More telling is the fold-by-fold record: across all 12 individual comparisons (three sparsity levels × four folds), not one favours the shared baseline. Under the null hypothesis of no systematic difference, the probability of that outcome is $(1/2)^{12} \approx 0.024\%$. The 12 comparisons are not fully independent (the three sparsity levels within each fold share the same test cases), but even under conservative adjustment the one-sided probability remains well below 0.01. At full annotation, the +0.5% gap is noise: the 4-fold SD for the baseline alone is ±0.0076, which swamps a difference that small. The per-rater advantage is concentrated in the annotation sparsity regime, where the shared posterior's gradient signal degrades.

The baseline GED degrades sharply under sparsity: from 0.1436 at full annotation to 0.2220 at one annotator (a 54.6% increase). Per-rater GED rises from 0.1429 to 0.1745 (a 22.1% increase). The per-rater model degrades under sparsity too, as it should when fewer annotators are providing training signal, but it degrades far less severely. The shared baseline loses 54.6% of its full-annotation GED performance at maximum sparsity; per-rater posteriors lose 22.1%. The degradation rate differs by 2.5× and widens at every step down in coverage, as Figure 4.3 shows.

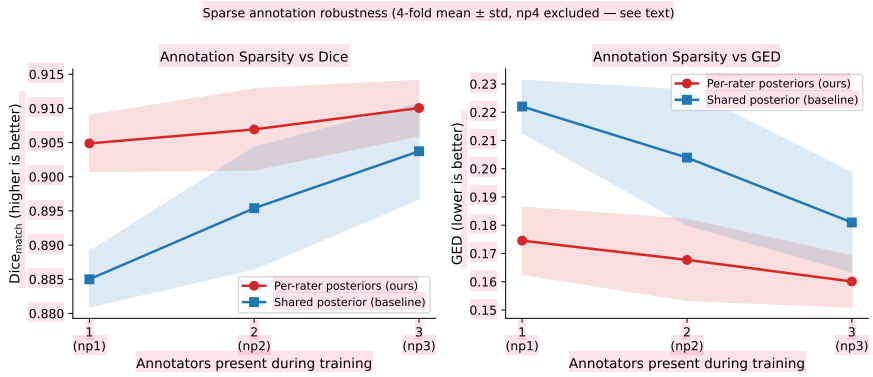

Figure 4.3 GED vs. annotators present per training image.

4.3.2 $\text{Dice}_{\text{match}}$ Under Sparse Annotation

GED penalises diversity loss while $\text{Dice}_{\text{match}}$ measures geometric closeness to individual annotator boundaries. When both metrics move in the same direction across every sparsity level in Table 4.7, a metric-specific artefact is ruled out.

Table 4.7 Per-rater vs. baseline $\text{Dice}_{\text{match}}$ across annotation coverage levels.

Annotators present	Baseline $\text{Dice}_{\text{match}}$	Per-rater $\text{Dice}_{\text{match}}$	Δ Dice
Full (np=4, reference)	$0.9126_{\pm.0033}$	$0.9130_{\pm.0060}$	+0.0004
np=3	$0.9037_{\pm.0070}$	$0.9101_{\pm.0041}$	+0.0064
np=2	$0.8954_{\pm.0089}$	$0.9069_{\pm.0059}$	+0.0115
np=1	$0.8850_{\pm.0041}$	$0.9049_{\pm.0041}$	+0.0199

$\text{Dice}_{\text{match}}$ shows the same monotonic pattern: +0.0004 at full annotation, rising to +0.0199 at one annotator. Both metrics agree on the direction at every sparsity level.

4.3.3 Fold 1 as the Sharpest Individual Evidence

Fold 1 is the only fold where the shared baseline outperforms per-rater posteriors at full annotation: baseline GED 0.1552 versus per-rater 0.1658, the baseline winning by 6.8%. This makes it the sharpest individual test of whether the advantage is structural.

At $\text{np} = 1$ in the same fold, on the same test cases, the picture reverses completely: baseline GED 0.2323 versus per-rater 0.1806. Per-rater wins by 22.3% $((0.2323 - 0.1806)/0.2323)$.

The model that is worse in Fold 1 at full annotation wins by 22.3% at maximum sparsity, on identical test cases. A general quality difference cannot explain that reversal. The reversal is tied specifically to the sparsity condition and is consistent with the mechanism described in §3.6: under $\text{np} = 1$, the shared baseline receives three zero-channel gradients for every one annotator gradient, collapsing the shared $\mathbf{z}$ toward empty-mask predictions. Per-rater posteriors avoid this entirely. At $\text{np} = 1$, the shared baseline's four prior samples converge to near-identical empty masks while the per-rater model's samples retain the spread visible across the four rater annotations, as Figure 4.4 shows side by side.

4.4 Gradient Alignment Analysis

The sparsity results establish that the shared baseline degrades far more than per-rater posteriors as annotation coverage decreases. The gradient alignment index (mean

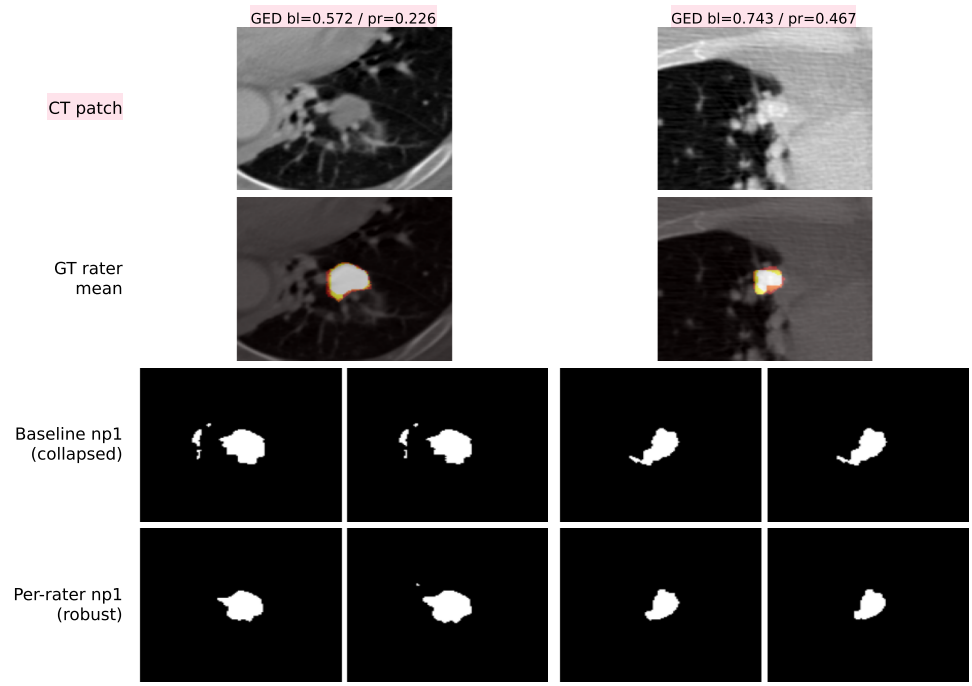

Figure 4.4 Qualitative comparison, single-annotator training ($n_p = 1$).

pairwise cosine similarity of per-rater reconstruction gradients at the shared latent code) turns the gradient conflict framework of Yu et al. [19] into a measurable scalar in the annotation sparsity setting. It examines whether the degeneration in the shared latent space is measurable, and whether it tracks the GED gap monotonically.

4.4.1 Results at Full Annotation

At full annotation, the mean pairwise cosine similarity of per-rater reconstruction gradients at the shared posterior mean $\bar{\mathbf{z}}$ is 0.167 across folds (fold-level standard deviation ± 0.026). The baseline alignment is positive and significantly different from zero in all four folds.

The within-fold distribution is highly heterogeneous: within a single fold's test set, the case-level standard deviation of alignment scores is approximately 0.439 (nearly three times the mean). Alignment ranges from strongly negative (opposing gradients, raters completely disagree on boundary location) to strongly positive (raters mostly agree). Approximately 62% of cases show positive alignment, 38% negative, and 17% exceed 0.5. Gradient conflict at full annotation is real but case-dependent, concentrated on specific nodules with ill-defined margins rather than universally presented.

The modest 4.2% GED improvement from per-rater posteriors at full annotation is consistent with gradient conflict being a significant problem only for a fraction of training cases.

4.4.2 Gradient Collapse Under Sparse Annotation

The alignment index rises in step with the GED gap, reaching near-unity at single-annotator coverage. Table 4.8 captures this collapse numerically.

Table 4.8 Gradient alignment index under sparse annotation.

Annotators	Baseline align.	SD (across folds)	Within-fold SD	Per-rater
Full (np=4)	0.167	0.026	≈0.439	0.000
np=3	0.291	0.074	—	0.000
np=2	0.463	0.039	—	0.000
np=1	0.976	0.012	≈0.023	0.000

The alignment index rises monotonically from 0.167 (full annotation) through 0.291 (three annotators) and 0.463 (two annotators) to 0.976 (one annotator).

The collapse at $\text{np} = 1$ is qualitatively different from the full-annotation condition, not a higher mean but a categorical shift in regime. At full annotation, the case-level within-fold standard deviation is approximately 0.439: gradient conflict varies substantially from nodule to nodule. At $\text{np} = 1$, the within-fold standard deviation drops to approximately 0.023 (an approximately 19-fold reduction; $0.439/0.023 = 19.1$), meaning almost every test nodule now shows near-complete gradient alignment regardless of how ambiguous its margin actually is. The shared posterior has entered a degenerate mode that is no longer sensitive to image content.

The fold-level consistency at $\text{np} = 1$ is also tight: standard deviation across folds is only 0.012, so the collapse is a property of the training regime rather than of one particular data split.

Per-rater alignment stays at 0.000 across all sparsity level. The zero value is expected: the architecture has no shared $\mathbf{z}$, so cross-rater gradient alignment at a single latent point is undefined.

4.4.3 Mechanism and Interpretation

The alignment data makes the zero-mask mechanism concrete. At $\text{np} = 1$, a 5-channel encoder receives three zero-mask channels and one real annotation channel simultaneously. The three empty-mask gradient directions are approximately collinear (all pushing $\mathbf{z}$ toward the same degenerate empty-segmentation prediction), while the one real gradient pushes in a structurally different direction. With three collinear gradients against one, mean pairwise cosine similarity is driven toward 1.0; the single real annotation gradient cannot break the degeneracy. Each step down in np increases the fraction of the training signal devoted to empty-mask pull, and the alignment index rises accordingly.

The alignment index is a diagnostic that measures the consequence of this collapse. The alignment index and the GED gap rise together, but one does not cause the other. Both trace back to zero-mask channels dominating the shared encoder's gradient as annotators are removed. Both the GED gap and the alignment index track the severity of that contamination, which is why they rise together as np decreases.

Figure 4.5 (GED gap and alignment on dual y-axes) shows this co-movement across the four annotation levels. The curves are not identical in shape (the alignment rises faster), but both are monotonically increasing in the same direction, reaching their peak at np = 1.

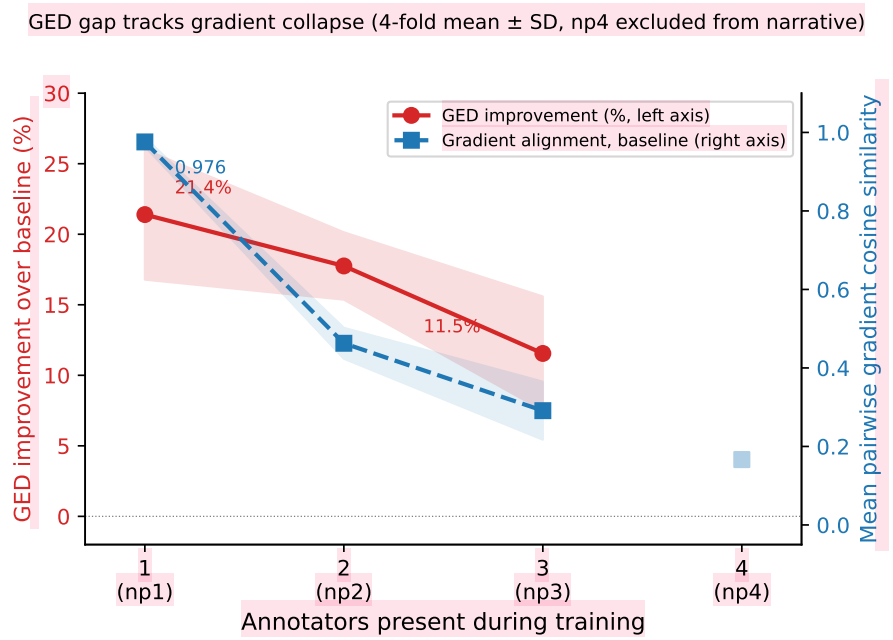

Figure 4.5 GED improvement and gradient alignment vs. annotators present.

4.5 Attribute Characterisation of Annotation Disagreement

The analysis in this section asks which nodule properties are associated with high inter-rater annotation disagreement in LIDC-IDRI [2]. It uses only the ground-truth masks and the nine per-rater clinical attribute ratings, and is independent of both segmentation models.

4.5.1 Results

Of the nine LIDC-IDRI clinical attributes, six reach statistical significance; Table 4.9 ranks them by effect size, separating the positive drivers of boundary ambiguity from the negative ones.

Table 4.9 Pearson r: attribute disagreement vs. inter-rater mask variance.

Attribute	**Pearson r**	**p-value**
Margin	$+0.318$	< 0.001
Lobulation	$+0.243$	< 0.001
Texture	$+0.210$	< 0.001
Spiculation	$+0.185$	< 0.001
Malignancy	-0.202	< 0.001
Subtlety	-0.155	< 0.001

Margin is the strongest predictor: $r = 0.318$, confirmed independently across all four folds (per-fold range: 0.238–0.400, all $p < 0.001$). Nodules with ill-defined, irregular margins cause the most inter-rater boundary disagreement. This is clinically expected: a sharp, well-defined margin leaves little room for interpretation; a gradual tissue transition at the nodule edge creates genuine ambiguity about where the nodule ends.

Lobulation ($r = 0.243$) and texture ($r = 0.210$) follow. Part-solid nodules (those with both solid and ground-glass opacity components) have unclear density transitions, and lobulated shapes mean the boundary is non-convex and harder to delineate consistently.

The two negative correlations are the more interesting finding. Malignancy ($r = -0.202$) and subtlety ($r = -0.155$) are negatively associated with mask variance. Higher inter-rater disagreement on a nodule's perceived malignancy does not predict higher mask variance; if anything, it slightly predicts lower mask variance. A nodule can be perceived as highly malignant while having a geometrically clear boundary, and a nodule can be ambiguous in appearance (hard to spot) while still being easy to delineate once found.

This separation matters clinically. A clinician looking for where uncertainty-aware segmentation is most needed should prioritise ill-defined, lobulated, part-solid nodules rather than ones rated most suspicious for malignancy. The raw scatter in Figure 4.6 confirms the $r = 0.318$ trend is genuine and not driven by outliers: margin disagreement predicts mask variance across the full range of cases with no discontinuity.

4.5.2 Four-Fold Confirmation

The margin correlation is particularly important to verify across folds because the attribute analysis is applied to 1,603 cases without further splitting: computing it once on the full dataset could in principle reflect a data-split artefact. Computing it independently within each fold's test set (fold sizes 450, 375, 412, 372) gives margin r values of 0.238, 0.305, 0.400, and 0.359 respectively. All four are positive and

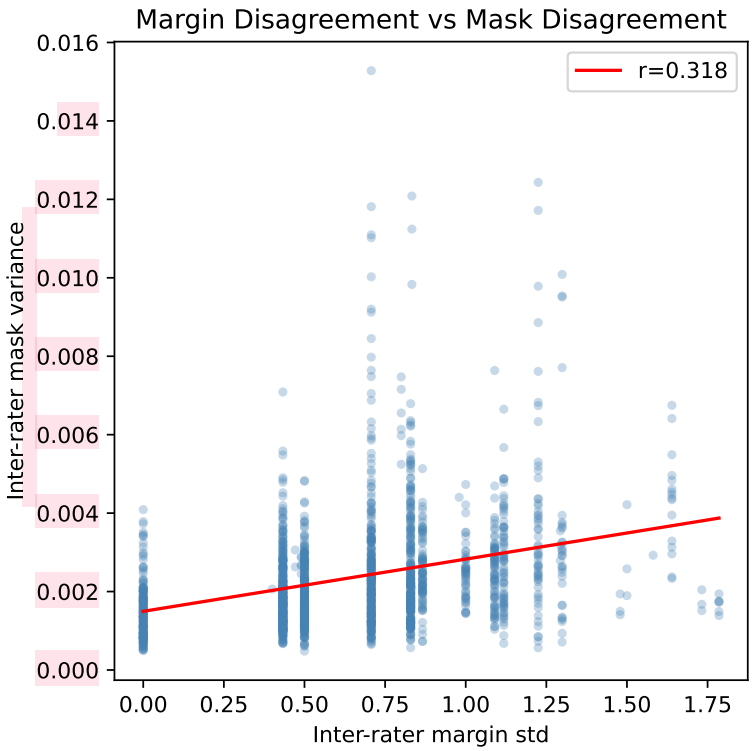

Figure 4.6 Margin disagreement vs. mask variance, LIDC-IDRI ($n = 1{,}603$).

significant. The result is stable across data splits, not an artefact of one particular test partition.

4.6 Discussion

4.6.1 Why Per-Rater Posteriors Work at Full Annotation

The t-SNE (Fig. 4.2) shows no distinct per-rater latent clusters; the four posterior means and the prior means overlap in the 6-dimensional latent space. Each encoder provides a directed gradient signal during training that the FComb decoder can distinguish from the other three, even when those signals point into the same latent region. The decoder learns rater-specific boundary responses from directional differences that t-SNE cannot reveal, since collapsing 6 dimensions to 2 discards the directional structure driving the effect.

The 0.5% full-annotation GED improvement reflects that gradient conflict at full annotation affects only a subset of training cases. Approximately 62% show positive alignment, 38% negative, and the within-fold standard deviation is 0.439. For cases where raters mostly agree, the shared and per-rater posteriors receive similar

training signals. The aggregate 4.2% GED gain comes disproportionately from the roughly 40% of cases where annotators diverge on the nodule boundary — the fraction where per-rater gradient isolation makes the biggest difference to what the shared decoder receives.

4.6.2 Why the Advantage Scales with Sparsity

At full annotation, gradient conflict is tied to specific nodules where raters disagree — it is bidirectional and case-dependent. Sparsity introduces a structurally different source of interference: zero-channel gradients from absent annotators all point the same way regardless of which nodule the encoder is processing. This directional uniformity is what makes sparse-regime degradation so much more severe than full-annotation conflict, and why the alignment index at $\text{np} = 1$ hits 0.976 rather than staying near its full-annotation value of 0.167.

As np decreases from 4 to 1, the proportion of the training signal coming from zero-channel gradients increases from 0 to 75%. The shared $\mathbf{z}$ has progressively less space to encode genuine annotation structure against this growing pressure. The gradient alignment index captures this, rising from 0.167 at full annotation (case-dependent conflict) to 0.976 at $\text{np} = 1$ (near-universal collapse). Per-rater posteriors are not affected by this mechanism at any sparsity level, since absent encoders do not run and contribute nothing to gradient update.

The Fold 1 reversal (§4.3.3) makes the sparsity-specificity of the advantage concrete at the individual fold level. A model that trails the shared baseline in a given fold at full annotation, then outperforms it by 22.3% in that same fold under single-annotator training, is responding to the sparsity condition specifically, not performing better in general.

4.6.3 Scope and Boundaries of the Results

On NPC-170, the GED gap falls within seed variance (§4.2.4); the method does not degrade on MRI nasopharyngeal carcinoma but does not improve either. The 0.5% full-annotation gap on LIDC-IDRI is similarly uninterpretable; the 4-fold SD for the baseline alone is ± 0.0076, which swamps a difference that small. On latent structure, the t-SNE shows no per-annotator clustering, which means the gain comes from gradient dynamics during training rather than from the model forming geometrically distinct modes for each annotator. The gradient alignment index is a diagnostic that characterises what happened to the shared encoder's training signal; it measures the consequence of zero-mask channel contamination rather than establishing a causal pathway from gradient collapse to GED degradation.

4.6.4 Clinical Implications

The benefit of per-rater posteriors concentrates in two specific deployment contexts.

First, datasets where not all radiologists annotate every image (which describes most large clinical collections). At one-annotator training coverage, the $+21.4\%$ GED improvement means the per-rater model preserves substantially more annotation diversity than the shared baseline. In a hospital setting where only one radiologist routinely annotates a given scan type, a model trained with per-rater posteriors remains calibrated to that radiologist's style without gradient contamination from absent annotators.

Second, the attribute analysis identifies where probabilistic methods matter most: ill-defined, lobulated, part-solid nodules. These nodules generate high inter-rater annotation disagreement by structural necessity, and a model calibrated to individual radiologist styles is most valuable for this group of cases. A deterministic segmentation model applied to well-defined, spiculated nodule does not lose much by ignoring annotation uncertainty. Applied to a nodule with an ill-defined margin (the attribute most correlated with inter-rater disagreement, $r = 0.318$), the absence of calibrated uncertainty is a real clinical problem.

4.7 Limitations

Dataset scope: All experiments use either LIDC-IDRI (CT, four radiologists, lung nodules) or NPC-170 (MRI, four annotators, nasopharyngeal carcinoma). Both datasets have $N = 4$ annotators by design. Generalisation to datasets with two annotators, ten annotators, or variable annotation counts per case has not been tested.

Training time: Per-rater Stage 1 training takes approximately 47 hours across all four LIDC-IDRI folds (4.5 times the shared baseline Stage 1 time of approximately 10.5 hours). In a setting with large-scale datasets or limited GPU access, this overhead is a practical barrier.

Epoch count confound in sparse experiments: Sparse models are trained for 300 epochs; full-annotation models for 100. The full-annotation reference row in Table 4.6 uses 100-epoch models. When sparse experiments are run at $\text{np} = 4$ (all annotators present) with 300-epoch training, the shared baseline degrades and wins only one of four folds (mean gap -3.1%). This is likely a training-artefact confound (longer training amplifies the shared posterior's degeneration under the full 5-channel input), but it cannot be ruled out as a genuine interaction without a controlled equal-epoch comparison at all sparsity levels.

Gradient alignment as diagnostic only: The alignment index is a diagnostic measurement computed on trained models. It characterises what happened to the shared $\mathbf{z}$

during training; causal proof that gradient conflict drove the GED gap would require a controlled intervention the experiments here cannot provide. An alternative explanation (that absent-rater information loss alone, without gradient mechanics, is responsible) cannot be definitively excluded.

Analysis C null result: We attempted to demonstrate that the per-rater GED improvement is concentrated on clinically ambiguous nodules by correlating per-rater improvement with per-attribute inter-rater disagreement. Pearson r between per-rater GED improvement and subtlety disagreement: $r = -0.042$ ($p = 0.093$); malignancy disagreement: $r = -0.023$ ($p = 0.356$); spiculation disagreement: $r = 0.035$ ($p = 0.157$). None is significant at $p < 0.05$. The claim that "per-rater posteriors help most in clinically ambiguous cases" is not supported statistically and is not made anywhere in this thesis.

Clinical validation: Neither LIDC-IDRI nor NPC-170 is a prospective clinical study. The claim that per-rater posteriors are useful in real deployment cannot be made without validation on prospectively collected multi-centre data with known radiologist agreement patterns.

CHAPTER 5

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

This thesis tested whether replacing the single shared posterior encoder with N independent per-rater encoders reduces gradient conflict enough to improve multi-rater segmentation in practice. At full annotation the GED advantage is 4.2%, real but modest, and at one annotator per training image it reaches 21.4%. The gap grows largest under annotation conditions most common in clinical practice, where complete multi-rater coverage is the exception rather than the rule.

5.1 Summary of Contributions

1. **Gradient isolation through per-rater posterior encoders.** D-Persona's Stage 1 shared encoder trains on all four annotators' masks simultaneously and receives conflicting gradient signals that partially cancel whenever raters disagree. The latent code encodes a compromise, and any downstream personalisation — Stage 2 projection heads, regularisation losses — operates on that already-averaged representation, working with gradient-cancelled residuals rather than the original per-rater signals.

   Four independent per-rater encoders — each receiving only the image and one annotator's mask — give the shared decoder four clean, rater-specific gradient signals during training. $\partial \mathcal{L}_i / \partial \mathbf{z}_j = 0$ for $i \neq j$ follows from the objective structure alone, with no regularisation term required. On LIDC-IDRI (1,609 nodule patches, 4-fold CV), the Stage 1-only per-rater model reaches GED 0.1444 ± 0.0141 and $\text{Dice}_{\text{match}}$ 0.9112 ± 0.0061 — a 4.2% GED reduction and 2.28% $\text{Dice}_{\text{match}}$ gain over the full D-Persona pipeline. All four per-rater Dice scores improve individually. $\text{Dice}_{\text{soft}}$ stays at 0.9015 for both models — the gain is in per-rater calibration, with average prediction quality unchanged.

2. **An ablation that rules out capacity, regularisation, and diversity loss as explanations.** Six modifications were tested against the D-Persona baseline. MiT-B2 adds parameters; GED is unchanged. An orthogonality loss worsens both metrics, because regularising the output of an encoder that trained on averaged gradients cannot undo the averaging that already occurred. The discretised prior bank ($k = 100$) produces the worst result in the table (GED 0.2212), removing

the stochastic sampling that GED rewards. Dual diversity loss ties the baseline.

Row 7 sharpens the interpretation: adding Stage 2 style vectors on top of per-rater Stage 1 raises GED by 27% (from 0.1444 to 0.1836). Stage 2 was designed to personalise a shared latent code. Applied after per-rater Stage 1 has already differentiated the gradient pathways, it imposes shared-posterior dynamics on a model that does not need them and collapses the diversity Stage 1 created.

3. **Sparsity experiments across three coverage levels, with a gradient collapse diagnostic.** At one annotator per training image, the shared baseline's 5-channel input carries three zero-mask channels. Three reconstruction gradients push the shared latent code toward empty-mask predictions; the single real annotation gradient competes against all three. The gradient alignment index — mean pairwise cosine similarity of per-rater gradients at the shared $\mathbf{z}$ — rises from 0.167 at full annotation to 0.976 at np $= 1$. The within-fold standard deviation drops 19-fold (from 0.439 to 0.023): the collapse is near-universal at maximum sparsity rather than case-dependent.

   Absent per-rater encoders do not execute and contribute nothing to the gradient update. GED advantage grows from $+11.5\%$ to $+21.4\%$ as coverage falls; all 12 per-fold comparisons at the three sparsity levels favour the per-rater model ($p < 0.001$, sign test). At full annotation the gap is 0.5% and within noise.

4. **Pearson correlation analysis linking nine nodule attributes to inter-rater mask variance.** Run on 1,603 LIDC-IDRI cases with complete attribute records, independent of both segmentation models. Margin clarity is the strongest predictor ($r = 0.318$, $p < 0.001$), confirmed across all four CV folds (r range: 0.238–0.400). Lobulation ($r = 0.243$) and texture ($r = 0.210$, capturing part-solid opacity) follow.

   The two negative correlations carry clinical weight: malignancy ($r = -0.202$) and subtlety ($r = -0.155$) are each negatively associated with mask variance. Boundary ambiguity and diagnostic severity are separable dimensions. A nodule perceived as highly malignant can have a geometrically clear edge, and a hard-to-detect nodule may be straightforward to delineate once found. Uncertainty-aware segmentation is most valuable for ill-defined, lobulated, part-solid nodules — a different population from those rated most suspicious for malignancy.

5.2 Overarching Finding

Under sparsity, absent annotators' zero-mask channels dominate the gradient at the shared latent code, and the model degrades accordingly. Stage 2 personalisation, a stronger backbone, and diversity regularisation all operate downstream of Stage 1 — after gradient averaging has already shaped the latent code. Stage 2 can only redistribute information already present in $\mathbf{z}$ — what gradient averaging removed before Stage 1's backward pass completed is gone from the latent code by the time Stage 2 runs. The orthogonality loss regularises the representation space after the gradient cancellation

has occurred. A larger backbone encodes a more expressive average of conflicting signals, with the conflict itself unaffected.

Mean pairwise cosine similarity at the shared code is 0.167 at full annotation — gradient conflict is real but concentrated on ambiguous nodules. At one-annotator coverage it reaches 0.976. The GED gap tracks the same trajectory: $+4.2\%$ at full annotation, $+11.5\%$, $+17.8\%$, and $+21.4\%$ as annotators drop to three, two, and one. Per-rater posteriors cut off the contamination at the encoder level, before it accumulates in $\mathbf{z}$, which is why the improvement scales with the contamination rather than being fixed.

5.3 Future Scope

Three open questions follow directly from the results.

1. **Pairwise repulsion between per-rater posterior means.** The t-SNE of the per-rater latent space (§4.2) shows that all four posterior means and the prior mean land in the same region of the 6-dimensional latent space, despite fully independent gradient pathways. The FComb decoder exploits directional differences between the per-rater signals — each encoder pushes in a rater-specific direction within a shared latent region. Stage 2 personalisation depends on per-rater modes being geometrically separated enough to project from, and those separations do not form.

   A repulsion term penalising proximity between μ_i and μ_j $(i \neq j)$ during Stage 1 training could push the modes apart while keeping per-encoder gradient isolation intact. If per-rater Stage 1 training produced distinct latent clusters, Stage 2 would have genuine rater-specific structure to steer from, and the Row 7 ablation result might reverse. Whether a repulsion term can achieve this without reintroducing gradient mixing remains untested.

2. **Attribute-conditioned training curricula.** The attribute analysis establishes that nodules with ill-defined margins, lobulated shapes, and part-solid texture account for the highest inter-rater annotation disagreement. Currently, all training images contribute equally to the per-rater ELBO, regardless of how much rater disagreement they carry. A curriculum that upweights high-disagreement cases (weighting reconstruction losses by inter-rater mask variance, or sampling high-disagreement cases more frequently in early epochs) could concentrate the model's learning capacity on the cases where boundary uncertainty is highest and calibrated probabilistic outputs matter most. The attribute correlation analysis provide a principled, clinically grounded basis for defining such a curriculum without requiring additional annotation effort.

3. **Scaling to variable annotator counts.** LIDC-IDRI has exactly four annotators per case by design. Real clinical datasets have variable annotator counts: some

images may have one expert annotation, others five or eight, depending on when and where they were collected. The per-rater formulation scales linearly in the number of encoders, but the training dynamics at $N = 2$ (where shared posterior gradient conflict is much less severe) and at $N = 8$ or more (where per-rater training cost grows substantially) have not been studied. Understanding how the gradient conflict advantage scales with N, and at what N the shared posterior becomes an adequate approximation, would establish the boundary conditions for when the per-rater design justifies its $4.5\times$ training cost overhead.

5.4 Social Impact

Per-rater posterior training was motivated by clinical settings where annotation resources are scarce and radiologist disagreement is structurally unavoidable. The results carry implications for deployment and for how annotation protocols are designed, alongside clear limits on what the experiments do and do not establish.

5.4.1 Direct Impact

The most concrete implication of the sparsity results concerns deployment in healthcare settings where specialist annotation coverage is limited. In tertiary centres with multiple radiologists on staff, multi-rater annotation of a CT study is operationally feasible. In rural district hospitals, community radiology practices, and healthcare systems in low-income settings, a single radiologist reading per scan is the norm rather than the exception. At single-annotator training coverage, the per-rater model achieves a 21.4% GED improvement over the shared baseline. A probabilistic model trained with per-rater posteriors on single-annotator data retains substantially more annotation diversity than one trained with a shared encoder under the same data conditions. At maximum sparsity, GED values remain 0.1745 (per-rater) and 0.2220 (shared baseline) — both represent real prediction variance, and neither substitutes for multi-annotator consensus in high-stakes decisions. For a model whose role is to flag boundary uncertainty to a reviewing clinician, the per-rater formulation provides better-calibrated diversity under single-annotator training.

Uncertainty-aware segmentation calibrated to individual radiologist styles enables presenting a range of plausible interpretations to a clinician reviewing a borderline case. A shared-posterior model, shaped by gradient averaging, collapses toward a distribution over gradient-averaged compromises; per-rater training preserves each annotator's gradient signal, so the resulting predictions can span the actual range of clinical judgment.

5.4.2 Indirect Impact

Nodule margin clarity ($r = 0.318$) being the dominant driver of inter-rater mask variance — more than malignancy or subtlety — points to a specific place in the annotation workflow where explicit guidance would reduce disagreement most efficiently. Annotation protocols that require radiologists to record a margin assessment (well-defined, lobulated, ill-defined) before drawing a contour make the margin judgment explicit and documentable before any contour is drawn, which targets the systematic component of inter-rater disagreement. Training programmes that address margin ambiguity at CT tissue transitions would focus on attribute most responsible for annotation variance.

Since malignancy ($r = -0.202$) and boundary ambiguity are separable dimensions, a clinical AI system that treats them as one will misassign uncertainty in both directions: high boundary uncertainty reported for a malignant-appearing nodule with a geometrically clear edge, and low boundary uncertainty for an ambiguous-margin nodule that looks benign. The attribute data support keeping segmentation uncertainty and diagnostic uncertainty as separate model outputs.

5.4.3 Limitations for Societal Deployment

All results come from LIDC-IDRI (CT, lung nodules, four radiologists) and NPC-170 (MRI, nasopharyngeal carcinoma, four annotators). Neither dataset is a prospective clinical study. The per-rater method has not been tested on multi-centre data with known radiologist agreement patterns, on datasets with more than four annotators, or on imaging modalities beyond chest CT and head-and-neck MRI. Clinical deployment requires prospective validation these experiments do not provide.

Per-rater Stage 1 training takes approximately 47 hours across four LIDC-IDRI folds (4.5$\times$ the shared baseline Stage 1 time of approximately 10.5 hours on Apple MPS hardware). For institutions without GPU access that overhead is a real barrier. Parameter sharing between encoders (shared early convolutional layers, diverging near the output) is one way to cut the cost, at the price of partially reintroducing gradient mixing. Distillation from a fully trained per-rater model into a smaller single-encoder student preserves more of the gradient isolation than early-layer sharing, though neither option has been evaluated on this architecture.

APPENDIX I

FOLD-LEVEL RESULTS: FULL ANNOTATION

Table I.1 reports per-fold GED and $\text{Dice}_{\text{match}}$ for both models evaluated at full annotation (all four annotators present, 100-epoch training). These values are read directly from `results/full_annotation_results.csv`. Bold entries mark the better-performing model in each fold.

The comparison here is between per-rater Stage 1 and the shared-posterior Stage 1 baseline — both trained without Stage 2 style vectors and at 100 epochs. The primary thesis comparison (Table 4.1) is between per-rater Stage 1 and the full D-Persona Stage 1+2 pipeline; per-fold data for the full D-Persona pipeline and for ProbUNet were not retained from the original training runs and are reported only as 4-fold means (±SD) in Chapter 4.

Fold 1 is the most instructive row: the shared baseline achieves lower GED (0.1552 vs. 0.1658), i.e. the per-rater model is worse at full annotation in this particular fold. This same fold shows the strongest per-rater advantage under annotation sparsity (Table II.1), which is the mechanistic point developed in §4.3.3.

Table I.1 Per-fold GED and $\text{Dice}_{\text{match}}$, full annotation.

	GED ↓				
Model	**Fold 0**	**Fold 1**	**Fold 2**	**Fold 3**	**Mean**
Per-Rater (ours)	**0.1371**	0.1658	**0.1268**	**0.1418**	**0.1429**
Shared Baseline	0.1426	**0.1552**	0.1338	0.1428	0.1436
	$\text{Dice}_{\text{match}}$ ↑				
Model	**Fold 0**	**Fold 1**	**Fold 2**	**Fold 3**	**Mean**
Per-Rater (ours)	**0.9143**	0.9031	**0.9193**	**0.9152**	**0.9130**
Shared Baseline	0.9132	**0.9074**	0.9165	0.9135	0.9127

APPENDIX II

FOLD-LEVEL RESULTS: SPARSE ANNOTATION

Tables II.1 and II.2 report per-fold GED and $\text{Dice}_{\text{match}}$ for all three sparsity levels (one, two, or three annotators present per training image). Both models were trained for 300 epochs at each sparsity level. All 12 per-fold comparisons in Table II.1 favour the per-rater model (lower GED); the expected probability of this pattern under the null hypothesis of equal models is $\left(\frac{1}{2}\right)^{12} \approx 0.024\%$.

The np = 4 rows (all four annotators present, 300-epoch sparse training) are included in Table II.3 for completeness but are *excluded* from the main results tables in Chapter 4. At 300 epochs with all four annotators, the shared baseline wins three of four folds (mean GED 0.1285 vs. 0.1328 for per-rater). This reversal is a training-duration artefact: 300-epoch training amplifies the degeneration of the shared-posterior objective under the full 5-channel input, producing a model that is worse than the 100-epoch reference. The full-annotation reference throughout this thesis uses 100-epoch models only; the np = 4 sparse row is neither the controlled full-annotation reference nor a clean sparsity result.

Table II.1 Per-fold GED under sparse annotation, 4-fold CV.

n_{present}	Model	Fold 0	Fold 1	Fold 2	Fold 3	Mean
np = 3	Per-Rater	**0.1632**	**0.1708**	**0.1459**	**0.1604**	**0.1601**
	Shared Baseline	0.1775	0.2085	0.1597	0.1782	0.1810
np = 2	Per-Rater	**0.1646**	**0.1912**	**0.1521**	**0.1630**	**0.1677**
	Shared Baseline	0.1979	0.2437	0.1803	0.1938	0.2039
np = 1	Per-Rater	**0.1640**	**0.1806**	**0.1625**	**0.1910**	**0.1745**
	Shared Baseline	0.2257	0.2323	0.2070	0.2231	0.2220

Table II.2 Per-fold $\text{Dice}_{\text{match}}$ under sparse annotation, 4-fold CV.

n_{present}	Model	Fold 0	Fold 1	Fold 2	Fold 3	Mean
np = 3	Per-Rater	**0.9083**	**0.9045**	**0.9154**	**0.9120**	**0.9101**
	Shared Baseline	0.9036	0.8927	0.9117	0.9069	0.9037
np = 2	Per-Rater	**0.9068**	**0.8975**	**0.9135**	**0.9099**	**0.9069**
	Shared Baseline	0.8966	0.8807	0.9041	0.9002	0.8954
np = 1	Per-Rater	**0.9070**	**0.8998**	**0.9104**	**0.9023**	**0.9049**
	Shared Baseline	0.8822	0.8805	0.8910	0.8863	0.8850

Table II.3 Per-fold GED and $\text{Dice}_{\text{match}}$, $n_p = 4$ sparse training.

Metric	Model	Fold 0	Fold 1	Fold 2	Fold 3	Mean
GED ↓	Per-Rater	**0.1242**	0.1567	0.1186	0.1315	0.1328
	Shared Baseline	0.1249	**0.1426**	**0.1153**	**0.1311**	**0.1285**
$\text{Dice}_{\text{match}}$ ↑	Per-Rater	**0.9169**	0.9003	**0.9212**	**0.9164**	0.9137
	Shared Baseline	0.9163	**0.9109**	0.9208	0.9163	**0.9161**

APPENDIX III

FOLD-LEVEL GRADIENT ALIGNMENT RESULTS

Table III.1 reports per-fold mean pairwise cosine similarity of per-rater reconstruction gradients at the shared latent code $\mathbf{z}$, for the shared-posterior baseline at all annotation coverage levels. Each fold value is the mean over $n = 100$ test cases. The per-rater model has zero gradient alignment by construction at all levels: the concept of a shared $\mathbf{z}$ does not exist in the per-rater design, so the measurement is undefined and reported as 0.000.

Two aspects of the full-annotation row are worth noting. The 4-fold mean (0.167) masks substantial case-level variability: within a single fold, the gradient alignment across the 100 test cases has a standard deviation of approximately 0.439 — roughly 38% of cases show negative pairwise cosine similarity (raters' gradients actually pointing in opposite directions) while 62% show positive alignment. At np = 1, this within-fold case spread collapses to approximately 0.023, an approximately 19-fold reduction. The collapse from case-specific conflict at full annotation to near-universal collapse at maximum sparsity is the central diagnostic finding of the gradient alignment analysis.

Table III.1 Gradient alignment per fold, baseline only.

n_{present}	Fold 0	Fold 1	Fold 2	Fold 3	4-fold mean	SD (across folds)
Full (4)	0.1366	0.1827	0.1999	0.1469	0.167	0.026
np = 3	0.2739	0.4163	0.2291	0.2454	0.291	0.074
np = 2	0.4706	0.5242	0.4267	0.4299	0.463	0.039
np = 1	0.9800	0.9807	0.9560	0.9867	0.976	0.012
Per-rater (all levels)	0.000	0.000	0.000	0.000	0.000	—

The np = 1 row shows the lowest across-fold SD (0.012) of any sparsity level — the opposite of what intuition might suggest. At full annotation, case-level variability is high (within-fold SD $\approx$ 0.439) and the 4-fold SD is moderate (0.026). As sparsity increases, the zero-channel gradient dominance overwhelms case-specific boundary differences: by np = 1, the alignment is near-1.0 in all four folds regardless of which test cases are in that fold. The collapse is systematic rather than fold-dependent, which is why the across-fold SD at np = 1 (0.012) is smaller than at full annotation (0.026).

APPENDIX IV

ABLATION STUDY: PER-FOLD BREAKDOWN

Per-fold GED and $\text{Dice}_{\text{match}}$ for the intermediate ablation rows (Rows 2–5 and Row 7 in Table 4.3) were not recorded separately during training. The ablation study tracked 4-fold mean results, and only the endpoint models (shared baseline Stage 1+2 and per-rater Stage 1) have per-fold evaluation data in the CSV output files. Table IV.1 provides the per-fold breakdown for these two rows for reference.

Table IV.1 Per-fold GED and $\text{Dice}_{\text{match}}$, ablation endpoints.

Metric	Row	Fold 0	Fold 1	Fold 2	Fold 3	Mean
GED ↓	Per-Rater Stage 1 (Row 6)	**0.1371**	0.1658	**0.1268**	**0.1418**	**0.1429**
	Shared Baseline Stage 1 only	0.1426	**0.1552**	0.1338	0.1428	0.1436
$\text{Dice}_{\text{match}}$ ↑	Per-Rater Stage 1 (Row 6)	**0.9143**	0.9031	**0.9193**	**0.9152**	**0.9130**
	Shared Baseline Stage 1 only	0.9132	**0.9074**	0.9165	0.9135	0.9127

Note on Row 1: The primary thesis comparison (Table 4.1) compares our model against the full D-Persona Stage 1+2 pipeline (mean GED 0.1507, $\text{Dice}_{\text{match}}$ 0.8909). The per-fold data above is for D-Persona Stage 1 only (the shared posterior without Stage 2 projection heads), which is the relevant baseline for the sparsity experiments. The Stage 1+2 pipeline per-fold values were not retained after the original training.

APPENDIX V

TRAINING HYPERPARAMETERS

Table V.1 lists every hyperparameter held fixed across all experiments. Values are identical for both the full-annotation and sparse-annotation conditions except the epoch count, which increases from 100 to 300 under sparsity to compensate for reduced gradient signal per epoch.

Table V.1 Training hyperparameters, all experiments.

Hyperparameter	**Value**
Backbone	ResNet34
Optimiser	Adam
Learning rate	1×10^{-4}
LR schedule	Cosine annealing
Latent dimension (D)	6
Batch size	12
Epochs (full annotation)	100
Epochs (sparse annotation)	300
β (KL weight)	0.5
Number of annotators (N)	4
Input size (LIDC-IDRI)	$128 \times 128 \times 1$
Input size (NPC-170)	3-channel (T1, T1CE, T2)
Hardware	Apple MPS
Training time (full, 4 folds)	$\sim$47 hours

APPENDIX VI

ATTRIBUTE CORRELATION PER FOLD

Table VI.1 reports the Pearson r between inter-rater margin standard deviation and inter-rater mask variance, computed independently within each fold's test set. This verifies that the margin finding ($r = 0.318$ overall) is not an artefact of any particular train/test partition.

Per-fold correlation analysis was conducted for margin only, the strongest predictor. For the remaining five attributes (lobulation, texture, spiculation, malignancy, subtlety), the Pearson r was computed on the full 1,603-case dataset without fold-level breakdown; those values are reported in Table 4.8 in Chapter 4.

Table VI.1 Per-fold Pearson r: margin disagreement vs. inter-rater mask variance.

Attribute	**Fold 0**	**Fold 1**	**Fold 2**	**Fold 3**	**4-fold range**	**Overall** ($n = 1603$)
Margin (r)	0.238	0.305	0.400	0.359	0.238–0.400	0.318

All four per-fold margin correlations are positive, consistent, and significant at $p < 0.001$. The range (0.238–0.400) spans slightly less than two-fold, which is the expected variability for a correlation estimated on 372–450 cases. Fold 2 has the strongest per-fold association ($r = 0.400$), likely reflecting the composition of that fold's test cases (412 patches, with a higher proportion of lobulated, part-solid nodules). Fold 0 has the lowest ($r = 0.238$), but remains clearly significant. The across-fold consistency confirms that the margin-uncertainty relationship is a structural property of the LIDC-IDRI dataset rather than a data-split artefact. Table VI.2 extends the analysis to all six attributes on the full dataset.

Table VI.2 All-attribute Pearson r, LIDC-IDRI full dataset ($n = 1{,}603$).

Attribute	**Pearson r**	p**-value**
Margin	$+0.318$	< 0.001
Lobulation	$+0.243$	< 0.001
Texture	$+0.210$	< 0.001
Spiculation	$+0.185$	< 0.001
Malignancy	-0.202	< 0.001
Subtlety	-0.155	< 0.001

turnitin Page 71 of 71 - Integrity Submission Submission ID trn:oid:::27535:140806921

Bindu Verma

KeshuShuklaDtuThesisPlagCheck_updated

bindu

Document Details

Submission ID
trn:oid:::27535:140806921

Submission Date
May 28, 2026, 6:57 PM GMT+5:30

Download Date
May 28, 2026, 7:12 PM GMT+5:30

File Name
KeshuShuklaDtuThesisPlagCheck_updated.pdf

File Size
1.8 MB

66 Pages

22,034 Words

119,689 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Take four radiologists reading the same CT slice of a right lower lobe pulmonary nodule. Three draw tight contours around the solid core. The fourth traces a wider boundary to include the surrounding ground-glass haze. Overlaid on the same image, the four delineations do not cluster around a common answer with minor scatter — they represent four distinct clinical judgements about where the lesion ends, and the imaging data does not settle the question.

At a lung nodule's margin, tissue density falls off gradually rather than sharply, ground-glass opacity has no edge a ruler could trace, and the convention for where the nodule stops differs between practitioners and between institutions. These are not correctable errors. Joskowicz et al. [6] reviewed CT annotation studies across anatomical structures and found rater disagreement large enough to change downstream clinical management, and for lung nodules in particular the disagreement is systematic enough that label variability overtakes image noise and class imbalance as the primary source of training uncertainty.

A segmentation model trained on one radiologist's masks learns that radiologist's boundary convention and nothing else. A model trained on the pixel-wise average of four masks chases a boundary that no radiologist actually drew — when two annotators went tight and two went wide, the average is a blurred intermediate that calibrates the model to a compromise opinion nobody holds. Both choices discard what a clinician in a second-opinion setting most needs: not a single boundary location, but a representation of how much expert judgment divides over this case and in which direction each expert resolved the ambiguity.

What the setting actually requires is a model whose output at test time is a distribution over plausible segmentations, calibrated to what the radiologist population produces. On a nodule with a clean, isolated, solid boundary where all four annotators converge, the model's samples should cluster. On a nodule with ill-defined margins, a part-solid texture, and four substantially different contours from four qualified radiologists, the samples should span that range. A single-output model has one forward

pass and one mask per input: the architecture gives the clinician reviewing the case a prediction, not a characterisation of how much the radiologist population divides on that particular boundary.

1.1.1 The LIDC-IDRI Dataset and the GED Metric

The primary dataset is LIDC-IDRI [1]: 1,018 CT scans, each with four independent radiologist annotations of pulmonary nodules, yielding 1,609 nodule patches after preprocessing (described in full in §2.6). Armato et al. [2] documented that in this dataset, which annotator's mask is treated as ground truth shapes model calibration more than image noise or class imbalance does — boundary extent differences between radiologists were that large. The difference between single-output models and probabilistic ones shows up in measured GED and $\text{Dice}_{\text{match}}$, and the clinical stakes of miscalibrated uncertainty in a nodule screening setting are well-documented.

The Generalised Energy Distance [11] captures both failure modes in a single scalar. When a model collapses to the same prediction on every forward pass, its within-sample distance term goes to zero regardless of how accurate that one prediction is — GED penalises this heavily. A model that spreads predictions widely but misses every annotator's mask also scores poorly. Throughout this thesis, GED is the primary metric and lower values are better. $\text{Dice}_{\text{match}}$ is the secondary metric: it uses optimal Hungarian assignment to match model samples to individual rater annotations and measures geometric closeness to each annotator separately, so diversity and per-rater accuracy can be tracked independently.

1.1.2 Probabilistic Segmentation Models

Probabilistic segmentation models address the distribution-output requirement by replacing the single-mask prediction with a stochastic latent variable that encodes annotation-level uncertainty. The Probabilistic U-Net (ProbUNet) [11], which combined a conditional variational autoencoder (VAE) [9] with the U-Net backbone [14], was the first model to demonstrate genuinely diverse segmentation hypotheses on LIDC-IDRI. Subsequent work extended this line in different directions: PHiSeg [3], which addressed coarse-versus-fine scale disagreement by maintaining separate latent variables at each spatial resolution, and Stochastic Segmentation Networks [12], which bypassed the latent variable altogether and parameterised uncertainty as a distribution over the full output mask. Each addressed different aspect of how annotation uncertainty should be represented. The qualitative result across all three is the same: diverse, plausible samples can be drawn at test time by sampling from a learned prior over the latent space.

In the multi-rater setting, however, training a single posterior encoder on all annotators' masks simultaneously introduces a problem that persists regardless of what

architectural choices are made above the posterior level. The problem originates in the backward pass: conflicting rater gradients accumulate at the shared latent code and the encoder's weights shift toward a compromise that satisfies none of the individual rater signals cleanly.

1.2 The Gradient Conflict Problem

In the shared posterior setup, a single encoder takes the image and all four annotators' masks as a concatenated 5-channel tensor and outputs a distribution over a shared latent code $\mathbf{z}$. A decoder maps samples from $\mathbf{z}$ to segmentation masks. At test time only the prior runs — the posterior encoder plays no role in inference — so the diversity of prior samples at test time is what GED ends up measuring.

Training this encoder requires a reconstruction loss for each annotator (a term that pushes the decoder's output toward each radiologist's mask). The gradient of each such term flows back through the decoder and accumulates at the shared latent code $\mathbf{z}$. When four radiologists annotated the same nodule with four different boundary decisions, their four reconstruction gradients push $\mathbf{z}$ in four different directions simultaneously. The shared encoder's weights are updated according to the sum of these conflicting signals.

Take radiologist 1, who drew a tight contour around the solid core, and radiologist 4, who extended the boundary outward into the surrounding opacity. The gradient from radiologist 1's reconstruction loss pushes $\mathbf{z}$ toward a code from which a tight prediction decodes. The gradient from radiologist 4's loss pushes $\mathbf{z}$ the other way. Both gradients hit the shared encoder at same backward pass, their effects partially cancel, and the encoder's weights shift toward a compromise that fits neither boundary decision. When two annotators label completely non-overlapping regions, the cancellation is exact and the net gradient at $\mathbf{z}$ is zero.

This problem is structurally equivalent to gradient conflict in multi-task learning. Yu et al. [19] showed this conflict can be measured through the cosine similarity between per-task gradients at the shared encoder, which quantifies how much those tasks compete rather than reinforce each other, and the resulting per-task performance loss tracks that cosine similarity. A negative cosine similarity means two tasks' gradients actively pull encoder weights in opposite directions, and the net update that the encoder receives is a blurred compromise that serves neither task well. Training a shared posterior on N annotators creates the same condition, with each annotator's reconstruction loss acting as a separate task competing over the shared encoder, and the net gradient encoding none of the competing boundary decisions cleanly.

1.2.1 D-Persona and the Limits of Stage 2 Personalisation

D-Persona [17] is the current state of the art in multi-rater segmentation and the primary baseline for this thesis. Its response to the shared posterior limitation is a two-stage training pipeline. Stage 1 trains the shared posterior encoder with all annotators' masks concatenated, producing a shared latent code $\mathbf{z}$. Stage 2 introduces per-rater projection heads that steer this shared code toward individual annotator styles at inference time. On LIDC-IDRI, the full two-stage pipeline achieves GED 0.1507 and $\text{Dice}_{\text{match}}$ 0.8909, a substantial improvement over ProbUNet (GED 0.2234, $\text{Dice}_{\text{match}}$ 0.8836).

Stage 2 was built on the assumption that a shared latent code contains enough per-rater structure for a small projection head to extract and amplify. A head applied to $\mathbf{z}$ can redistribute whatever information is already there — it can rotate dimensions, amplify or suppress components, and bias the output toward given annotator's boundary style. The problem is that Stage 1's gradient averaging removes per-rater structure from $\mathbf{z}$ before Stage 2 ever sees it. Each annotator's gradient signal gets partially cancelled by the others during Stage 1's backward pass, and Stage 2's heads have nothing rater-specific to work with.

The ablation study in Chapter 4 makes this concrete. Adding Stage 2 projection heads on top of a Stage 1 trained with per-rater posteriors (which already provides clean, rater-specific gradients) raises GED by 27%, from 0.1444 to 0.1836. When Stage 2 is applied on top of a Stage 1 that already provides clean, rater-specific gradients, it introduces shared-posterior dynamics on a model that does not need them. The projection heads find no rater-specific structure to amplify and instead collapse the diversity Stage 1 created.

1.2.2 The Architectural Capacity Hypothesis

Before arriving at the per-rater formulation, a simpler hypothesis should be ruled out: if the shared posterior's bottleneck is representational capacity rather than training objective, a more expressive encoder should help regardless of how the objective is structured. Transformer-based architectures, specifically the Mix Transformer (MiT-B2) from Xie et al. [18], represent exactly this class of increased capacity. MiT-B2's attention mechanism is not limited by local kernel size; it can, in principle, relate any two spatial locations regardless of distance. If annotation-ambiguous regions produce interacting gradient signals that a convolutional encoder cannot disentangle because of its limited receptive field, MiT-B2 should show improvement over ResNet34 in shared posterior setting. Substituting MiT-B2 into the shared posterior encoder tests this: if the problem is architectural expressiveness, a stronger encoder should close the GED gap.

The experiment is described in full in §4.2.3 (Table 4.4, Row 2). MiT-B2 yields GED 0.1531 versus the ResNet34 baseline's 0.1507, with no change in $\text{Dice}_{\text{match}}$. A more expressive encoder learns a more expressive average of conflicting

rater gradients; the conflict itself is unaffected by how large the encoder is. Increasing encoder expressiveness changes nothing about gradient conflict, because the conflict is a property of the training objective (specifically which gradient signals reach the encoder) and not of the encoder's capacity.

1.2.3 Per-Rater Posterior Encoders: The Proposed Solution

Replacing the shared posterior with N independent encoders, each dedicated to one annotator, removes the conflict where it originates. Each encoder receives a 2-channel input: the image and that annotator's mask only. Its gradient signal reflects only that annotator's boundary decisions, with no contribution from any other annotator's loss. The shared decoder and the shared prior receive N rater-specific signals rather than one conflict-averaged signal.

The training objective decomposes into N per-rater ELBO terms, one per annotator. Since each ELBO term depends only on its own encoder's output, $\partial \mathscr{L}_i / \partial \mathbf{z}_j = 0$ for $i \neq j$ — the gradient from encoder i carries no information from encoder j, and no regularisation is needed to enforce this. It follows from the objective structure alone.

At test time, the N posterior encoders are dropped. Inference is identical to the D-Persona baseline: sample $\mathbf{z}$ from the shared prior, decode. The training cost is approximately $4.5\times$ higher; inference cost is unchanged. Every performance difference in Chapter 4 originates from training-time gradient quality, since nothing else differs between the two models.

On LIDC-IDRI, this Stage 1-only per-rater model achieves GED 0.1444 ± 0.0141 (-4.2%) and $\text{Dice}_{\text{match}}$ 0.9112 ± 0.0061 ($+2.28\%$) compared to the full D-Persona two-stage pipeline, with per-rater Dice improving for each individual expert. Cross-dataset validation on NPC-170 (nasopharyngeal carcinoma MRI, 4 annotators, 3-seed evaluation) produces a GED gap of 0.0011, which falls within the seed variance of ± 0.0085, indicating neither improvement nor degradation.

1.3 The Annotation Sparsity Extension

LIDC-IDRI guarantees four annotations per case by design. Most clinical imaging archives do not. A retrospective dataset assembled from routine reads may have three annotators on some scans, one on others, and the coverage is uneven in ways that reflect staffing history rather than data quality decisions. The full-annotation results matter, but the more operationally relevant test is how each model degrades when annotation coverage is patchy.

1.3.1 Sparse Annotation in the Shared Posterior

In the shared posterior model, an absent annotator's mask channel is set to zero. The encoder still receives a full 5-channel input; the absent annotator's channel contains only zeros. Two things follow from this. First, the zero-mask channel produces a reconstruction gradient that pushes the latent code toward predicting an empty segmentation for the absent annotator (a gradient that competes with the present annotators' reconstruction gradients). Second, as more annotators become absent, the proportion of zero-mask gradient signals grows. When only one of four annotators is present, the shared encoder receives three zero-mask gradients and one real annotation gradient. The latent code is simultaneously pulled toward the present annotator's boundary decision and toward predicting nothing for three absent annotators. With three absent annotators versus one present, the zero-mask gradients outnumber the real annotation gradient three to one, and the net update at $\mathbf{z}$ is pulled predominantly toward predicting empty masks.

1.3.2 Sparse Annotation in Per-Rater Posteriors

In the per-rater model, absent annotators are handled through omission: if annotators 1, 3, and 4 are absent for a given image, their three encoders simply do not execute. The gradient update for that training step comes only from encoder 2, carrying only encoder 2's rater-specific signal. Nothing in the architecture forces a zero-channel placeholder into the gradient computation; the absent raters leave no trace in the parameter update at all. Chapter 4's sparsity results follow directly from that difference in how absence is represented.

1.3.3 Experimental Evidence

At one annotator per training image, the per-rater model achieves $+21.4\%$ GED over the shared baseline. At two annotators, $+17.8\%$. At three, $+11.5\%$. The gap widens at every step down in coverage, and across all 12 per-fold comparisons (three sparsity levels times four folds), the per-rater model wins every single one. Under a null hypothesis of no systematic difference, $(1/2)^{12} \approx 0.024\%$ is the probability of that sweep.

A gradient alignment diagnostic provides a mechanistic picture of what is happening inside the shared baseline as annotation coverage decreases. The mean pairwise cosine similarity of per-rater reconstruction gradients at the shared latent code rises from 0.167 at full annotation to 0.976 at one annotator present (a near-complete collapse to a single gradient direction, as three zero-mask signals drive all gradients toward the same degenerate prediction of an empty mask). The within-fold spread of this alignment measure collapses approximately 19-fold (standard deviation from 0.439

at full annotation to 0.023 at one annotator present). At maximum sparsity, the collapse is nearly universal across test cases, not limited to specific high-disagreement nodules. The per-rater model's gradient alignment stays at 0.000 at all sparsity levels, since the architecture has no shared $\mathbf{z}$ to measure alignment on.

At full annotation the per-rater GED advantage is 0.5%, inside the noise. Fold 1 makes the sparsity-specificity of the advantage concrete: at full annotation the shared baseline actually beats the per-rater model in that fold (GED 0.1552 vs. 0.1658, baseline wins by 6.8%), yet at one annotator per training image the same fold reverses completely (0.2323 vs. 0.1806, per-rater wins by 22.3%). A model that was worse at full annotation in Fold 1 outperforms the shared baseline by 22.3% in the same fold once annotation drops to a single rater. The reversal is tied to the sparsity condition, not to any overall quality difference between the two models: the shared baseline's zero-channel gradients dominate the latent code under single-annotator training in ways that do not occur at full annotation.

1.4 Thesis Contributions

Four contributions follow from this work.

1. **Per-rater posterior encoders that isolate each annotator's gradient at Stage 1.** Giving each annotator a dedicated posterior encoder, trained only on that annotator's mask, eliminates gradient conflict at the source. The isolation property ($\partial \mathcal{L}_i / \partial \mathbf{z}_j = 0$ for $i \neq j$) falls out of the per-rater ELBO decomposition without any additional regularisation. On LIDC-IDRI (1,609 nodule patches, 4-fold cross-validation), this Stage 1-only design achieves GED 0.1444 ± 0.0141 and $\text{Dice}_{\text{match}}$ 0.9112 ± 0.0061, a 4.2% GED reduction and 2.28% $\text{Dice}_{\text{match}}$ gain over the full D-Persona two-stage pipeline; every individual annotator's per-rater Dice improves. On NPC-170 (nasopharyngeal carcinoma MRI), the GED gap is 0.0011, within seed variance, confirming the method transfers without LIDC-specific tuning.

2. **A seven-row ablation that isolates the posterior encoder as the load-bearing component.** Four design alternatives were tested before the per-rater formulation: a transformer backbone (MiT-B2), an orthogonality regularisation loss, a discretised prior bank ($k = 100$), and a dual diversity loss. None improves GED consistently, because none changes what gradients the encoder receives during training. Adding Stage 2 style vectors on top of per-rater Stage 1 (Row 7) makes things worse — GED rises from 0.1444 to 0.1836 — confirming that Stage 2 is only useful when Stage 1 has failed to encode rater signals, and actively harmful in the opposite case.

3. **Sparsity experiments spanning three annotation coverage levels, with a gradient collapse diagnostic.** Training both models at $\text{np} \in \{1, 2, 3\}$ annotators per image shows the per-rater advantage widening as coverage drops: $+11.5\%$

at three annotators, $+17.8\%$ at two, $+21.4\%$ at one, with the per-rater model winning all 12 per-fold comparisons. The collapse mechanism in the shared baseline is measurable: mean pairwise gradient cosine similarity at the shared latent code rises from 0.167 at full annotation to 0.976 at one annotator, with within-fold standard deviation shrinking 19-fold (0.439 to 0.023) — gradient conflict goes from case-specific to effectively universal.

4. **Pearson correlation analysis linking nine LIDC-IDRI nodule attributes to inter-rater mask variance.** Run on 1,603 cases with complete attribute records, the analysis finds nodule margin clarity ($r = 0.318$, $p < 0.001$, confirmed in all four folds with per-fold r ranging from 0.238 to 0.400) as the strongest predictor of boundary disagreement. Malignancy is negatively correlated ($r = -0.202$): a nodule perceived as highly suspicious need not have an ambiguous contour, and the two dimensions do not necessary travel together. Uncertainty-aware methods are most relevant for ill-defined, lobulated, part-solid nodules rather than for the cases rated most dangerous.

1.5 Thesis Organisation

The literature review in Chapter 2 traces the gradient-conflict problem through probabilistic segmentation and multi-annotator learning, identifying where each strand leaves the shared-encoder bottleneck intact. Chapter 3 formalises the per-rater ELBO, proves gradient isolation, and sets up both the sparse annotation protocol and the two diagnostic analyses. Experimental results, the ablation, and the sparsity study are in Chapter 4, together with a discussion of scope and limitations. Conclusions, open questions, and deployment implications for settings where radiologist coverage is thin are in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

The work reviewed here spans deterministic segmentation, probabilistic models for annotation diversity, multi-annotator label fusion, and gradient conflict in shared-encoder training. Each strand advanced part of the multi-rater segmentation problem. The shared posterior encoder's training dynamics — how conflicting rater gradients accumulate at $\mathbf{z}$ and what that does to the latent code — remained unaddressed across all of them. Chapters 3 and 4 take up that gap directly.

2.1 Medical Image Segmentation and the Uncertainty Problem

Medical image segmentation assigns each pixel to an anatomical structure or pathological region, and that assignment feeds directly into clinical decisions. Ronneberger et al.'s U-Net [14] made dense pixel prediction practical on limited annotated data by pairing a contracting encoder with skip connections that carried high-resolution spatial detail into the expanding decoder. Within two years of the 2015 paper it had become the default architecture across modalities and anatomies — a rapid adoption driven by a simple fit between the architecture and the setting it was designed for, where each training image came with one reference mask.

That assumption breaks down in practice. A radiologist delineating a lung nodule, a cardiologist tracing a ventricular wall, or a radiation oncologist contouring a tumour volume all face the same problem: tissue boundaries in medical images are gradual rather than sharp, imaging noise obscures the transition, and where exactly the structure ends depends on conventions that differ between institutions and between individual practitioners. Multiple trained experts examining the same scan can produce genuinely different boundaries, all of them defensible.

2.1.1 Two Kinds of Uncertainty

Kendall and Gal [7] separated uncertainty into two categories: the kind that shrinks as more training data accumulates (epistemic, about model parameters), and the kind that

persists regardless of how much data is collected (aleatoric, in the observation itself).

Expert disagreement at nodule margins falls in the second category. The four radiologists in LIDC-IDRI are responding to a tissue boundary that the CT image does not resolve to a single location — their disagreement is a property of the image, and collecting more annotated scans would produce more instances of that same disagreement rather than resolving it. A model minimising expected loss against one annotator's masks is calibrated to that annotator's convention and nothing else. What clinical deployment actually requires is a distribution over plausible boundaries, matched to the spread the expert population produces on that specific image.

A single predicted boundary for an ambiguous structure presents one outcome as certain when the clinical reality is a range of defensible contours. The distribution over boundaries carries clinical information too: which boundaries are plausible and how much the experts diverge. A nodule where all four radiologists agree has a fundamentally different uncertainty profile from one where they draw four substantially different contours. A model that collapses both to identical single predictions treats the two situations as equivalent, even though the annotation spread is the clinically meaningful difference between them.

2.1.2 The Scale of Inter-Rater Variability

Joskowicz et al. [6] reviewed CT annotation studies across anatomical structures and found rater disagreement large enough, in many cases, to change clinical management decisions rather than scatter narrowly around a stable consensus. For lung nodules, where ground-glass opacity and tissue density gradients complicate boundary definition, the disagreement grows substantial enough that the annotation label itself becomes the dominant source of training uncertainty — larger in practice than image noise or class imbalance.

The LIDC-IDRI dataset [1], described in detail in §2.6, makes this concrete. Four radiologists independently annotated 1,018 CT scans, and the annotation disagreement is large enough that treating any single radiologist mask as ground truth introduces systematic bias: the resulting model is calibrated to one radiologist's style and uncalibrated to all others (this is established experimentally in §1.2).

2.1.3 Why Deterministic Models Are Insufficient

U-Net and its variants produce single point estimate for each image, one mask per forward pass calibrated to one assumed correct answer. This works well for the problems they were designed to solve, cases with a single clear annotation. For ambiguous anatomy with multiple annotators, this design is structurally incomplete.

A good U-Net can approximate the mean radiologist decision well enough on average. The problem is that it has no mechanism to represent the variability around that mean. It cannot tell a clinician: “for this nodule, two of the four radiologists drew a tight contour around the core and two included the surrounding ground-glass region. My prediction represents one of these choices, not the consensus.” That kind of output — quantified boundary uncertainty, attributed to specific annotator decisions — is what a point-estimate architecture is structurally incapable of generating, regardless of how well the single prediction tracks the mean annotator.

The model the setting requires outputs a distribution over masks, generating multiple distinct hypotheses each reflecting a plausible radiologist decision. Kohl et al., Baumgartner et al., and Monteiro et al. each built toward this goal from different angles.

2.2 Probabilistic Segmentation Models

Moving from deterministic to probabilistic segmentation meant changing what the model’s output represents. Rather than one mask per image, the goal is a learned distribution $p(\hat{y} \mid \mathbf{x})$ from which diverse samples can be drawn at test time. Three architectures developed in this direction, each solving a different part of the problem.

2.2.1 Probabilistic U-Net

Kohl et al. [11] combined a conditional variational autoencoder [9] with the U-Net backbone to create the Probabilistic U-Net (ProbUNet). Its key structural property is that the training pathway and the inference pathway are different. During training, a posterior encoder receives the image together with the annotation, using that annotation signal to supervise the latent variable $\mathbf{z}$. At test time, the posterior encoder is not used at all; $\mathbf{z}$ is sampled from a prior that is conditioned on the image alone. The KL term in the training objective pulls this prior toward the annotation distribution during training, so that prior samples at test time span the range of plausible boundary decisions. This was the first model to produce genuinely varied predictions for the same input. Four samples from ProbUNet span qualitatively different boundary decisions on the same lung nodule; the deterministic U-Net rules this out entirely, since its single forward pass produces one fixed prediction per input with no mechanism to represent boundary ambiguity.

However, ProbUNet has a specific limitation that becomes apparent when multiple annotators are involved. The original model was designed and evaluated with one annotation per training image (randomly selected from among the available annotators). Applying it to a multi-rater dataset requires a decision about what the posterior encoder should receive. In D-Persona’s implementation [17], the posterior is adapted to receive all four rater masks simultaneously. As discussed below, this

adaptation introduces the bottleneck that this thesis targets.

A separate issue is that ProbUNet's 6-dimensional latent space is a single global code for the entire image patch. It captures patch-level uncertainty (whether the overall annotation is large or small, tight or loose) but has limited capacity to represent spatially structured disagreement where radiologists agree on some regions and disagree on others. This limitation motivated the hierarchical extension discussed next.

2.2.2 PHiSeg: Hierarchical Probabilistic Segmentation

Hierarchical latent variable models for segmentation were proposed concurrently by Kohl et al. [10] (Hierarchical ProbUNet) and Baumgartner et al. [3] (PHiSeg). Both start from the observation that annotation uncertainty is not scale-invariant: two radiologists may agree on which region contains a nodule but disagree on where exactly its boundary falls. A single global latent variable cannot separately represent these two distinct levels of disagreement: it conflates a coarse placement decision with a fine boundary decision into one low-dimensional code. PHiSeg's response is to maintain a separate latent variable at each spatial resolution of the encoder, so that disagreements at different scales can be independently captured and sampled.

PHiSeg demonstrated improved sample diversity and calibration on LIDC-IDRI compared to the single-latent ProbUNet, and its samples can differ at fine spatial scales while remaining consistent at coarse scales — a structure that matches expert disagreement better than a single global code.

The gradient conflict problem, however, operates at every level of the hierarchy. Each hierarchical posterior still receives all annotators' masks together, and each scale's shared latent code accumulates the same sum of competing per-rater gradients. Adding more levels multiplies the representational capacity but leaves the training dynamics at each level unchanged. A more expressive hierarchy with same shared-posterior training objective inherits the same bottleneck.

2.2.3 Stochastic Segmentation Networks

Monteiro et al. [12] took a structurally different approach. Rather than compressing annotation uncertainty into a low-dimensional latent variable, they parameterised a distribution directly over the full output mask. This lets SSN represent spatially-structured uncertainty: a nodule core where radiologists agree behaves differently in the output distribution than a margin region where they diverge. A global 6-dimensional latent code cannot represent this distinction, because it collapses the entire image into one low-dimensional uncertainty estimate. SSN makes this spatial parameterisation computationally tractable through a structured covariance approximation. There is no posterior/prior split: the model has one pathway for both training and inference.

SSN remains a single-distribution model, however. It learns one distribution $p(\hat{y} \mid \mathbf{x})$ over all annotations pooled together, without distinguishing individual annotator styles. Sampling from SSN produces masks that span the annotation distribution in aggregate, but individual samples cannot be attributed to specific annotators. In a setting where the clinical goal is to produce four predictions each reflecting one radiologist's style, SSN provides a distribution over the pooled annotation space with no mechanism to steer individual samples toward a specific radiologist's boundary convention.

2.2.4 The Common Bottleneck

ProbUNet, PHiSeg, and SSN each solve part of the problem, producing diverse predictions that outperform deterministic baselines on LIDC-IDRI, yet all three were designed assuming a single training pathway processes all available annotations together. The question of whether mixing annotators' gradients inside a shared encoder is the right design choice was not part of how any of them were built.

Each model, when applied to multi-rater data, encodes annotation information through a single pathway (a latent code, a hierarchical latent code, or a direct output distribution) that must simultaneously accommodate every annotator's mask. The training signal that reaches this shared representation is an aggregate over all annotators. When annotators disagree, these aggregated signals partially cancel each other. The shared representation is pulled toward a weighted compromise across four conflicting annotation signals, one that fits no individual annotator's style well. Methods that explicitly model individual annotators — STAPLE, confusion matrices, calibrated consensus — take a different approach, though as the following review shows, the shared-encoder gradient problem persists in all of them.

2.3 Multi-Annotator Learning

A separate literature addressed multi-annotator learning from a different angle: modelling the annotation process explicitly, asking who annotated what, how reliable each annotator is, and how to aggregate or disentangle their contributions. Zhang et al. [21] provide a recent taxonomy. Five approaches from this line are directly relevant to the per-rater posterior contribution of this thesis.

2.3.1 STAPLE: Simultaneous Truth and Performance Level Estimation

Warfield et al. [16] formalised multi-annotator fusion as an expectation-maximisation problem over two coupled unknowns: what the true segmentation is, and how reliable each annotator is relative to that truth. The E-step estimates per-annotator reliability given the current consensus; the M-step updates the consensus given those reliability

estimates. Iteration continues until both quantities stabilise. The final output is a single weighted consensus mask, with annotators whose labels track the group receiving more influence over the result than those who deviate.

STAPLE is still used in clinical settings where a consensus reference standard is needed for downstream training. For the purposes of this thesis, a single consensus label answers a different question from the one being asked: the goal here is a distribution over plausible annotations, and a consensus collapses exactly the disagreement that distribution is meant to preserve. It models annotator reliability as a property of the annotator relative to an assumed ground truth, not as a source of clinically meaningful variation. For lung nodule segmentation, where expert disagreement reflects genuine boundary ambiguity rather than annotator error, assuming a single hidden true contour that all four radiologists are noisily trying to identify is not appropriate.

2.3.2 Annotator Confusion Models

Rodrigues et al. [13] and Tanno et al. [15] developed more flexible models of annotator behaviour. Rodrigues et al. placed Gaussian process classifiers over each annotator's labels with per-annotator latent functions, marginalising over annotator-specific noise to estimate a true label. Tanno et al. [15] took a discriminative approach: each annotator is characterised by a per-class confusion matrix encoding how systematically they deviate from the true label. This gives each annotator a learned behavioural fingerprint (rather than a single reliability scalar), so the model can distinguish an annotator who consistently over-segments from one whose labels are correct on average but noisy at boundaries. Both models learn annotator-specific behaviour and produce consensus estimates that account for those learned properties.

Both models move beyond STAPLE's single reliability scalar. Annotator-specific parameterisation lets the model weight each annotator's contribution in proportion to how much their labels agree with the others, so a radiologist who consistently over-segments gets lower weight than one who tracks the group more closely. Tanno et al.'s confusion matrix formulation has been applied to medical image segmentation, where the per-annotator reliability estimate produces more accurate consensus labels than treating all annotators equally.

The limitation is the same one that affects STAPLE: the output is a consensus label, not a distribution over plausible annotations. The confusion matrix framework also assumes there exists a true class label that annotators are confusedly trying to identify. This assumption is defensible for tasks like pathology slide classification where a ground truth (the tissue pathology) exists in principle, even if it is difficult to observe. For lung nodule boundary delineation, the assumption breaks down because the tissue transition at the nodule edge is gradual enough that four qualified radiologists can draw four genuinely different contours, each defensible, with no hidden ground truth contour that any of them is approximating.

2.3.3 Multi-Rater Calibration

Rather than averaging annotator decisions uniformly, Ji et al. [5] weight each annotator's contribution proportionally to how much that annotator's label matches the broader population at each pixel, producing a consensus more representative of the annotation distribution than a simple mean.

For many clinical applications that calibrated consensus is useful — a model that recognises when four radiologists converge versus when they split two-against-two carries more information than one that always produces the same confidence. Producing four distinct predictions each reflecting one radiologist's boundary style, spanning the full range of clinically defensible decisions, is a different goal and one that weighted consensus methods leave unaddressed.

2.3.4 Disentangling Human Error from Ground Truth

Zhang et al. [20] introduced a framework that separated annotation variance into two structurally distinct components: variability driven by the image itself (present regardless of which annotator is involved) and variability driven by individual annotator tendencies (idiosyncratic errors or biases specific to each person). Because these two sources have different structures (one is a property of the image, the other is a property of the annotator), pooling them into a single noise term wastes model capacity and produces poorly calibrated outputs. Treating them separately allows the model to assign image-level uncertainty to cases where any expert would struggle, and annotator-level uncertainty to cases where one annotator systematically diverges from the others.

This decomposition is the closest conceptually to what per-rater posterior encoders achieve. By modelling each annotator's contribution separately and identifying the annotator-specific component, Zhang et al. move toward individual-level annotation modelling rather than population-level averaging. The limitation is that the framework produces a discriminative model (a classifier with per-annotator noise components) rather than a generative model that can sample from the distribution of plausible annotations. GED evaluation requires drawing multiple diverse prior samples at test time, and a discriminative classifier with no prior has no sampling mechanism to support this.

2.3.5 D-Persona: Per-Rater Personalisation of Shared Posteriors

D-Persona [17] introduced a two-stage training pipeline that combined the latent-space diversity of ProbUNet with per-rater personalisation — the first method to target individual radiologist style from a probabilistic segmentation base.

In Stage 1, a shared posterior encoder receives all four annotators' masks simultaneously as a single multi-channel input and produces one shared latent code. This adapts ProbUNet to the multi-rater setting by widening the posterior from one annotation to four. An auxiliary range loss pushes the prior to cover the full breadth of annotator decisions rather than collapsing toward the mean. Feeding all four masks into one encoder produces gradient conflict at the shared latent code.

Stage 2 adds per-rater projection heads on top of the frozen Stage 1 prior. Stage 1 parameters are not updated. Each head takes the shared $\mathbf{z}$ and applies a small rater-specific transformation, steering samples toward that annotator's boundary style at inference time. Stage 2 can steer samples toward a rater's style only if that style is already present somewhere in the shared latent code. Whether that condition holds depends on what Stage 1's training objective actually encodes, and the gradient analysis in §3.2.3 examines this.

On LIDC-IDRI, D-Persona substantially outperforms ProbUNet on both GED and $\text{Dice}_{\text{match}}$; full numerical comparisons are in Table 4.2.

D-Persona represents the state of the art in per-rater prediction, but its Stage 1 training introduces a structural bottleneck that Stage 2 cannot compensate for. Because the shared encoder processes all four annotators' masks simultaneously, the gradient update at the shared latent code is an average of four per-rater terms. On nodules where all four radiologists draw similar boundaries, these terms reinforce each other and the shared code receives a clean signal. At ill-defined nodule margins — the region that matters clinically — individual rater gradients point in different directions and partially cancel, pulling the latent code toward a weighted average that fits none of the four boundary decisions cleanly; what the shared $\mathbf{z}$ encodes is shaped by all four radiologists at once rather than by any single annotator's boundary convention. The full derivation of this gradient conflict is in §3.2.3.

Stage 2 personalisation then has to extract per-rater information from a $\mathbf{z}$ that does not contain it. Whatever the projection heads h_i do to the shared code, they are working with the gradient-averaged residual from Stage 1 — whatever was cancelled out during Stage 1's backward pass is gone before h_i ever runs. Stage 1's gradient averaging is the bottleneck, and Stage 2 operates too late in the pipeline to address it.

The ablation in §4.2.3 confirms this empirically: adding Stage 2 style vectors after a Stage 1 trained with per-rater posteriors worsens GED by 27% (from 0.1444 to 0.1836). Stage 2 projection heads find something to steer only when Stage 1 has left per-rater structure underspecified in $\mathbf{z}$. After a per-rater Stage 1, that structure is already encoded cleanly, and the heads collapse it.

Giving each annotator a dedicated posterior encoder — trained on that annotator's mask alone — removes the gradient averaging from Stage 1's training objective, the point where the information loss actually occurs.

2.4 Variational Inference and the ELBO

The per-rater posterior formulation rests on standard variational inference. Two properties of the ELBO are relevant: that the per-rater objective is a valid variational lower bound, and that the $\frac{1}{N}$ normalisation keeps KL regularisation pressure comparable to the baseline — without it, four KL terms together would over-regularise and collapse the latent diversity GED requires. §2.5 then connects these mathematical foundations to the gradient conflict problem.

2.4.1 Variational Autoencoders and the Evidence Lower Bound

Kingma and Welling [9] introduced the variational autoencoder (VAE) as a framework for learning latent-variable generative models. The difficulty is tractability: the true posterior $p(\mathbf{z} \mid x)$ requires integrating over all possible $\mathbf{z}$, which has no closed form for continuous latent variables. The VAE resolves this through amortised inference: a learned encoder $q_\phi(\mathbf{z} \mid x)$ stands in for the intractable posterior, and the training objective is constructed so that pushing this approximation closer to the true posterior simultaneously improves the generative model.

The ELBO is derived by applying Jensen's inequality to the log-marginal likelihood:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(\mathbf{z}|x)}\big[\log p_\theta(x \mid \mathbf{z})\big] - \mathrm{KL}\big(q_\phi(\mathbf{z}|x) \,\|\, p(\mathbf{z})\big). \tag{2.1}$$

The reconstruction term measures how well the decoder recovers the original observation from a latent sample. The KL term keeps the approximate posterior anchored near the prior: an unconstrained posterior would collapse each observation to a near-deterministic code, fitting training data well but producing incoherent outputs at test time when the prior is sampled instead. Maximising the bound jointly over encoder and decoder parameters drives both toward a useful equilibrium.

In ProbUNet and D-Persona, the observation is the segmentation mask y, the latent variable is $\mathbf{z}$, and both the prior and posterior are conditioned on the image $\mathbf{x}$:

$$\log p(y \mid \mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x},y)}\big[\log p(y \mid \mathbf{z}, \mathbf{x})\big] - \mathrm{KL}\big(q(\mathbf{z}|\mathbf{x},y) \,\|\, p(\mathbf{z}|\mathbf{x})\big). \tag{2.2}$$

2.4.2 The AxisAlignedConvGaussian Encoder

Both the prior $p(\mathbf{z} \mid \mathbf{x})$ and the posterior $q(\mathbf{z} \mid \mathbf{x}, y)$ in ProbUNet are implemented as AxisAlignedConvGaussian networks [11]: convolutional encoders whose output is a mean μ and a log-variance $\log \sigma^2$, one scalar per latent dimension. The diagonal covariance assumption is an efficiency choice: a full $D \times D$ covariance would be

impractical to parameterise and invert, while a diagonal Gaussian with flexible means and variances retains enough expressiveness for the 6-dimensional latent space used here.

Backpropagating through a stochastic sample would ordinarily block gradient flow, since drawing from a distribution is not a differentiable operation. Kingma and Welling's reparameterisation [9] resolves this: by expressing the sample as $\mathbf{z} = \mu + \sigma \odot \varepsilon$, the stochastic draw is factored into a deterministic function of encoder outputs and a noise variable $\varepsilon \sim \mathcal{N}(0, I)$ that carries no parameter dependence. Gradients reach μ and σ through the deterministic path; ε is never differentiated.

2.4.3 Additivity of the ELBO for Independent Observations

The mathematical foundation for the per-rater ELBO is a standard property of the ELBO under independent observations. If N observations $\{y_i\}_{i=1}^{N}$ are conditionally independent given $\mathbf{z}$ and $\mathbf{x}$, then the log-joint likelihood decomposes:

$$\log p(y_1, \ldots, y_N \mid \mathbf{x}) = \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}). \tag{2.3}$$

Applying the ELBO bound to each term independently and summing:

$$\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}) \geq \sum_{i=1}^{N} \Big[\mathbb{E}_{q_i(\mathbf{z}|\mathbf{x},y_i)} \big[\log p(y_i \mid \mathbf{z}, \mathbf{x}) \big] - \mathrm{KL}\big(q_i \,\|\, p(\mathbf{z}|\mathbf{x}) \big) \Big]. \tag{2.4}$$

The right-hand side of Equation 2.4 is the sum of N valid lower bounds, each for a different posterior q_i and observation y_i; since linearity preserves the bound, their sum is also a valid lower bound. Normalising by N and adding the auxiliary range loss $\mathscr{L}_{\text{bound}}$ gives the per-rater ELBO objective of Equation 3.5 (Chapter 3). No approximation beyond the standard variational one is needed; the per-rater ELBO follows from ELBO additivity without any additional modelling assumption.

The $\frac{1}{N}$ normalisation has a practical consequence that follows directly from the term count. Without it, N KL terms each weighted 1 exert $4\times$ the regularisation pressure of the baseline's single KL term, which would over-regularise the posteriors toward the prior and collapse the latent space diversity GED requires. Dividing by N keeps the total KL pressure equal to the baseline — each rater's posterior pulls the shared prior with the same weight as in the original formulation.

2.5 Gradient Conflict in Multi-Task Learning

Each annotator's reconstruction loss can be treated as a separate task competing over the shared encoder — the same structural situation that motivated gradient conflict

research in multi-task learning. Multi-task learning's findings about shared-encoder gradient interference translate directly to the shared posterior setting and provide both the diagnostic tools and the architectural vocabulary for addressing it.

2.5.1 Gradient Surgery and the Interference Problem

Yu et al. [19] showed that a shared encoder trained on tasks with opposing gradients receives a net update that is less informative for each individual task than a task-specific update would be. The per-task performance loss tracks the cosine similarity between task gradients. The more conflicted the gradients, the worse each task learns. Their practical contribution was a gradient projection fix: modifying each task's gradient to remove the component that conflicts with the other tasks before the shared encoder update is applied. What matters for our purposes is diagnostic framework: gradient conflict in shared encoders is measurable and its magnitude predicts how much each task suffers.

2.5.2 Multi-Rater Segmentation as Multi-Task Learning

In the D-Persona Stage 1 setting, the shared posterior encoder processes all four annotators' labels simultaneously, which maps directly onto the multi-task learning structure that motivated Gradient Surgery. The reconstruction loss for the i-th annotator is one task, and the four tasks share the single posterior encoder as their joint feature extractor. When two annotators disagree on a boundary, the reconstruction gradients from their respective tasks point in opposite directions in the latent space, creating exactly the conflict that Gradient Surgery was designed to mitigate.

Per-rater posterior encoders are the multi-task analogue of task-specific feature extractors: each annotator's task gets a dedicated encoder, and gradient conflict is eliminated because no two tasks share the same encoder parameters. The decoder (FComb) remains shared, but at the decoder level the gradients are N separate signals arriving from independent encoders, not N conflicting signals mixed inside a shared encoder.

Per-rater posteriors apply this principle to probabilistic multi-rater segmentation, where the variational framework and ELBO additivity provide the mathematical justification for treating each annotator as an independent task.

The connection also makes the solution's failure mode predictable. In multi-task learning, task-specific encoders are most valuable when tasks conflict. When tasks align, a shared encoder can be more data-efficient because all tasks reinforce the same representation. By analogy, per-rater posteriors should provide the largest advantage when annotators disagree most: under annotation sparsity (where absent annotators' zero-channel gradients introduce a particularly severe form of conflict) and for nodules

with ill-defined margins (where annotation disagreement is structurally highest). Chapter 4 reports both effects, with the sparsity advantage growing monotonically from $+11.5\%$ at three annotators to $+21.4\%$ at one.

2.6 The LIDC-IDRI Dataset and Evaluation Metrics

2.6.1 LIDC-IDRI

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [1] is the standard benchmark for multi-rater lung nodule segmentation. It contains 1,018 low-dose CT scans from seven academic medical centres, each read by four thoracic radiologists under a two-phase protocol: an initial blinded read where each radiologist marked lesions independently, followed by a revision pass in which the four readers could see each other's markings before locking their final contours. Because every radiologist had access to the others' boundaries during the revision phase, contours that still differ after revision reflect deliberate interpretive choices rather than isolated reading errors.

Armato et al. [2] quantified the surviving disagreement and found it large enough, and consistent enough across the full scan collection, to rule out fatigue or ambiguous instructions as explanations. What persists is the tissue itself: density gradients at a pulmonary nodule's periphery do not resolve to a single edge in CT, and four readers trained on the same guidelines can still place four defensible contours at different locations on the same case.

Following D-Persona's preprocessing protocol [17], cases with fewer than four complete rater annotations are excluded, and the remaining cases are cropped to 128×128 patches centred on each annotated nodule, yielding 1,609 patches. Each patch has four binary segmentation masks, one per radiologist, and nine per-radiologist clinical attribute ratings (malignancy, texture, spiculation, lobulation, margin, sphericity, calcification, internalStructure, subtlety) on an integer scale of 1–5. These attribute ratings are used in the annotation disagreement characterisation analysis (§4.5).

The dataset is split into four cross-validation folds following D-Persona's exact partitions, with test fold sizes of 450, 375, 412, and 372 patches respectively. All methods are trained and evaluated on identical splits; all comparisons are therefore direct.

2.6.2 NPC-170

The NPC-170 dataset [17] provides a cross-dataset validation context. It contains 170 nasopharyngeal carcinoma MRI cases, each annotated by four independent annotators

on 3-channel MRI (T1, T1CE, T2 sequences). The training set comprises 2,405 slices; validation and test sets each contain 20 cases. NPC-170 differs from LIDC-IDRI in modality (MRI vs CT), anatomy (head and neck vs thorax), and imaging protocol (multi-channel vs single-channel). Using it as a secondary validation tests whether the per-rater design transfers across these differences.

2.6.3 Generalised Energy Distance

The Generalised Energy Distance (GED) [11] is the primary evaluation metric throughout this thesis. Its value as a metric comes from what it jointly rewards: a model must produce predictions that are both geometrically close to individual annotator masks and spread across the range those annotators cover. Collapsing to one prediction, no matter how accurate, drives the within-sample diversity term to zero and GED climbs; spreading samples widely without tracking any rater's boundary is equally penalised. GED reaches zero only when the model's sampling distribution and the annotation distribution are identical. The formal definition is in §3.1.

2.7 Summary of Research Gaps

Each of the four literature strands above advanced multi-rater segmentation in one direction while leaving a specific gap open — gap that this thesis addresses.

Every probabilistic multi-rater segmentation model reviewed in §2.2 and §2.3 trains with a shared posterior that encodes all annotators together. The reconstruction gradient at the shared latent code is a sum of per-rater terms; when annotators disagree, these partially cancel and $\mathbf{z}$ drifts toward a compromise that fits no individual style. Giving each annotator a dedicated posterior encoder — one that receives only that annotator's mask and therefore carries only that annotator's gradient — is a design the literature reviewed here leaves unexplored.

Building on this gap, D-Persona's Stage 2 projection heads attempt to personalise a shared $\mathbf{z}$ by steering it toward individual rater styles at inference time. Projection heads can only redistribute structure already in the latent code. Rater-specific information that Stage 1's gradient averaging removed is gone before Stage 2 runs. The ablation in §4.2.3 confirms the consequence: adding Stage 2 on top of a per-rater Stage 1 — where the latent code already has clean rater-specific gradients — actively worsens GED.

A third gap concerns annotation coverage. Real clinical datasets rarely have every image annotated by all radiologists, yet every probabilistic multi-rater method in this literature is evaluated under complete coverage. How absent annotators degrade a shared posterior — specifically, zero-channel inputs generating empty-mask gradients that contaminate $\mathbf{z}$ — has received no systematic treatment.

Finally, LIDC-IDRI provides nine clinical attribute ratings per nodule per radiologist, but no systematic analysis has asked which attributes drive inter-rater boundary disagreement. Such an analysis would show where uncertainty-aware methods matter most and where a deterministic model is adequate. §3.7 takes this up, with results in §4.5.

turnitin Page 24 of 68 - AI Writing Submission Submission ID trn:oid:::27535:140806921

CHAPTER 3

METHODOLOGY

Understanding why per-rater posteriors work requires first seeing what goes wrong inside the shared baseline during training. The gradient conflict analysis in §3.2 establishes the baseline bottleneck; §3.4 then derives the per-rater ELBO, proves the gradient isolation property, and specifies the architecture. Sparse annotation training (§3.5) and the two diagnostic analyses — gradient alignment (§3.6) and nodule attribute characterisation (§3.7) — follow from the same structural difference between the two designs.

3.1 Problem Formulation

Let $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ denote an input medical image, and let $\{y_i\}_{i=1}^{N}$ be the binary segmentation masks provided by N independent annotators, each $y_i \in \{0,1\}^{H\times W}$. For LIDC-IDRI, $H = W = 128$, $C = 1$ (grayscale CT), and $N = 4$. For NPC-170, $C = 3$ (T1, T1CE, T2 MRI channels) with the same N.

The goal is to learn a model that, given $\mathbf{x}$ at test time, can generate a set of diverse segmentation predictions that reflect the true spread of annotation disagreement: not one averaged prediction, and not N identical samples. Formally, we want a learned distribution $p(\hat{y} \mid \mathbf{x})$ from which samples span the range of plausible annotator decisions.

3.1.1 The Generalised Energy Distance

We adopt the Generalised Energy Distance (GED) [11] as the primary metric because it formalises this joint requirement in a single score. For a set of S model predictions $\{\hat{y}_s\}$ and a set of N ground-truth annotations $\{y_i\}$:

$$\text{GED} = 2\,\mathbb{E}\big[d(\hat{y},y)\big] - \mathbb{E}\big[d(\hat{y},\hat{y}')\big] - \mathbb{E}\big[d(y,y')\big] \tag{3.1}$$

where $d(\cdot,\cdot)$ is a ground distance (here, $1-\text{IoU}$), and the expectations are over independently drawn samples from the model distribution and the annotation set respectively.

GED equals zero only when the model distribution and the annotation distribution are identical. The second term penalises a model that collapses to a single prediction: if all $\hat{y}_s$ are identical, $\mathbb{E}[d(\hat{y},\hat{y}')] = 0$, and GED reduces to twice the average prediction error. A model can achieve low prediction error on each individual rater mask and still score poorly on GED if its predictions are insufficiently diverse.

Diversity and accuracy are ordinarily at odds, since a model that improves GED by producing more diverse predictions would normally sacrifice per-rater accuracy ($\text{Dice}_{\text{match}}$). Per-rater posteriors improve both simultaneously, as the results in Chapter 4 show — an outcome that points to better gradient signal quality during training rather than a diversity-accuracy trade-off, since the latter would move the two metrics in opposite directions.

3.1.2 Probabilistic Formulation

We follow the variational framework of the Probabilistic U-Net [11]. A latent variable $\mathbf{z} \in \mathbb{R}^D$ ($D = 6$ throughout) captures annotation uncertainty. An image-conditioned prior $p(\mathbf{z} \mid \mathbf{x})$ defines the sampling distribution at test time: drawing $\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})$ and passing it through a decoder $f_{\text{comb}}(\mathbf{z}, \mathbf{x})$ produces one predicted mask. Diversity over multiple draws then reflects uncertainty captured in $p(\mathbf{z} \mid \mathbf{x})$. Training requires a posterior $q(\mathbf{z} \mid \mathbf{x}, \text{annotation})$ that pushes $p(\mathbf{z} \mid \mathbf{x})$ toward the annotation distribution. The design of this posterior is the crux of both the baseline and the proposed method. Table 3.1 collects the symbols and variables that appear from this point on.

Table 3.1 Notation index for Chapters 3 and 4.

Symbol	Meaning
$\mathbf{x}$	Input image
y_i	Segmentation mask from annotator i
Y	All annotation masks $\{y_1, \ldots, y_N\}$
N	Number of annotators (4 throughout)
$\mathbf{z}$, $\mathbf{z}_i$	Shared / per-rater latent variable
D	Latent dimension ($D = 6$)
$q(\mathbf{z} \mid \mathbf{x}, Y)$	Shared posterior (baseline)
$q_i(\mathbf{z} \mid \mathbf{x}, y_i)$	Per-rater posterior for annotator i (proposed)
$p(\mathbf{z} \mid \mathbf{x})$	Image-conditioned prior
f_{comb}	Segmentation decoder (FComb)
β	KL weighting coefficient ($\beta = 0.5$)
$\mathcal{L}_{\text{bound}}$	Auxiliary range loss (D-Persona)
np	Number of annotators present during training

3.2 Baseline: D-Persona with a Shared Posterior

D-Persona [17] is the baseline. Among published methods for per-rater personalised probabilistic segmentation it achieves the best reported GED and $\text{Dice}_{\text{match}}$ on LIDC-IDRI, and its two-stage architecture contains the gradient conflict that the per-rater posterior is designed to remove.

3.2.1 Architecture

D-Persona builds on the Probabilistic U-Net [11]. The backbone is a ResNet34 encoder [4] that extracts image features. The prior network is an AxisAlignedConvGaussian that takes the image $\mathbf{x}$ alone and outputs a diagonal Gaussian $p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mu_p, \text{diag}(\sigma_p^2))$. The posterior network has the same architecture but takes a concatenation of the image and *all four* rater masks as input (a 5-channel tensor (x, y_1, y_2, y_3, y_4)). It outputs a shared posterior $q(\mathbf{z} \mid \mathbf{x}, Y) = \mathcal{N}(\mu_q, \text{diag}(\sigma_q^2))$.

A lightweight FComb decoder [11] takes two inputs: the U-Net's spatial feature maps (passed through skip connections) and a single sample drawn from the latent distribution. The spatial features supply local boundary detail; the latent sample globally biases which boundary decision the decoder resolves to. This is Stage 1. Stage 2 adds four per-rater projection heads $\{h_i\}_{i=1}^{N}$, each a small MLP that takes the shared $\mathbf{z}$ and produces a rater-specific style offset, which is then added to $\mathbf{z}$ before decoding. Stage 1 parameters are frozen before Stage 2 begins; the projection heads operate on whatever $\mathbf{z}$ Stage 1 produced, with no path back to the Stage 1 training objective.

3.2.2 The Stage 1 Training Objective

Stage 1 is trained with the following evidence lower bound (ELBO) [9]:

$$\begin{aligned} \mathcal{L}_{\text{base}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x},Y)}\big[\log p(Y \mid \mathbf{z}, \mathbf{x})\big] \\ - \text{KL}\big(q(\mathbf{z}|\mathbf{x},Y) \,\|\, p(\mathbf{z}|\mathbf{x})\big) + \beta\, \mathcal{L}_{\text{bound}}. \end{aligned} \tag{3.2}$$

The first term is the reconstruction likelihood: the posterior sample $\mathbf{z}$ must decode to all four rater masks well. The second term is the KL divergence that pulls the posterior toward the prior. This keeps the latent space coherent for test-time sampling. The third term, $\mathcal{L}_{\text{bound}}$, is an auxiliary diversity floor loss from D-Persona. It operates by drawing samples from the prior at training time and applying a hinge penalty whenever the spread of the decoded outputs falls below the observed spread of the four rater masks. Unlike the reconstruction term, it acts on prior samples rather than posterior samples, shaping the model's test-time output distribution. We retain this term unchanged with $\beta = 0.5$.

3.2.3 The Gradient Conflict Problem

The reconstruction likelihood in Equation 3.2 requires the single posterior sample $\mathbf{z}$ to simultaneously explain all four rater masks. Backpropagating through the first term gives the gradient at $\mathbf{z}$:

$$\frac{\partial \mathcal{L}_{\text{base}}}{\partial \mathbf{z}} = \sum_{i=1}^{N} \frac{\partial \log p(y_i \mid \mathbf{z}, \mathbf{x})}{\partial \mathbf{z}}. \tag{3.3}$$

This gradient is a sum of N per-rater terms. When all four annotators agree, these terms reinforce each other. But when they disagree on a boundary, as they routinely do at ill-defined nodule margins in LIDC-IDRI, the per-rater gradients point in different directions. For a pixel that rater i labels as lesion and rater j labels as background, the gradient $\partial \log p(y_i \mid \mathbf{z}, \mathbf{x})/\partial \mathbf{z}$ pushes $\mathbf{z}$ toward predicting 1 for that pixel, while $\partial \log p(y_j \mid \mathbf{z}, \mathbf{x})/\partial \mathbf{z}$ pushes it toward 0. In the extreme case of complete disagreement ($y_i = 1 - y_j$ everywhere), the two terms cancel exactly and the net gradient at $\mathbf{z}$ is zero.

In practice the cancellation is partial, not total, and it is case-dependent. At full annotation, we measure the mean pairwise cosine similarity between per-rater reconstruction gradients at $\mathbf{z}$: it is 0.167, distributed across test cases with a within-fold standard deviation of 0.439 (meaning the conflict is much worse for some cases than others). At full annotation the shared posterior still receives a net signal — the conflict does not zero out the gradient entirely — but the latent space degrades: rather than encoding the distribution of annotation styles, the shared $\mathbf{z}$ is pulled toward a compromise that fits no individual rater well.

3.2.4 Why Stage 2 Cannot Fix This

Stage 2 projection heads h_i steer samples from the shared $\mathbf{z}$ toward individual annotator styles by applying a small rater-specific offset before decoding. For this to work, Stage 1 must encode per-rater structure into $\mathbf{z}$ in the first place. Whatever per-rater information gradient averaging removed during Stage 1's backward pass is absent from $\mathbf{z}$ before h_i ever executes; the projection heads operate on the gradient-averaged residual, not on the original rater signals.

The ablation study in §4.2.3 makes this failure mode concrete. When Stage 2 is applied on top of a Stage 1 trained with per-rater posteriors — where $\mathbf{z}$ already carries clean, rater-specific gradient signals — GED rises from 0.1444 to 0.1836, a 27% degradation. Stage 2 disrupts the diversity that Stage 1 already encoded, because it was designed to compensate for a deficit that no longer exists.

3.3 Transformer Encoder Investigation

The gradient conflict analysis in §3.2.3 establishes that the shared posterior's training signal is a sum of conflicting per-rater gradients. Before concluding that the fix must operate at the training objective level, a simpler alternative must be ruled out: that the bottleneck is *representational capacity*. ResNet34 was designed for image classification and its convolutional structure constrains how far apart two spatial locations can interact within one forward pass. If the gradient blurring is in fact a capacity problem (a representation too limited to encode four raters' styles simultaneously), then a more expressive backbone should alleviate it regardless of how the training objective is structured. The experiment tests whether an architectural fix suffices or whether the training objective itself must change.

We test this hypothesis by replacing the ResNet34 backbone in the shared posterior encoder with MiT-B2 [18], a Mix Transformer with a substantially larger parameter count and attention-based feature extraction. The encoder architecture, input channels, and training objective are otherwise unchanged.

The result, reported in §4.2.3 (Table 4.4, Row 2), is unambiguous. MiT-B2 achieves GED 0.1531, marginally *worse* than the ResNet34 baseline at 0.1507, with no change in $\text{Dice}_{\text{match}}$. Attention mechanisms and long-range spatial reasoning do not alleviate the gradient conflict problem. A more expressive encoder learns a more expressive weighted average of conflicting rater signals; the conflict itself is a property of the training objective, not of the encoder's receptive field. Both encoders receive the same 5-channel shared-posterior input and the same sum of per-rater reconstruction gradients — the training signal reaching each backbone is structurally the same, whatever the backbone's capacity for representing it.

3.4 Per-Rater Posterior Encoders

Instead of one shared encoder receiving all four masks simultaneously, each annotator gets a dedicated posterior encoder conditioned only on that annotator's mask.

3.4.1 Design

We replace the single shared posterior with N independent encoders, each an AxisAlignedConvGaussian identical in architecture to the original:

$$q_i(\mathbf{z} \mid \mathbf{x}, y_i) = \mathcal{N}\big(\mu_i, \text{diag}(\sigma_i^2)\big), \quad i = 1, \ldots, N. \tag{3.4}$$

Each encoder takes a 2-channel input: the image $\mathbf{x}$ and rater i's mask y_i. The input width drops from 5 channels (baseline) to 2 channels per encoder. No weights are shared

between encoders. The prior $p(\mathbf{z} \mid \mathbf{x})$ and the FComb decoder f_{comb} are identical to the baseline: the only change is in how the posterior signals are collected during training.

At test time, there are no posterior encoders in the loop. Predictions are drawn by sampling $\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})$ and passing through $f_{\text{comb}}(\mathbf{z}, \mathbf{x})$, exactly as in the baseline. The inference procedure is unchanged, and inference cost is identical.

What separates the two architectures is where in the computation graph the four raters' gradients first meet. In the baseline they meet inside the shared encoder, before any rater-specific information can be preserved; in the proposed design they meet only at the decoder, after each encoder has independently updated its representation from single rater's mask. Figure 3.1 shows both paths side by side.

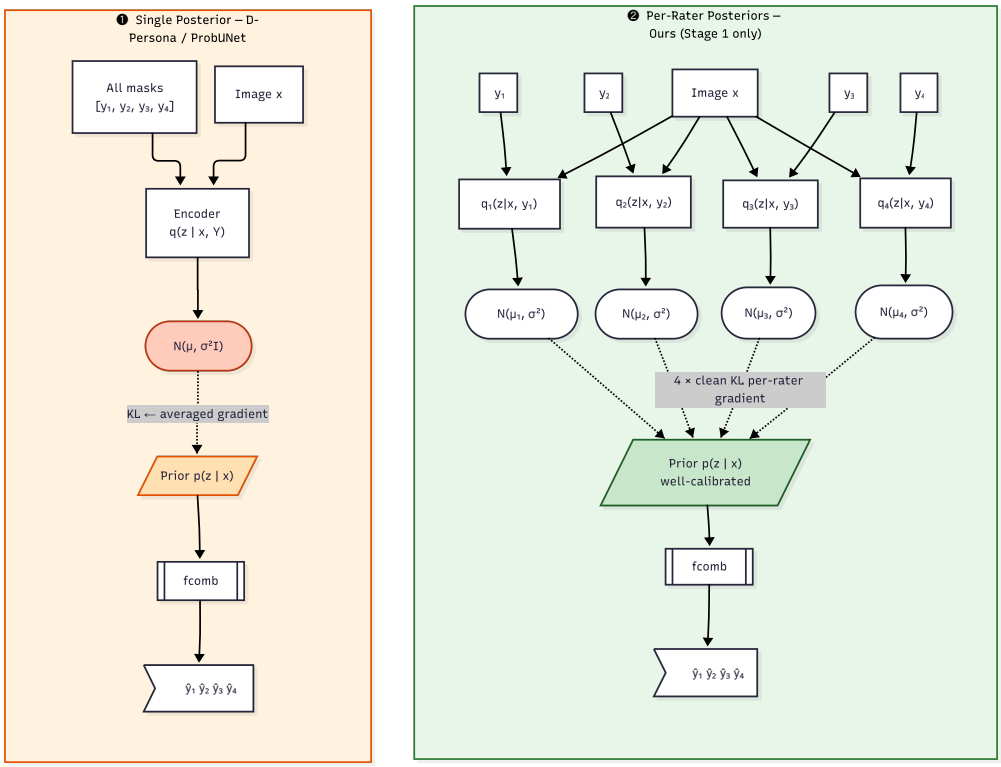

Figure 3.1 Architecture: shared posterior baseline (left) vs. per-rater design (right).

3.4.2 Training Objective

With N independent posteriors, the joint training objective decomposes into N separate ELBO terms:

$$\begin{aligned}\mathscr{L}_{\text{ours}} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{q_i}\big[\log p(y_i \mid \mathbf{z},\mathbf{x})\big] \\ &-\frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}\big(q_i \,\|\, p(\mathbf{z}|\mathbf{x})\big) + \beta\,\mathscr{L}_{\text{bound}}.\end{aligned} \tag{3.5}$$

Each summand is a standard ELBO for rater i. Because a sum of valid lower bounds is itself a valid lower bound (by linearity), $\mathscr{L}_{\text{ours}}$ is a proper variational lower bound on $\sum_{i=1}^{N}\log p(y_i \mid \mathbf{x})$.

The $\frac{1}{N}$ normalisation matters practically. Without it, the N KL terms together exert N times the regularisation pressure of the baseline's single KL term, which would over-regularise the posteriors toward the prior and collapse the latent space diversity that GED requires. Dividing by N keeps the total KL pressure equal to the baseline — each rater's posterior pulls the shared prior with the same weight as the single posterior in the original formulation, and no single annotator's distribution dominates.

3.4.3 Gradient Isolation

With N independent encoders, per-rater gradients cannot reach each other's parameters.

Proposition. *For the per-rater ELBO $\mathscr{L}_{\text{ours}}$, the gradient of the i-th reconstruction term with respect to the j-th latent variable is zero for all $i \neq j$: $\partial\mathscr{L}_i/\partial\mathbf{z}_j = 0$.*

Proof. Let $\mathscr{L}_i = \mathbb{E}_{q_i(\mathbf{z}_i|\mathbf{x},y_i)}[\log p(y_i \mid \mathbf{z}_i,\mathbf{x})]$. The expectation is taken over $\mathbf{z}_i \sim q_i$, which depends only on $(\mathbf{x},y_i)$. Since q_i does not depend on q_j for $j \neq i$, and $\mathbf{z}_i$ is drawn independently from $\mathbf{z}_j$, the term $\mathscr{L}_i$ is a function only of the parameters of q_i. By the chain rule, $\partial\mathscr{L}_i/\partial\mathbf{z}_j = 0$ for all $j \neq i$. □

The reconstruction gradient that reaches encoder q_i reflects rater i's mask exclusively, a structural consequence of the objective decomposition with no regularisation term required to enforce it. What rater j labelled has no effect on how encoder q_i is trained. The FComb decoder still receives the gradients from all N reconstruction terms, but they arrive as N separate, rater-specific signals — each encoder has already processed its own rater's mask independently before anything reaches the decoder. The decoder learns to reconcile them through usual gradient accumulation that any multi-task network experiences, but each signal is clean and rater-specific when it arrives.

To see why this matters, contrast with Equation 3.3. In the baseline, the N rater gradients are summed before reaching any part of the encoder; in the proposed objective, they are summed only at the decoder, after each encoder has been updated with its own rater's signal. What changes between the two designs is what information the encoder is trained to represent, not how the decoder processes it afterward.

3.4.4 Architectural Details and Implementation

The implementation adds $N - 1 = 3$ additional AxisAlignedConvGaussian encoder networks to the D-Persona Stage 1 architecture. Each has same convolutional structure as the original posterior encoder and is initialised independently. No weight tying, no shared convolutional layer. The prior network and FComb decoder are taken from the D-Persona codebase without modification. Stage 2 projection heads are not used.

Training follows the D-Persona hyperparameters exactly to ensure fair comparison: Adam optimiser [8] with learning rate 10^{-4}, cosine annealing schedule, batch size 12, latent dimension $D = 6$, and 100 training epochs per fold. Four-fold cross-validation is used on LIDC-IDRI, with fold test sizes of 450, 375, 412, and 372 nodule patches respectively.

The additional encoders increase training time significantly. Stage 1 with per-rater posteriors takes approximately 47 hours across all four folds on Apple MPS hardware (4.5 times the baseline Stage 1 training time of approximately 10.5 hours, and 2.5 times the full D-Persona two-stage pipeline of approximately 18.8 hours). We acknowledge this as a genuine practical barrier, particularly for institutions without GPU access. Inference cost, however, is identical to the baseline: one sample from $p(\mathbf{z} \mid \mathbf{x})$ and one decoder forward pass.

3.5 Sparse Annotation Training

The experiments in Chapter 4 include a systematic study of how both models behave when annotation coverage is incomplete.

3.5.1 Motivation

In clinical datasets, it is common for different radiologists to annotate different subsets of cases. The LIDC-IDRI dataset used for the main experiments enforces complete four-rater annotation, but this is atypical of real deployment conditions. A model that relies on complete annotation coverage would be fragile in practice. Most large clinical collections do not have every image annotated by all radiologists. The sparse annotation experiments ask whether per-rater posteriors handle this condition better than the shared

baseline.

The sparsity level is controlled by a parameter $\text{np} \in \{1,2,3\}$ specifying how many annotators provide masks for each training image at each training step. Full annotation corresponds to $\text{np} = 4$ and is used as the reference point. In each training step, $N - \text{np}$ annotators are chosen at random and treated as absent.

3.5.2 The `drop_raters()` Mechanism and Its Asymmetric Effect

Absent annotators are handled asymmetrically by the two architectures, and this asymmetry drives the sparsity results.

In the shared baseline, an absent annotator's mask channel is set to zero. At $\text{np} = 2$ with annotators a and b present, the 5-channel input is $(x, y_a, y_b, 0, 0)$. The encoder processes all five channels — including the two zeros — and the reconstruction gradient from the zero channels pushes $\mathbf{z}$ toward predicting empty segmentations for the absent raters. At $\text{np} = 1$, three of the four gradient contributions point toward empty-mask predictions and only one carries real annotation signal. The latent code is simultaneously pulled toward the present annotator's boundary decision and toward predicting nothing on three channels that carry no real information.

In the per-rater model, if rater j is absent for a given training step, encoder q_j does not run. There is no zero-channel input and no loss term for that rater. The gradient for that step comes only from the np present encoders, each providing a clean signal from its own rater's mask, and absent encoders contribute nothing: no zero-mask contamination and no competing gradient. This follows structurally from the per-encoder architecture.

3.5.3 Training Configuration for Sparse Experiments

Sparse model are trained for 300 epochs rather than the 100 used at full annotation. Fewer annotators per step means fewer gradient updates per epoch that carry meaningful segmentation signal; 300 epochs compensate for this reduced information density. All other hyperparameters (learning rate, cosine schedule, batch size, latent dimension) remain identical to the full-annotation setting.

Evaluation is always performed with all four annotators' masks, regardless of the training sparsity level. This ensures that sparse-training GED values are measured on the same distribution as full-annotation GED values and can be compared on the same scale. Results at $\text{np} = 4$ with 300-epoch sparse training are excluded from the main comparisons: the shared baseline degrades at 300 epochs with full annotation (winning only one of four folds), which is a training-artefact confound introduced by the longer schedule, not a genuine phenomenon of the per-rater design. We compare

the 100-epoch full-annotation models as the reference and the 300-epoch sparse models at np $\in \{1, 2, 3\}$ as the sparsity conditions.

3.6 Gradient Alignment Measurement

The gradient conflict hypothesis predicts that as annotators are removed during training, the gradients at the shared $\mathbf{z}$ become more aligned. The mechanism is zero-channel dominance: absent annotators each contribute an empty-mask gradient, and as their count grows these degenerate signals increasingly pull $\mathbf{z}$ in the same direction. To measure whether this collapse is detectable, we compute a scalar diagnostic index for each fold.

3.6.1 Measurement Procedure

For each test image $\mathbf{x}$ with four rater masks $\{y_i\}_{i=1}^{4}$, we perform the following:

1. Forward pass through the trained baseline to obtain the shared posterior mean $\bar{\mathbf{z}} = \mu_q(\mathbf{x}, Y)$.
2. For each rater i, compute the per-rater reconstruction gradient: $\mathbf{g}_i = \partial \log p(y_i \mid \bar{\mathbf{z}}, \mathbf{x}) / \partial \bar{\mathbf{z}}$.
3. Compute pairwise cosine similarity across all $\binom{4}{2} = 6$ pairs: $\cos(\mathbf{g}_i, \mathbf{g}_j) = (\mathbf{g}_i \cdot \mathbf{g}_j)/(\|\mathbf{g}_i\|\|\mathbf{g}_j\|)$.
4. Average across pairs and across all test cases in the fold (100 cases sampled per fold).

The resulting scalar summarises how aligned the per-rater reconstruction gradients are at the shared posterior mean. A value near 0 indicates independent gradients pointing in different directions in the 6-dimensional latent space. A value near 1 indicates all gradients have collapsed to approximately the same direction. The latent code receives the same update signal regardless of which annotator it tries to represent.

3.6.2 Why Per-Rater Alignment Is Zero by Construction

There is no shared posterior mean in the per-rater model. Each encoder has its own $\mathbf{z}_i$, and by Proposition 3.4.3, $\partial \mathcal{L}_i / \partial \mathbf{z}_j = 0$ for $i \neq j$. The concept of cross-rater gradient alignment at a shared latent point does not exist in the per-rater architecture. The measured alignment for per-rater posteriors is 0.000 at all sparsity levels, by construction rather than by measurement.

3.6.3 What the Diagnostic Measures, and What It Does Not

The alignment index is a diagnostic: it measures whether degeneration has occurred, not why. The mechanism is zero-mask channel dominance in the shared encoder (§3.5.2). As fewer annotators are present, zero channels constitute a larger fraction of the 5-channel input, and their gradients increasingly dominate the update at $\mathbf{z}$. When $\text{np} = 1$, the single present annotator provides one gradient, and the three absent annotators each contribute a gradient directed toward predicting an empty mask. Three of the four gradient directions point the same way: all pushing the latent code toward predicting empty masks for absent annotators. The one real gradient from the present rater is outnumbered three to one. The pairwise cosine similarity across all six pairs therefore rises sharply, reaching 0.976 at $\text{np} = 1$.

The alignment rise and the GED gap both grow monotonically as np decreases. This co-movement is not coincidental; both are downstream of same mechanism. The alignment is a symptom of gradient collapse, and the GED degradation is a consequence of the same collapse degrading latent space quality — two observed effects of zero-mask channel contamination rather than a direct causal relationship between the two metrics. Stating that "gradient collapse causes higher GED" would therefore overstate what the diagnostic can show.

3.7 Attribute Characterisation of Annotation Disagreement

The final methodological component is an analysis that is independent of our model entirely. LIDC-IDRI provides not only segmentation masks but also nine per-rater per-nodule attribute ratings: malignancy, texture, spiculation, lobulation, margin, sphericity, calcification, internalStructure, and subtlety, each on an integer scale. We use these ratings to ask which nodule attributes most strongly predict how much the four radiologists will disagree about a nodule's boundary.

The analysis uses only the LIDC-IDRI ground-truth annotations; neither model is involved. Clinically, this identifies the nodule types where annotation uncertainty is structurally highest and uncertainty-aware methods are most relevant.

3.7.1 Setup

For each nodule case k, we compute the inter-rater attribute standard deviation across the four radiologists' ratings, $\sigma_{\text{attr}}^{(k)}$, and the inter-rater mask variance $\sigma_{\text{mask}}^{(k)}$ (variance in binary segmentation decisions across the four masks). We then compute the Pearson correlation coefficient r between these two quantities across all cases with complete attribute records.

Out of the 1,609 LIDC-IDRI nodule patches used in the main experiments, 1,603 have complete attribute data for all nine attributes and all four annotators. The remaining 6 cases are excluded from this analysis. The correlation is computed independently within each of the four cross-validation folds to verify that the result is not a data-split artefact.

3.7.2 What Was Attempted and Dropped

Two further analyses were attempted during development and both had to be abandoned.

Analysis A (GED by nodule margin quartile) stratified the test-set GED comparison by the inter-rater margin disagreement quartile, aiming to show that per-rater posteriors improve most on the most ambiguous nodules. This analysis was dropped for two reasons. First, the third quartile (Q3, moderately ambiguous nodules) showed a -5.0% reversal where the baseline outperformed per-rater, directly contradicting the intended narrative. Second, the GED formula in the stratification script used mean squared error rather than the IoU-based formulation of Equation 3.1. The computed values (0.001–0.004) are incomparable with the main GED results (0.14–0.22). The analysis is not included anywhere in this thesis.

Analysis C attempted to demonstrate that the GED improvement from per-rater posteriors is concentrated on clinically ambiguous nodules, specifically that the per-rater advantage correlates with per-attribute inter-rater disagreement. The Pearson r between per-rater GED improvement and inter-rater attribute disagreement was: subtlety $r = -0.042$ ($p = 0.093$), malignancy $r = -0.023$ ($p = 0.356$), spiculation $r = 0.035$ ($p = 0.157$). None of these is significant at $p < 0.05$. The claim has no statistical support with the available data and is reported in the Limitations section (§4.7). It is not treated as a finding.

3.8 Summary of Design Decisions

Table 3.2 captures every training-time difference between the two architectures. Both models use an identical inference procedure — sampling from $p(\mathbf{z} \mid \mathbf{x})$ and decoding — so any performance difference in Chapter 4 traces back to the training-time change rather than to differences in sampling or decoding strategy.

Table 3.2 Design decisions: shared baseline vs. per-rater.

Design aspect	**Shared baseline**	**Per-rater (proposed)**
Posterior encoders	1 shared encoder	$N = 4$ independent encoders
Encoder input	$N + 1 = 5$ channels	2 channels (image + 1 mask)
Training objective	Single joint ELBO	Sum of N per-rater ELBOs
Stage 2	Used	Not used
Absent rater handling	Zero-channel input	Encoder not executed
Gradient at posterior	Averaged over N raters	Pure single-rater signal
Inference	Sample from $p(\mathbf{z}\|\mathbf{x})$	Identical
Training time	$\sim$10.5 hours (4 folds)	$\sim$47 hours (4 folds)

The proposed method makes one change to the training procedure, giving each annotator a dedicated posterior encoder. Everything else (prior, decoder, loss weighting, hyperparameters, inference) is held constant. This isolation allows any performance difference in Chapter 4 to be attributed to the per-rater encoder design rather than to differences in optimiser, backbone, or loss function.

CHAPTER 4

RESULTS AND DISCUSSION

Full-annotation comparisons and the ablation are in §4.2; the sparsity results, which show the larger performance gap, are in §4.3. §4.5 reports the attribute correlation analysis, which uses only LIDC-IDRI ground-truth masks and attribute ratings and is independent of both segmentation models.

4.1 Experimental Setup

4.1.1 Datasets

LIDC-IDRI [1]: Dataset details and preprocessing are described in §2.6. Briefly: 1,609 nodule patches across four cross-validation folds (test sizes 450, 375, 412, 372), each patch carrying four independent radiologist masks and nine per-rater clinical attribute ratings. All experiments are trained and evaluated independently on each fold; results are 4-fold means $\pm$ standard deviation unless stated otherwise.

NPC-170 [17]: 170 nasopharyngeal carcinoma MRI cases, each annotated by four annotators. The input is 3-channel (T1, T1CE, T2 MRI). Training uses 2,405 slices, with 20 validation and 20 test cases on a single train/test split. Three random seeds are used to account for split variance. The NPC-170 experiment is a cross-dataset check on whether the per-rater approach transfers to a different modality and anatomy.

4.1.2 Evaluation Metrics

Generalised Energy Distance (GED $\downarrow$) [11]: defined formally in Equation 3.1. Rewards predictions that are both geometrically close to individual rater masks and spread across the range those raters cover; lower is better. The formal properties and training implications are discussed in §3.1.

Dice$_{\text{match}}$ ($\uparrow$) [17]: Hungarian-matched assignment of model predictions to rater an-

notations, averaged over the four raters. Measures geometric closeness to individual annotator masks.

Dice$_{\text{soft}}$ (↑): Soft Dice between the mean model prediction and the mean rater annotation. Measures average accuracy without accounting for per-rater variation.

4.1.3 Baselines

ProbUNet [11]: trained on the pixel-wise mean of all four rater annotations as a single target. This sets the lower bound: collapsing four annotations into one mean target discards all per-rater structure, and any multi-rater method that fails to beat it substantially has gained nothing from using multiple annotations.

D-Persona Stage 1+2 [17]: the full two-stage pipeline: shared posterior Stage 1 followed by per-rater projection head Stage 2. This is the primary baseline.

Proposed (per-rater, Stage 1 only): four independent posterior encoders, per-rater ELBO, no Stage 2. Stage 1 only.

For the sparsity experiments, the comparison collapses to shared posterior (Stage 1 baseline trained with `drop_raters()`) versus per-rater Stage 1. Stage 2 is not applicable in sparse settings where annotators are routinely absent during training.

4.1.4 Implementation Details

All hyperparameters match D-Persona's published settings to ensure fair comparison: ResNet34 backbone [4], Adam optimiser [8] with learning rate 10^{-4}, cosine annealing, batch size 12, latent dimension $D = 6$. Full-annotation models are trained for 100 epochs per fold. Sparse-annotation models are trained for 300 epochs to compensate for the reduced information density at each step. All experiments run on Apple MPS hardware; the full configuration is given in Table 4.1.

Table 4.1 Hyperparameter configuration for all experiments.

Setting	Value
Backbone	ResNet34
Optimiser	Adam, lr $= 10^{-4}$
LR schedule	Cosine annealing
Latent dimension D	6
Batch size	12
β (KL weight)	0.5
Epochs (full annotation)	100 per fold
Epochs (sparse annotation)	300 per fold
Number of annotators N	4
Training hardware	Apple MPS

4.2 Full-Annotation Performance

4.2.1 Main Comparison on LIDC-IDRI

Table 4.2 lists GED, $\text{Dice}_{\text{match}}$, and $\text{Dice}_{\text{soft}}$ for all three methods. GED and $\text{Dice}_{\text{match}}$ both improve in the per-rater model while $\text{Dice}_{\text{soft}}$ holds at exactly 0.9015, unchanged from D-Persona's value. Diversity and per-rater accuracy improve together, with no loss in average prediction quality.

Table 4.2 LIDC-IDRI segmentation results, 4-fold CV.

Method	GED ↓	$\text{Dice}_{\text{match}}$ ↑	$\text{Dice}_{\text{soft}}$ ↑
ProbUNet [11]	$0.2234_{\pm.0211}$	$0.8836_{\pm.0111}$	$0.8827_{\pm.0135}$
D-Persona (S1+S2) [17]	$0.1507_{\pm.0088}$	$0.8909_{\pm.0037}$	$0.9015_{\pm.0039}$
Ours (per-rater, S1)	$\mathbf{0.1444}_{\pm.0141}$	$\mathbf{0.9112}_{\pm.0061}$	$0.9015_{\pm.0066}$

GED falls from 0.1507 ± 0.0088 (D-Persona) to 0.1444 ± 0.0141 (per-rater), a 4.2% reduction. $\text{Dice}_{\text{match}}$ rises from 0.8909 to 0.9112, a +2.28% gain. Both metrics move in the right direction simultaneously. Models that improve diversity by spreading predictions more widely usually pay for it with worse per-rater geometric accuracy; per-rater posteriors move both metrics in the same direction, which points to improved gradient signal quality rather than a redistribution of the diversity-accuracy budget.

$\text{Dice}_{\text{soft}}$ sits at exactly 0.9015 for both D-Persona and the per-rater model, meaning average prediction quality against the mean annotation is unchanged. The gain is in per-rater calibration: the model approximates individual annotator boundary decisions better, without any shift in centrist prediction.

The per-rater Stage 1 alone outperforms the full D-Persona Stage 1+2

pipeline. Stage 2 adds roughly eight hours of training across four folds and makes GED worse.

ProbUNet (0.2234 GED, 0.8836 $\text{Dice}_{\text{match}}$, 0.8827 $\text{Dice}_{\text{soft}}$) is the lower bound — training on the mean of four masks and generating near-identical samples. The gap between ProbUNet and both D-Persona variants confirms that multi-rater probabilistic training adds real value.

The per-rater model's cross-fold $\text{Dice}_{\text{match}}$ variance (± 0.0061) is wider than D-Persona's (± 0.0037). Some folds contain more nodules with ill-defined margins where gradient conflict is severe and per-rater encoders help more; other folds are more homogeneous. The wider variance tracks this fold-level heterogeneity rather than any instability in the training.

4.2.2 Per-Expert Dice Breakdown

Table 4.3 decomposes the $\text{Dice}_{\text{match}}$ improvement by individual annotator, averaged across all four folds.

Table 4.3 Per-expert $\text{Dice}_{\text{match}}$ breakdown on LIDC-IDRI, 4-fold average.

Method	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**
D-Persona (S1+S2)	0.8816	0.8947	0.8986	0.8888
Ours (per-rater, S1)	0.9062	0.9147	0.9182	0.9058
Improvement	+0.0246	+0.0200	+0.0196	+0.0170

The gain ranges from $+0.0170$ (Expert 4) to $+0.0246$ (Expert 1) across all four annotators; no single rater drives the aggregate while others regress. Each radiologist's predictions are better approximated under per-rater training than under the shared baseline.

Expert 1 gains the most ($+0.0246$). Their boundary style is better captured when their gradients reach the encoder without being pooled with three other raters. Expert 4 shows the smallest gain at $+0.0170$ but still improves, and no single annotator drives the aggregate at the expense of the others.

4.2.3 Ablation Study

Table 4.4 evaluates the D-Persona baseline against six independent design modifications. Each row is a separate standalone experiment, not a cumulative stack. All ablation results use 4-fold cross-validation on LIDC-IDRI.

Table 4.4 Ablation study on LIDC-IDRI, 4-fold CV.

Method	**GED** ↓	**Dice$_{\text{match}}$** ↑
(1) D-Persona baseline (S1+S2)	$0.1507_{\pm.0088}$	$0.8909_{\pm.0037}$
(2) + MiT-B2 backbone	$0.1531_{\pm.0127}$	$0.8909_{\pm.0068}$
(3) + Orthogonality loss	$0.1519_{\pm.0070}$	$0.8890_{\pm.0033}$
(4) + Prior bank ($k = 100$)	$0.2212_{\pm.0055}$	$0.8816_{\pm.0016}$
(5) + Dual diversity loss	$0.1509_{\pm.0070}$	$0.8889_{\pm.0032}$
(6) **Per-rater posteriors (S1)**	$\mathbf{0.1444}_{\pm.0141}$	$\mathbf{0.9112}_{\pm.0061}$
(7) Per-rater + Style vectors (S1+S2)	$0.1836_{\pm.0108}$	$0.8876_{\pm.0049}$

Rows 2 through 5 represent four independent attempts to improve over the baseline before arriving at the per-rater formulation.

Substituting MiT-B2 for ResNet34 (Row 2) tests the architectural capacity hypothesis directly. The attention mechanism in MiT-B2 can relate any two spatial locations regardless of distance — considerably more expressive than a convolutional backbone. GED comes out at 0.1531 versus the baseline's 0.1507, with Dice$_{\text{match}}$ unchanged. A more expressive transformer learns a more expressive average of conflicting rater gradients; the conflict itself is a property of which signals reach the encoder, not how much capacity the encoder has.

Row 3, Orthogonality loss: An explicit loss term penalising cosine similarity between per-rater posteriors (intended to push the shared encoder toward more orthogonal rater representations) marginally worsens both metrics (0.1519 GED, 0.8890 Dice$_{\text{match}}$). The gradient averaging inside the shared encoder happens before the orthogonality loss ever sees the encoder's output — the loss is applied to a representation already shaped by conflicting gradients, so the per-rater structure it is trying to enforce was never encoded in the first place.

Row 4, Prior bank ($k = 100$): Replacing the continuous prior with $k = 100$ discrete k-means centroids produces the worst result in the table. GED reaches 0.2212, worse than ProbUNet. GED rewards diversity from stochastic prior sampling; replacing that with deterministic prototype retrieval eliminates the sample variance the metric is measuring.

Row 5 adds a dual diversity loss penalising pairwise Dice similarity between prior samples, reaching GED 0.1509 — essentially tying the baseline. Diversity pressure applied to the prior cannot create rater-specific structure that Stage 1's gradient averaging never encoded in the first place.

Across all four alternatives, none changes how gradients reach the posterior encoder. That is the consistent reason for failure. Per-rater posteriors (Row 6) are the only modification that improves both GED and Dice$_{\text{match}}$ simultaneously: 0.1444 GED and 0.9112 Dice$_{\text{match}}$ with Stage 1 only.

Row 7 resolves the question of whether Stage 2 can be layered on top of a per-rater Stage 1. GED rises from 0.1444 to 0.1836 — a 27% increase. Stage 2 was designed to personalise a shared latent code that lacks rater-specific structure. Applied after per-rater Stage 1, where each annotator already has a distinct gradient pathway, it introduces shared-posterior dynamics on top of a model that does not need them and collapses the diversity that Stage 1 created.

4.2.4 Cross-Dataset Validation on NPC-170

On NPC-170 the GED gap is 0.0011, which falls within the seed variance of ±0.0085. Table 4.5 reports the full results.

Table 4.5 Cross-dataset validation on NPC-170, 3-seed mean ± std.

Method	**GED** ↓	**Dice$_{\text{match}}$** ↑
Baseline (shared posterior, S1)	0.1824	0.8249
Ours (per-rater, S1)	$0.1813_{\pm.0085}$	$0.8225_{\pm.0008}$

On NPC-170, per-rater posteriors and the shared baseline perform at essentially the same level: GED 0.1813 versus 0.1824, a gap of 0.0011 against a seed variance of ±0.0085.

Two possible explanations are worth considering but cannot be confirmed without further experiments. First, NPC-170 has a single train/test split rather than four-fold cross-validation; the lack of independent validation folds means the baseline GED (0.1824) is itself a point estimate with unknown variance. Any small gap is therefore uninterpretable. Second, NPC-170 has 2,405 training slices across 170 cases (a smaller dataset relative to the complexity of 3-channel MRI input); the per-rater encoders may need more training cases to distinguish four annotators' styles on 3-channel MRI.

What the NPC-170 result does establish is that per-rater posteriors do not degrade on different modality and anatomy. The method transfers without special adaptation to 3-channel MRI and a different disease site, and does not require LIDC-specific tuning.

4.2.5 Qualitative Analysis

The diversity gap is most visible at maximum rater disagreement; two fold 3 cases with the widest annotation spread are shown in Figure 4.1 (four prior samples each). ProbUNet's four samples cluster near-identically in both cases. The model has learned a single average representation and sampling from the prior produces negligible variation (despite 100 prior samples being available), so the within-model distance term in GED

approaches zero and ProbUNet is penalised heavily regardless of how accurate each individual sample is.

Our model's four samples span the range visible across the four rater masks. Where radiologists differ on whether the ground-glass halo around the nodule core should be included, some of our samples include it and some do not. The samples match the annotation spread. The prior is calibrated to produce this range by per-rater training signals.

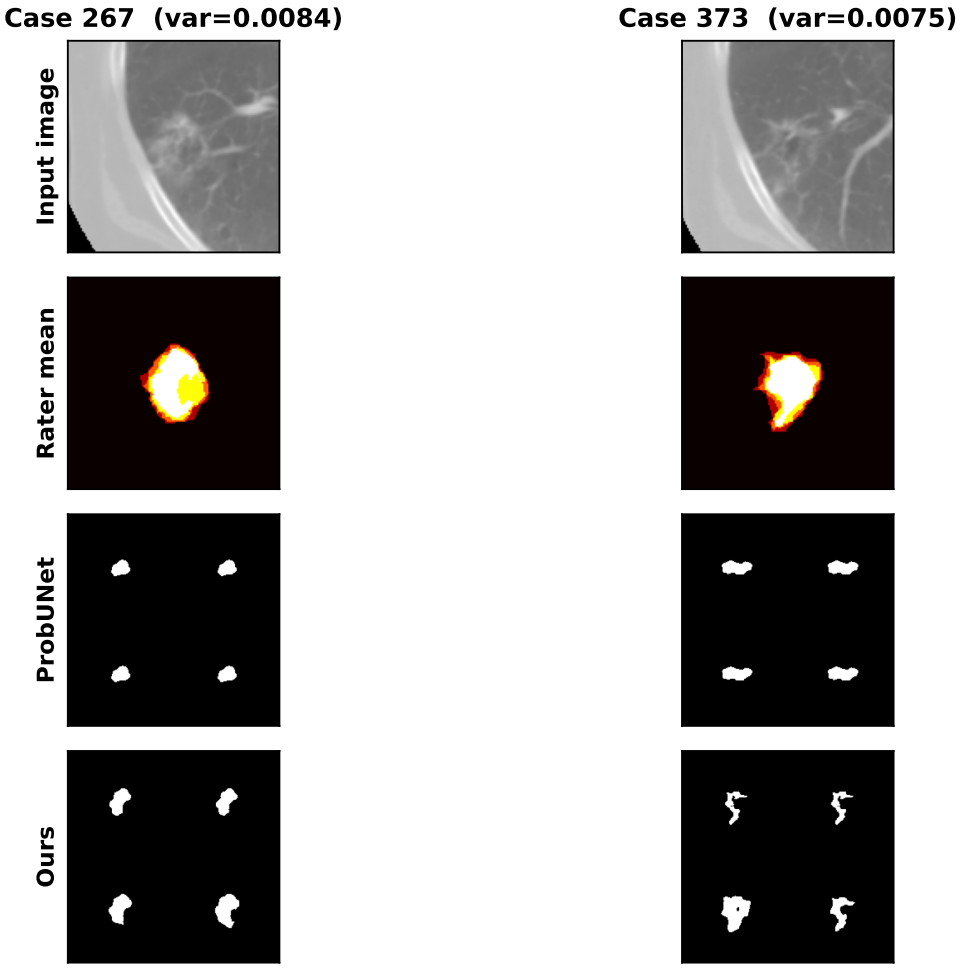

Figure 4.1 Qualitative comparison, high-disagreement LIDC-IDRI fold 3 cases.

Figure 4.2 shows all four posterior clouds and the prior mean overlapping in the t-SNE projection of 412 fold 3 test cases — no distinct per-rater clusters form. Each encoder pushes the prior in a rater-specific direction within a shared latent region rather than separating the modes. Stage 2 personalisation depends on per-rater modes being geometrically separated enough to steer from; that separation is absent here, which explains the Row 7 GED degradation.

The FComb decoder recovers rater-specific boundary decisions from directional differences between encoder signals, without the modes being globally separated

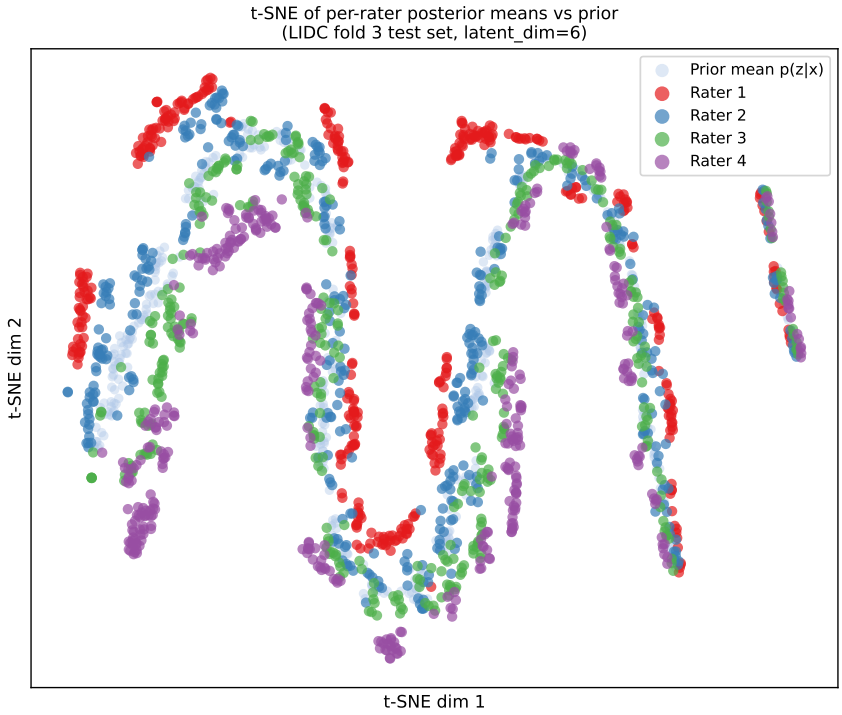

Figure 4.2 t-SNE of posterior and prior means, fold 3.

in the 6-dimensional space. Collapsing 6 dimensions to 2 for t-SNE discards these directional differences while preserving global structure, so the overlap in Figure 4.2 is consistent with — not a contradiction of — the GED gain. The improvement comes from what the training gradients encode, not from where the posterior means land.

4.3 Annotation Sparsity Robustness

At full annotation, per-rater posteriors improve GED and $\text{Dice}_{\text{match}}$ by 4.2% and 2.28% respectively — real but modest. Both models were then retrained at np $\in \{1, 2, 3\}$ annotators per training step, always evaluated on the full four-rater annotation set, to test how each degrades when coverage is incomplete.

4.3.1 GED Under Sparse Annotation

As coverage drops from four raters to one, the shared baseline loses more than half its GED performance while the per-rater model degrades far less severely; Table 4.6 quantifies this divergence (reference row: 100-epoch models; sparse rows: 300-epoch models). Two features of this table require explicit note. First, the baseline here is D-Persona Stage 1 only (GED 0.1436), not the Stage 1+2 pipeline of Table 4.2 (GED 0.1507): Stage 2 is omitted because it cannot be applied when annotators are routinely

absent during training. Second, the per-rater full-annotation GED (0.1429) differs slightly from the main-table value (0.1444) because these models were trained under the sparse experimental protocol and the main experiments were independent training runs; both are 100-epoch per-rater Stage 1 models, and the difference is within the cross-run variance expected from random initialisation.

Table 4.6 Per-rater vs. baseline GED across annotation coverage levels.

Annotators present	Baseline GED	Per-rater GED	Improvement
Full (np=4, reference)	$0.1436_{\pm.0076}$	$0.1429_{\pm.0143}$	+0.5% (n.s.)
np=3 (3 annotators)	$0.1810_{\pm.0175}$	$0.1601_{\pm.0090}$	**+11.5%**
np=2 (2 annotators)	$0.2039_{\pm.0239}$	$0.1677_{\pm.0144}$	**+17.8%**
np=1 (1 annotator)	$0.2220_{\pm.0093}$	$0.1745_{\pm.0119}$	**+21.4%**

The gap is monotonic: +11.5%, +17.8%, +21.4% as annotators drop from three to two to one, with no reversal at any level. More telling is the fold-by-fold record: across all 12 individual comparisons (three sparsity levels $\times$ four folds), not one favours the shared baseline. Under the null hypothesis of no systematic difference, the probability of that outcome is $(1/2)^{12} \approx 0.024\%$. The 12 comparisons are not fully independent (the three sparsity levels within each fold share the same test cases), but even under conservative adjustment the one-sided probability remains well below 0.01. At full annotation, the +0.5% gap is noise: the 4-fold SD for the baseline alone is ± 0.0076, which swamps a difference that small. The per-rater advantage is concentrated in the annotation sparsity regime, where the shared posterior's gradient signal degrades.

The baseline GED degrades sharply under sparsity: from 0.1436 at full annotation to 0.2220 at one annotator (a 54.6% increase). Per-rater GED rises from 0.1429 to 0.1745 (a 22.1% increase). The per-rater model degrades under sparsity too, as it should when fewer annotators are providing training signal, but it degrades far less severely. The shared baseline loses 54.6% of its full-annotation GED performance at maximum sparsity; per-rater posteriors lose 22.1%. The degradation rate differs by 2.5$\times$ and widens at every step down in coverage, as Figure 4.3 shows.

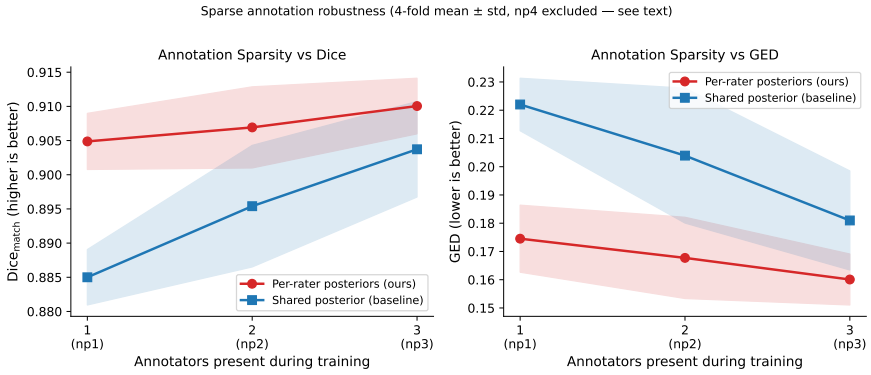

Figure 4.3 GED vs. annotators present per training image.

4.3.2 $\text{Dice}_{\text{match}}$ Under Sparse Annotation

GED penalises diversity loss while $\text{Dice}_{\text{match}}$ measures geometric closeness to individual annotator boundaries. When both metrics move in the same direction across every sparsity level in Table 4.7, a metric-specific artefact is ruled out.

Table 4.7 Per-rater vs. baseline $\text{Dice}_{\text{match}}$ across annotation coverage levels.

Annotators present	**Baseline $\text{Dice}_{\text{match}}$**	**Per-rater $\text{Dice}_{\text{match}}$**	**Δ Dice**
Full (np=4, reference)	$0.9126_{\pm.0033}$	$0.9130_{\pm.0060}$	$+0.0004$
np=3	$0.9037_{\pm.0070}$	$0.9101_{\pm.0041}$	$+0.0064$
np=2	$0.8954_{\pm.0089}$	$0.9069_{\pm.0059}$	$+0.0115$
np=1	$0.8850_{\pm.0041}$	$0.9049_{\pm.0041}$	$+0.0199$

$\text{Dice}_{\text{match}}$ shows the same monotonic pattern: $+0.0004$ at full annotation, rising to $+0.0199$ at one annotator. Both metrics agree on the direction at every sparsity level.

4.3.3 Fold 1 as the Sharpest Individual Evidence

Fold 1 is the only fold where the shared baseline outperforms per-rater posteriors at full annotation: baseline GED 0.1552 versus per-rater 0.1658, the baseline winning by 6.8%. This makes it the sharpest individual test of whether the advantage is structural.

At $\text{np} = 1$ in the same fold, on the same test cases, the picture reverses completely: baseline GED 0.2323 versus per-rater 0.1806. Per-rater wins by 22.3% $((0.2323 - 0.1806)/0.2323)$.

The model that is worse in Fold 1 at full annotation wins by 22.3% at maximum sparsity, on identical test cases. A general quality difference cannot explain that reversal. The reversal is tied specifically to the sparsity condition and is consistent with the mechanism described in §3.6: under $\text{np} = 1$, the shared baseline receives three zero-channel gradients for every one annotator gradient, collapsing the shared $\mathbf{z}$ toward empty-mask predictions. Per-rater posteriors avoid this entirely. At $\text{np} = 1$, the shared baseline's four prior samples converge to near-identical empty masks while the per-rater model's samples retain the spread visible across the four rater annotations, as Figure 4.4 shows side by side.

4.4 Gradient Alignment Analysis

The sparsity results establish that the shared baseline degrades far more than per-rater posteriors as annotation coverage decreases. The gradient alignment index (mean

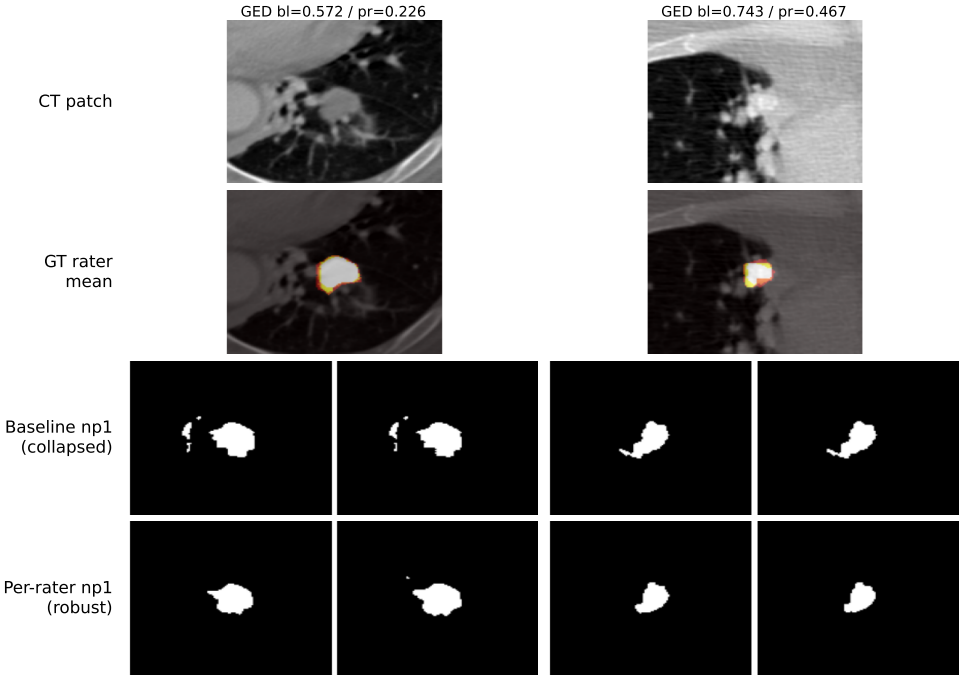

Figure 4.4 Qualitative comparison, single-annotator training ($n_p = 1$).

pairwise cosine similarity of per-rater reconstruction gradients at the shared latent code) turns the gradient conflict framework of Yu et al. [19] into a measurable scalar in the annotation sparsity setting. It examines whether the degeneration in the shared latent space is measurable, and whether it tracks the GED gap monotonically.

4.4.1 Results at Full Annotation

At full annotation, the mean pairwise cosine similarity of per-rater reconstruction gradients at the shared posterior mean $\bar{\mathbf{z}}$ is 0.167 across folds (fold-level standard deviation ± 0.026). The baseline alignment is positive and significantly different from zero in all four folds.

The within-fold distribution is highly heterogeneous: within a single fold's test set, the case-level standard deviation of alignment scores is approximately 0.439 (nearly three times the mean). Alignment ranges from strongly negative (opposing gradients, raters completely disagree on boundary location) to strongly positive (raters mostly agree). Approximately 62% of cases show positive alignment, 38% negative, and 17% exceed 0.5. Gradient conflict at full annotation is real but case-dependent, concentrated on specific nodules with ill-defined margins rather than universally presented.

The modest 4.2% GED improvement from per-rater posteriors at full annotation is consistent with gradient conflict being a significant problem only for a fraction of training cases.

4.4.2 Gradient Collapse Under Sparse Annotation

The alignment index rises in step with the GED gap, reaching near-unity at single-annotator coverage. Table 4.8 captures this collapse numerically.

Table 4.8 Gradient alignment index under sparse annotation.

Annotators	Baseline align.	SD (across folds)	Within-fold SD	Per-rater
Full (np=4)	0.167	0.026	≈0.439	0.000
np=3	0.291	0.074	—	0.000
np=2	0.463	0.039	—	0.000
np=1	0.976	0.012	≈0.023	0.000

The alignment index rises monotonically from 0.167 (full annotation) through 0.291 (three annotators) and 0.463 (two annotators) to 0.976 (one annotator).

The collapse at $\text{np} = 1$ is qualitatively different from the full-annotation condition, not a higher mean but a categorical shift in regime. At full annotation, the case-level within-fold standard deviation is approximately 0.439: gradient conflict varies substantially from nodule to nodule. At $\text{np} = 1$, the within-fold standard deviation drops to approximately 0.023 (an approximately 19-fold reduction; $0.439/0.023 = 19.1$), meaning almost every test nodule now shows near-complete gradient alignment regardless of how ambiguous its margin actually is. The shared posterior has entered a degenerate mode that is no longer sensitive to image content.

The fold-level consistency at $\text{np} = 1$ is also tight: standard deviation across folds is only 0.012, so the collapse is a property of the training regime rather than of one particular data split.

Per-rater alignment stays at 0.000 across all sparsity level. The zero value is expected: the architecture has no shared $\mathbf{z}$, so cross-rater gradient alignment at a single latent point is undefined.

4.4.3 Mechanism and Interpretation

The alignment data makes the zero-mask mechanism concrete. At $\text{np} = 1$, a 5-channel encoder receives three zero-mask channels and one real annotation channel simultaneously. The three empty-mask gradient directions are approximately collinear (all pushing $\mathbf{z}$ toward the same degenerate empty-segmentation prediction), while the one real gradient pushes in a structurally different direction. With three collinear gradients against one, mean pairwise cosine similarity is driven toward 1.0; the single real annotation gradient cannot break the degeneracy. Each step down in np increases the fraction of the training signal devoted to empty-mask pull, and the alignment index rises accordingly.

The alignment index is a diagnostic that measures the consequence of this collapse. The alignment index and the GED gap rise together, but one does not cause the other. Both trace back to zero-mask channels dominating the shared encoder's gradient as annotators are removed. Both the GED gap and the alignment index track the severity of that contamination, which is why they rise together as np decreases.

Figure 4.5 (GED gap and alignment on dual y-axes) shows this co-movement across the four annotation levels. The curves are not identical in shape (the alignment rises faster), but both are monotonically increasing in the same direction, reaching their peak at np = 1.

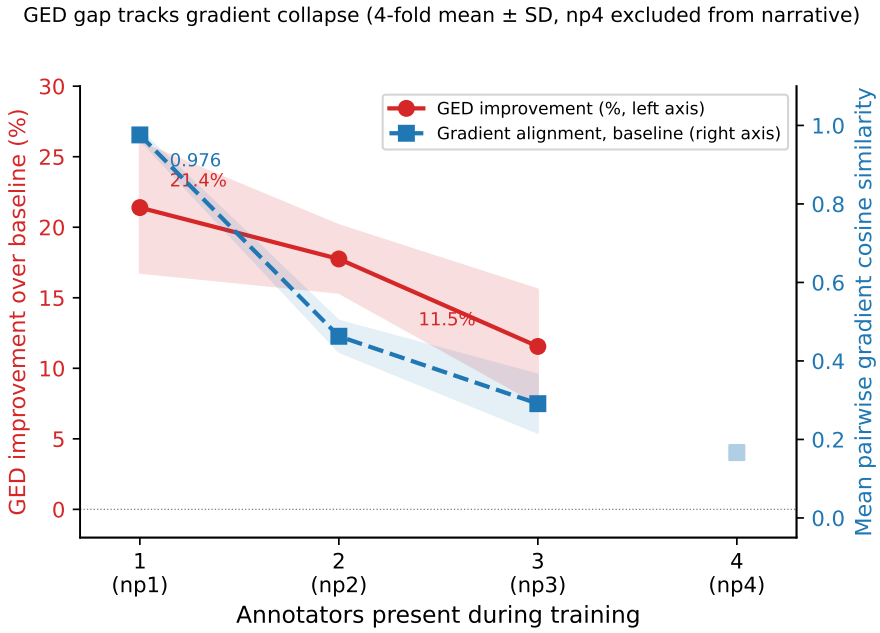

Figure 4.5 GED improvement and gradient alignment vs. annotators present.

4.5 Attribute Characterisation of Annotation Disagreement

The analysis in this section asks which nodule properties are associated with high inter-rater annotation disagreement in LIDC-IDRI [2]. It uses only the ground-truth masks and the nine per-rater clinical attribute ratings, and is independent of both segmentation models.

4.5.1 Results

Of the nine LIDC-IDRI clinical attributes, six reach statistical significance; Table 4.9 ranks them by effect size, separating the positive drivers of boundary ambiguity from the negative ones.

Table 4.9 Pearson r: attribute disagreement vs. inter-rater mask variance.

Attribute	**Pearson** r	p**-value**
Margin	$+0.318$	< 0.001
Lobulation	$+0.243$	< 0.001
Texture	$+0.210$	< 0.001
Spiculation	$+0.185$	< 0.001
Malignancy	-0.202	< 0.001
Subtlety	-0.155	< 0.001

Margin is the strongest predictor: $r = 0.318$, confirmed independently across all four folds (per-fold range: 0.238–0.400, all $p < 0.001$). Nodules with ill-defined, irregular margins cause the most inter-rater boundary disagreement. This is clinically expected: a sharp, well-defined margin leaves little room for interpretation; a gradual tissue transition at the nodule edge creates genuine ambiguity about where the nodule ends.

Lobulation ($r = 0.243$) and texture ($r = 0.210$) follow. Part-solid nodules (those with both solid and ground-glass opacity components) have unclear density transitions, and lobulated shapes mean the boundary is non-convex and harder to delineate consistently.

The two negative correlations are the more interesting finding. Malignancy ($r = -0.202$) and subtlety ($r = -0.155$) are negatively associated with mask variance. Higher inter-rater disagreement on a nodule's perceived malignancy does not predict higher mask variance; if anything, it slightly predicts lower mask variance. A nodule can be perceived as highly malignant while having a geometrically clear boundary, and a nodule can be ambiguous in appearance (hard to spot) while still being easy to delineate once found.

This separation matters clinically. A clinician looking for where uncertainty-aware segmentation is most needed should prioritise ill-defined, lobulated, part-solid nodules rather than ones rated most suspicious for malignancy. The raw scatter in Figure 4.6 confirms the $r = 0.318$ trend is genuine and not driven by outliers: margin disagreement predicts mask variance across the full range of cases with no discontinuity.

4.5.2 Four-Fold Confirmation

The margin correlation is particularly important to verify across folds because the attribute analysis is applied to 1,603 cases without further splitting: computing it once on the full dataset could in principle reflect a data-split artefact. Computing it independently within each fold's test set (fold sizes 450, 375, 412, 372) gives margin r values of 0.238, 0.305, 0.400, and 0.359 respectively. All four are positive and

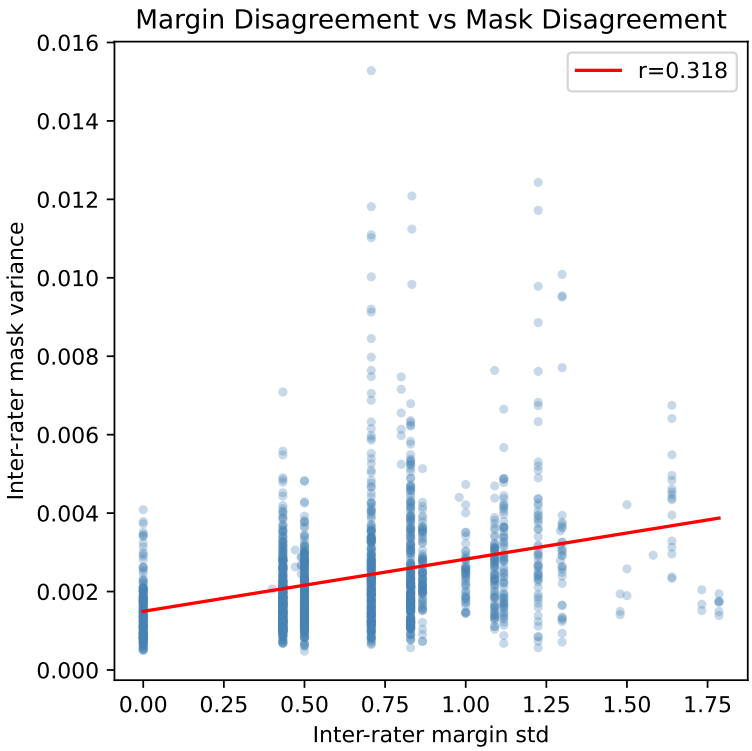

Figure 4.6 Margin disagreement vs. mask variance, LIDC-IDRI ($n = 1{,}603$).

significant. The result is stable across data splits, not an artefact of one particular test partition.

4.6 Discussion

4.6.1 Why Per-Rater Posteriors Work at Full Annotation

The t-SNE (Fig. 4.2) shows no distinct per-rater latent clusters; the four posterior means and the prior means overlap in the 6-dimensional latent space. Each encoder provides a directed gradient signal during training that the FComb decoder can distinguish from the other three, even when those signals point into the same latent region. The decoder learns rater-specific boundary responses from directional differences that t-SNE cannot reveal, since collapsing 6 dimensions to 2 discards the directional structure driving the effect.

The 0.5% full-annotation GED improvement reflects that gradient conflict at full annotation affects only a subset of training cases. Approximately 62% show positive alignment, 38% negative, and the within-fold standard deviation is 0.439. For cases where raters mostly agree, the shared and per-rater posteriors receive similar

training signals. The aggregate 4.2% GED gain comes disproportionately from the roughly 40% of cases where annotators diverge on the nodule boundary — the fraction where per-rater gradient isolation makes the biggest difference to what the shared decoder receives.

4.6.2 Why the Advantage Scales with Sparsity

At full annotation, gradient conflict is tied to specific nodules where raters disagree — it is bidirectional and case-dependent. Sparsity introduces a structurally different source of interference: zero-channel gradients from absent annotators all point the same way regardless of which nodule the encoder is processing. This directional uniformity is what makes sparse-regime degradation so much more severe than full-annotation conflict, and why the alignment index at $\text{np} = 1$ hits 0.976 rather than staying near its full-annotation value of 0.167.

As np decreases from 4 to 1, the proportion of the training signal coming from zero-channel gradients increases from 0 to 75%. The shared $\mathbf{z}$ has progressively less space to encode genuine annotation structure against this growing pressure. The gradient alignment index captures this, rising from 0.167 at full annotation (case-dependent conflict) to 0.976 at $\text{np} = 1$ (near-universal collapse). Per-rater posteriors are not affected by this mechanism at any sparsity level, since absent encoders do not run and contribute nothing to gradient update.

The Fold 1 reversal (§4.3.3) makes the sparsity-specificity of the advantage concrete at the individual fold level. A model that trails the shared baseline in a given fold at full annotation, then outperforms it by 22.3% in that same fold under single-annotator training, is responding to the sparsity condition specifically, not performing better in general.

4.6.3 Scope and Boundaries of the Results

On NPC-170, the GED gap falls within seed variance (§4.2.4); the method does not degrade on MRI nasopharyngeal carcinoma but does not improve either. The 0.5% full-annotation gap on LIDC-IDRI is similarly uninterpretable; the 4-fold SD for the baseline alone is ± 0.0076, which swamps a difference that small. On latent structure, the t-SNE shows no per-annotator clustering, which means the gain comes from gradient dynamics during training rather than from the model forming geometrically distinct modes for each annotator. The gradient alignment index is a diagnostic that characterises what happened to the shared encoder's training signal; it measures the consequence of zero-mask channel contamination rather than establishing a causal pathway from gradient collapse to GED degradation.

4.6.4 Clinical Implications

The benefit of per-rater posteriors concentrates in two specific deployment contexts.

First, datasets where not all radiologists annotate every image (which describes most large clinical collections). At one-annotator training coverage, the $+21.4\%$ GED improvement means the per-rater model preserves substantially more annotation diversity than the shared baseline. In a hospital setting where only one radiologist routinely annotates a given scan type, a model trained with per-rater posteriors remains calibrated to that radiologist's style without gradient contamination from absent annotators.

Second, the attribute analysis identifies where probabilistic methods matter most: ill-defined, lobulated, part-solid nodules. These nodules generate high inter-rater annotation disagreement by structural necessity, and a model calibrated to individual radiologist styles is most valuable for this group of cases. A deterministic segmentation model applied to well-defined, spiculated nodule does not lose much by ignoring annotation uncertainty. Applied to a nodule with an ill-defined margin (the attribute most correlated with inter-rater disagreement, $r = 0.318$), the absence of calibrated uncertainty is a real clinical problem.

4.7 Limitations

Dataset scope: All experiments use either LIDC-IDRI (CT, four radiologists, lung nodules) or NPC-170 (MRI, four annotators, nasopharyngeal carcinoma). Both datasets have $N = 4$ annotators by design. Generalisation to datasets with two annotators, ten annotators, or variable annotation counts per case has not been tested.

Training time: Per-rater Stage 1 training takes approximately 47 hours across all four LIDC-IDRI folds (4.5 times the shared baseline Stage 1 time of approximately 10.5 hours). In a setting with large-scale datasets or limited GPU access, this overhead is a practical barrier.

Epoch count confound in sparse experiments: Sparse models are trained for 300 epochs; full-annotation models for 100. The full-annotation reference row in Table 4.6 uses 100-epoch models. When sparse experiments are run at $\text{np} = 4$ (all annotators present) with 300-epoch training, the shared baseline degrades and wins only one of four folds (mean gap -3.1%). This is likely a training-artefact confound (longer training amplifies the shared posterior's degeneration under the full 5-channel input), but it cannot be ruled out as a genuine interaction without a controlled equal-epoch comparison at all sparsity levels.

Gradient alignment as diagnostic only: The alignment index is a diagnostic measurement computed on trained models. It characterises what happened to the shared $\mathbf{z}$

during training; causal proof that gradient conflict drove the GED gap would require a controlled intervention the experiments here cannot provide. An alternative explanation (that absent-rater information loss alone, without gradient mechanics, is responsible) cannot be definitively excluded.

Analysis C null result: We attempted to demonstrate that the per-rater GED improvement is concentrated on clinically ambiguous nodules by correlating per-rater improvement with per-attribute inter-rater disagreement. Pearson r between per-rater GED improvement and subtlety disagreement: $r = -0.042$ ($p = 0.093$); malignancy disagreement: $r = -0.023$ ($p = 0.356$); spiculation disagreement: $r = 0.035$ ($p = 0.157$). None is significant at $p < 0.05$. The claim that "per-rater posteriors help most in clinically ambiguous cases" is not supported statistically and is not made anywhere in this thesis.

Clinical validation: Neither LIDC-IDRI nor NPC-170 is a prospective clinical study. The claim that per-rater posteriors are useful in real deployment cannot be made without validation on prospectively collected multi-centre data with known radiologist agreement patterns.

CHAPTER 5

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

This thesis tested whether replacing the single shared posterior encoder with N independent per-rater encoders reduces gradient conflict enough to improve multi-rater segmentation in practice. At full annotation the GED advantage is 4.2%, real but modest, and at one annotator per training image it reaches 21.4%. The gap grows largest under annotation conditions most common in clinical practice, where complete multi-rater coverage is the exception rather than the rule.

5.1 Summary of Contributions

1. **Gradient isolation through per-rater posterior encoders.** D-Persona's Stage 1 shared encoder trains on all four annotators' masks simultaneously and receives conflicting gradient signals that partially cancel whenever raters disagree. The latent code encodes a compromise, and any downstream personalisation — Stage 2 projection heads, regularisation losses — operates on that already-averaged representation, working with gradient-cancelled residuals rather than the original per-rater signals.

   Four independent per-rater encoders — each receiving only the image and one annotator's mask — give the shared decoder four clean, rater-specific gradient signals during training. $\partial \mathcal{L}_i / \partial \mathbf{z}_j = 0$ for $i \neq j$ follows from the objective structure alone, with no regularisation term required. On LIDC-IDRI (1,609 nodule patches, 4-fold CV), the Stage 1-only per-rater model reaches GED 0.1444 ± 0.0141 and $\text{Dice}_{\text{match}}$ 0.9112 ± 0.0061 — a 4.2% GED reduction and 2.28% $\text{Dice}_{\text{match}}$ gain over the full D-Persona pipeline. All four per-rater Dice scores improve individually. $\text{Dice}_{\text{soft}}$ stays at 0.9015 for both models — the gain is in per-rater calibration, with average prediction quality unchanged.

2. **An ablation that rules out capacity, regularisation, and diversity loss as explanations.** Six modifications were tested against the D-Persona baseline. MiT-B2 adds parameters; GED is unchanged. An orthogonality loss worsens both metrics, because regularising the output of an encoder that trained on averaged gradients cannot undo the averaging that already occurred. The discretised prior bank ($k = 100$) produces the worst result in the table (GED 0.2212), removing

the stochastic sampling that GED rewards. Dual diversity loss ties the baseline.

Row 7 sharpens the interpretation: adding Stage 2 style vectors on top of per-rater Stage 1 raises GED by 27% (from 0.1444 to 0.1836). Stage 2 was designed to personalise a shared latent code. Applied after per-rater Stage 1 has already differentiated the gradient pathways, it imposes shared-posterior dynamics on a model that does not need them and collapses the diversity Stage 1 created.

3. **Sparsity experiments across three coverage levels, with a gradient collapse diagnostic.** At one annotator per training image, the shared baseline's 5-channel input carries three zero-mask channels. Three reconstruction gradients push the shared latent code toward empty-mask predictions; the single real annotation gradient competes against all three. The gradient alignment index — mean pairwise cosine similarity of per-rater gradients at the shared $\mathbf{z}$ — rises from 0.167 at full annotation to 0.976 at np $= 1$. The within-fold standard deviation drops 19-fold (from 0.439 to 0.023): the collapse is near-universal at maximum sparsity rather than case-dependent.

   Absent per-rater encoders do not execute and contribute nothing to the gradient update. GED advantage grows from $+11.5\%$ to $+21.4\%$ as coverage falls; all 12 per-fold comparisons at the three sparsity levels favour the per-rater model ($p < 0.001$, sign test). At full annotation the gap is 0.5% and within noise.

4. **Pearson correlation analysis linking nine nodule attributes to inter-rater mask variance.** Run on 1,603 LIDC-IDRI cases with complete attribute records, independent of both segmentation models. Margin clarity is the strongest predictor ($r = 0.318$, $p < 0.001$), confirmed across all four CV folds (r range: 0.238–0.400). Lobulation ($r = 0.243$) and texture ($r = 0.210$, capturing part-solid opacity) follow.

   The two negative correlations carry clinical weight: malignancy ($r = -0.202$) and subtlety ($r = -0.155$) are each negatively associated with mask variance. Boundary ambiguity and diagnostic severity are separable dimensions. A nodule perceived as highly malignant can have a geometrically clear edge, and a hard-to-detect nodule may be straightforward to delineate once found. Uncertainty-aware segmentation is most valuable for ill-defined, lobulated, part-solid nodules — a different population from those rated most suspicious for malignancy.

5.2 Overarching Finding

Under sparsity, absent annotators' zero-mask channels dominate the gradient at the shared latent code, and the model degrades accordingly. Stage 2 personalisation, a stronger backbone, and diversity regularisation all operate downstream of Stage 1 — after gradient averaging has already shaped the latent code. Stage 2 can only redistribute information already present in $\mathbf{z}$ — what gradient averaging removed before Stage 1's backward pass completed is gone from the latent code by the time Stage 2 runs. The orthogonality loss regularises the representation space after the gradient cancellation

has occurred. A larger backbone encodes a more expressive average of conflicting signals, with the conflict itself unaffected.

Mean pairwise cosine similarity at the shared code is 0.167 at full annotation — gradient conflict is real but concentrated on ambiguous nodules. At one-annotator coverage it reaches 0.976. The GED gap tracks the same trajectory: $+4.2\%$ at full annotation, $+11.5\%$, $+17.8\%$, and $+21.4\%$ as annotators drop to three, two, and one. Per-rater posteriors cut off the contamination at the encoder level, before it accumulates in $\mathbf{z}$, which is why the improvement scales with the contamination rather than being fixed.

5.3 Future Scope

Three open questions follow directly from the results.

1. **Pairwise repulsion between per-rater posterior means.** The t-SNE of the per-rater latent space (§4.2) shows that all four posterior means and the prior mean land in the same region of the 6-dimensional latent space, despite fully independent gradient pathways. The FComb decoder exploits directional differences between the per-rater signals — each encoder pushes in a rater-specific direction within a shared latent region. Stage 2 personalisation depends on per-rater modes being geometrically separated enough to project from, and those separations do not form.

   A repulsion term penalising proximity between μ_i and μ_j $(i \neq j)$ during Stage 1 training could push the modes apart while keeping per-encoder gradient isolation intact. If per-rater Stage 1 training produced distinct latent clusters, Stage 2 would have genuine rater-specific structure to steer from, and the Row 7 ablation result might reverse. Whether a repulsion term can achieve this without reintroducing gradient mixing remains untested.

2. **Attribute-conditioned training curricula.** The attribute analysis establishes that nodules with ill-defined margins, lobulated shapes, and part-solid texture account for the highest inter-rater annotation disagreement. Currently, all training images contribute equally to the per-rater ELBO, regardless of how much rater disagreement they carry. A curriculum that upweights high-disagreement cases (weighting reconstruction losses by inter-rater mask variance, or sampling high-disagreement cases more frequently in early epochs) could concentrate the model's learning capacity on the cases where boundary uncertainty is highest and calibrated probabilistic outputs matter most. The attribute correlation analysis provide a principled, clinically grounded basis for defining such a curriculum without requiring additional annotation effort.

3. **Scaling to variable annotator counts.** LIDC-IDRI has exactly four annotators per case by design. Real clinical datasets have variable annotator counts: some

images may have one expert annotation, others five or eight, depending on when and where they were collected. The per-rater formulation scales linearly in the number of encoders, but the training dynamics at $N = 2$ (where shared posterior gradient conflict is much less severe) and at $N = 8$ or more (where per-rater training cost grows substantially) have not been studied. Understanding how the gradient conflict advantage scales with N, and at what N the shared posterior becomes an adequate approximation, would establish the boundary conditions for when the per-rater design justifies its $4.5\times$ training cost overhead.

5.4 Social Impact

Per-rater posterior training was motivated by clinical settings where annotation resources are scarce and radiologist disagreement is structurally unavoidable. The results carry implications for deployment and for how annotation protocols are designed, alongside clear limits on what the experiments do and do not establish.

5.4.1 Direct Impact

The most concrete implication of the sparsity results concerns deployment in healthcare settings where specialist annotation coverage is limited. In tertiary centres with multiple radiologists on staff, multi-rater annotation of a CT study is operationally feasible. In rural district hospitals, community radiology practices, and healthcare systems in low-income settings, a single radiologist reading per scan is the norm rather than the exception. At single-annotator training coverage, the per-rater model achieves a 21.4% GED improvement over the shared baseline. A probabilistic model trained with per-rater posteriors on single-annotator data retains substantially more annotation diversity than one trained with a shared encoder under the same data conditions. At maximum sparsity, GED values remain 0.1745 (per-rater) and 0.2220 (shared baseline) — both represent real prediction variance, and neither substitutes for multi-annotator consensus in high-stakes decisions. For a model whose role is to flag boundary uncertainty to a reviewing clinician, the per-rater formulation provides better-calibrated diversity under single-annotator training.

Uncertainty-aware segmentation calibrated to individual radiologist styles enables presenting a range of plausible interpretations to a clinician reviewing a borderline case. A shared-posterior model, shaped by gradient averaging, collapses toward a distribution over gradient-averaged compromises; per-rater training preserves each annotator's gradient signal, so the resulting predictions can span the actual range of clinical judgment.

5.4.2 Indirect Impact

Nodule margin clarity ($r = 0.318$) being the dominant driver of inter-rater mask variance — more than malignancy or subtlety — points to a specific place in the annotation workflow where explicit guidance would reduce disagreement most efficiently. Annotation protocols that require radiologists to record a margin assessment (well-defined, lobulated, ill-defined) before drawing a contour make the margin judgment explicit and documentable before any contour is drawn, which targets the systematic component of inter-rater disagreement. Training programmes that address margin ambiguity at CT tissue transitions would focus on attribute most responsible for annotation variance.

Since malignancy ($r = -0.202$) and boundary ambiguity are separable dimensions, a clinical AI system that treats them as one will misassign uncertainty in both directions: high boundary uncertainty reported for a malignant-appearing nodule with a geometrically clear edge, and low boundary uncertainty for an ambiguous-margin nodule that looks benign. The attribute data support keeping segmentation uncertainty and diagnostic uncertainty as separate model outputs.

5.4.3 Limitations for Societal Deployment

All results come from LIDC-IDRI (CT, lung nodules, four radiologists) and NPC-170 (MRI, nasopharyngeal carcinoma, four annotators). Neither dataset is a prospective clinical study. The per-rater method has not been tested on multi-centre data with known radiologist agreement patterns, on datasets with more than four annotators, or on imaging modalities beyond chest CT and head-and-neck MRI. Clinical deployment requires prospective validation these experiments do not provide.

Per-rater Stage 1 training takes approximately 47 hours across four LIDC-IDRI folds (4.5$\times$ the shared baseline Stage 1 time of approximately 10.5 hours on Apple MPS hardware). For institutions without GPU access that overhead is a real barrier. Parameter sharing between encoders (shared early convolutional layers, diverging near the output) is one way to cut the cost, at the price of partially reintroducing gradient mixing. Distillation from a fully trained per-rater model into a smaller single-encoder student preserves more of the gradient isolation than early-layer sharing, though neither option has been evaluated on this architecture.

APPENDIX I

FOLD-LEVEL RESULTS: FULL ANNOTATION

Table I.1 reports per-fold GED and $\text{Dice}_{\text{match}}$ for both models evaluated at full annotation (all four annotators present, 100-epoch training). These values are read directly from `results/full_annotation_results.csv`. Bold entries mark the better-performing model in each fold.

The comparison here is between per-rater Stage 1 and the shared-posterior Stage 1 baseline — both trained without Stage 2 style vectors and at 100 epochs. The primary thesis comparison (Table 4.1) is between per-rater Stage 1 and the full D-Persona Stage 1+2 pipeline; per-fold data for the full D-Persona pipeline and for ProbUNet were not retained from the original training runs and are reported only as 4-fold means (±SD) in Chapter 4.

Fold 1 is the most instructive row: the shared baseline achieves lower GED (0.1552 vs. 0.1658), i.e. the per-rater model is worse at full annotation in this particular fold. This same fold shows the strongest per-rater advantage under annotation sparsity (Table II.1), which is the mechanistic point developed in §4.3.3.

Table I.1 Per-fold GED and $\text{Dice}_{\text{match}}$, full annotation.

	GED ↓				
Model	**Fold 0**	**Fold 1**	**Fold 2**	**Fold 3**	**Mean**
Per-Rater (ours)	**0.1371**	0.1658	**0.1268**	**0.1418**	**0.1429**
Shared Baseline	0.1426	**0.1552**	0.1338	0.1428	0.1436
	$\text{Dice}_{\text{match}}$ ↑				
Model	**Fold 0**	**Fold 1**	**Fold 2**	**Fold 3**	**Mean**
Per-Rater (ours)	**0.9143**	0.9031	**0.9193**	**0.9152**	**0.9130**
Shared Baseline	0.9132	**0.9074**	0.9165	0.9135	0.9127

APPENDIX II

FOLD-LEVEL RESULTS: SPARSE ANNOTATION

Tables II.1 and II.2 report per-fold GED and $\text{Dice}_{\text{match}}$ for all three sparsity levels (one, two, or three annotators present per training image). Both models were trained for 300 epochs at each sparsity level. All 12 per-fold comparisons in Table II.1 favour the per-rater model (lower GED); the expected probability of this pattern under the null hypothesis of equal models is $\left(\frac{1}{2}\right)^{12} \approx 0.024\%$.

The np = 4 rows (all four annotators present, 300-epoch sparse training) are included in Table II.3 for completeness but are *excluded* from the main results tables in Chapter 4. At 300 epochs with all four annotators, the shared baseline wins three of four folds (mean GED 0.1285 vs. 0.1328 for per-rater). This reversal is a training-duration artefact: 300-epoch training amplifies the degeneration of the shared-posterior objective under the full 5-channel input, producing a model that is worse than the 100-epoch reference. The full-annotation reference throughout this thesis uses 100-epoch models only; the np = 4 sparse row is neither the controlled full-annotation reference nor a clean sparsity result.

Table II.1 Per-fold GED under sparse annotation, 4-fold CV.

n_{present}	Model	Fold 0	Fold 1	Fold 2	Fold 3	Mean
np = 3	Per-Rater	**0.1632**	**0.1708**	**0.1459**	**0.1604**	**0.1601**
	Shared Baseline	0.1775	0.2085	0.1597	0.1782	0.1810
np = 2	Per-Rater	**0.1646**	**0.1912**	**0.1521**	**0.1630**	**0.1677**
	Shared Baseline	0.1979	0.2437	0.1803	0.1938	0.2039
np = 1	Per-Rater	**0.1640**	**0.1806**	**0.1625**	**0.1910**	**0.1745**
	Shared Baseline	0.2257	0.2323	0.2070	0.2231	0.2220

Table II.2 Per-fold $\text{Dice}_{\text{match}}$ under sparse annotation, 4-fold CV.

n_{present}	Model	Fold 0	Fold 1	Fold 2	Fold 3	Mean
$\text{np} = 3$	Per-Rater	**0.9083**	**0.9045**	**0.9154**	**0.9120**	**0.9101**
	Shared Baseline	0.9036	0.8927	0.9117	0.9069	0.9037
$\text{np} = 2$	Per-Rater	**0.9068**	**0.8975**	**0.9135**	**0.9099**	**0.9069**
	Shared Baseline	0.8966	0.8807	0.9041	0.9002	0.8954
$\text{np} = 1$	Per-Rater	**0.9070**	**0.8998**	**0.9104**	**0.9023**	**0.9049**
	Shared Baseline	0.8822	0.8805	0.8910	0.8863	0.8850

Table II.3 Per-fold GED and $\text{Dice}_{\text{match}}$, $n_p = 4$ sparse training.

Metric	Model	Fold 0	Fold 1	Fold 2	Fold 3	Mean
GED $\downarrow$	Per-Rater	**0.1242**	0.1567	0.1186	0.1315	0.1328
	Shared Baseline	0.1249	**0.1426**	**0.1153**	**0.1311**	**0.1285**
$\text{Dice}_{\text{match}}$ $\uparrow$	Per-Rater	**0.9169**	0.9003	**0.9212**	**0.9164**	0.9137
	Shared Baseline	0.9163	**0.9109**	0.9208	0.9163	**0.9161**

APPENDIX III

FOLD-LEVEL GRADIENT ALIGNMENT RESULTS

Table III.1 reports per-fold mean pairwise cosine similarity of per-rater reconstruction gradients at the shared latent code $\mathbf{z}$, for the shared-posterior baseline at all annotation coverage levels. Each fold value is the mean over $n = 100$ test cases. The per-rater model has zero gradient alignment by construction at all levels: the concept of a shared $\mathbf{z}$ does not exist in the per-rater design, so the measurement is undefined and reported as 0.000.

Two aspects of the full-annotation row are worth noting. The 4-fold mean (0.167) masks substantial case-level variability: within a single fold, the gradient alignment across the 100 test cases has a standard deviation of approximately 0.439 — roughly 38% of cases show negative pairwise cosine similarity (raters' gradients actually pointing in opposite directions) while 62% show positive alignment. At np = 1, this within-fold case spread collapses to approximately 0.023, an approximately 19-fold reduction. The collapse from case-specific conflict at full annotation to near-universal collapse at maximum sparsity is the central diagnostic finding of the gradient alignment analysis.

Table III.1 Gradient alignment per fold, baseline only.

n_{present}	Fold 0	Fold 1	Fold 2	Fold 3	4-fold mean	SD (across folds)
Full (4)	0.1366	0.1827	0.1999	0.1469	0.167	0.026
np = 3	0.2739	0.4163	0.2291	0.2454	0.291	0.074
np = 2	0.4706	0.5242	0.4267	0.4299	0.463	0.039
np = 1	0.9800	0.9807	0.9560	0.9867	0.976	0.012
Per-rater (all levels)	0.000	0.000	0.000	0.000	0.000	—

The np = 1 row shows the lowest across-fold SD (0.012) of any sparsity level — the opposite of what intuition might suggest. At full annotation, case-level variability is high (within-fold SD $\approx$ 0.439) and the 4-fold SD is moderate (0.026). As sparsity increases, the zero-channel gradient dominance overwhelms case-specific boundary differences: by np = 1, the alignment is near-1.0 in all four folds regardless of which test cases are in that fold. The collapse is systematic rather than fold-dependent, which is why the across-fold SD at np = 1 (0.012) is smaller than at full annotation (0.026).

APPENDIX IV

ABLATION STUDY: PER-FOLD BREAKDOWN

Per-fold GED and $\text{Dice}_{\text{match}}$ for the intermediate ablation rows (Rows 2–5 and Row 7 in Table 4.3) were not recorded separately during training. The ablation study tracked 4-fold mean results, and only the endpoint models (shared baseline Stage 1+2 and per-rater Stage 1) have per-fold evaluation data in the CSV output files. Table IV.1 provides the per-fold breakdown for these two rows for reference.

Table IV.1 Per-fold GED and $\text{Dice}_{\text{match}}$, ablation endpoints.

Metric	Row	Fold 0	Fold 1	Fold 2	Fold 3	Mean
GED ↓	Per-Rater Stage 1 (Row 6)	**0.1371**	0.1658	**0.1268**	**0.1418**	**0.1429**
	Shared Baseline Stage 1 only	0.1426	**0.1552**	0.1338	0.1428	0.1436
$\text{Dice}_{\text{match}}$ ↑	Per-Rater Stage 1 (Row 6)	**0.9143**	0.9031	**0.9193**	**0.9152**	**0.9130**
	Shared Baseline Stage 1 only	0.9132	**0.9074**	0.9165	0.9135	0.9127

Note on Row 1: The primary thesis comparison (Table 4.1) compares our model against the full D-Persona Stage 1+2 pipeline (mean GED 0.1507, $\text{Dice}_{\text{match}}$ 0.8909). The per-fold data above is for D-Persona Stage 1 only (the shared posterior without Stage 2 projection heads), which is the relevant baseline for the sparsity experiments. The Stage 1+2 pipeline per-fold values were not retained after the original training.

APPENDIX V

TRAINING HYPERPARAMETERS

Table V.1 lists every hyperparameter held fixed across all experiments. Values are identical for both the full-annotation and sparse-annotation conditions except the epoch count, which increases from 100 to 300 under sparsity to compensate for reduced gradient signal per epoch.

Table V.1 Training hyperparameters, all experiments.

Hyperparameter	**Value**
Backbone	ResNet34
Optimiser	Adam
Learning rate	1×10^{-4}
LR schedule	Cosine annealing
Latent dimension (D)	6
Batch size	12
Epochs (full annotation)	100
Epochs (sparse annotation)	300
β (KL weight)	0.5
Number of annotators (N)	4
Input size (LIDC-IDRI)	$128 \times 128 \times 1$
Input size (NPC-170)	3-channel (T1, T1CE, T2)
Hardware	Apple MPS
Training time (full, 4 folds)	$\sim$47 hours

APPENDIX VI

ATTRIBUTE CORRELATION PER FOLD

Table VI.1 reports the Pearson r between inter-rater margin standard deviation and inter-rater mask variance, computed independently within each fold's test set. This verifies that the margin finding ($r = 0.318$ overall) is not an artefact of any particular train/test partition.

Per-fold correlation analysis was conducted for margin only, the strongest predictor. For the remaining five attributes (lobulation, texture, spiculation, malignancy, subtlety), the Pearson r was computed on the full 1,603-case dataset without fold-level breakdown; those values are reported in Table 4.8 in Chapter 4.

Table VI.1 Per-fold Pearson r: margin disagreement vs. inter-rater mask variance.

Attribute	**Fold 0**	**Fold 1**	**Fold 2**	**Fold 3**	**4-fold range**	**Overall** ($n = 1603$)
Margin (r)	0.238	0.305	0.400	0.359	0.238–0.400	0.318

All four per-fold margin correlations are positive, consistent, and significant at $p < 0.001$. The range (0.238–0.400) spans slightly less than two-fold, which is the expected variability for a correlation estimated on 372–450 cases. Fold 2 has the strongest per-fold association ($r = 0.400$), likely reflecting the composition of that fold's test cases (412 patches, with a higher proportion of lobulated, part-solid nodules). Fold 0 has the lowest ($r = 0.238$), but remains clearly significant. The across-fold consistency confirms that the margin-uncertainty relationship is a structural property of the LIDC-IDRI dataset rather than a data-split artefact. Table VI.2 extends the analysis to all six attributes on the full dataset.

Table VI.2 All-attribute Pearson r, LIDC-IDRI full dataset ($n = 1{,}603$).

Attribute	**Pearson** r	p**-value**
Margin	$+0.318$	< 0.001
Lobulation	$+0.243$	< 0.001
Texture	$+0.210$	< 0.001
Spiculation	$+0.185$	< 0.001
Malignancy	-0.202	< 0.001
Subtlety	-0.155	< 0.001