

Thesis

by Riya Thesis

Submission date: 29-May-2026 11:51AM (UTC+0530)

Submission ID: 2971826826

File name: Riya_Thesis.pdf (1.22M)

Word count: 7285

Character count: 41778

Chapter 1

INTRODUCTION

In Automatic ³image captioning — the task of generating natural language descriptions from visual contents of the image — is a ³fundamental challenge at the intersection of computer vision (CV) and natural language processing (NLP). A captioning systems must not only the recognize the objects, attributes, and the relationships present in a image, but also express them in fluent, grammatically correct natural language. This dual requirement makes image captioning a uniquely demanding multimodal tasks. Its applications is wide-ranging and include accessibility services for visually impaired individuals, medical imaging report generation, autonomous vehicle scenes understanding, social media content moderation, and image retrieval systems [1].

1.1 Background and Motivation

In the field of image captioning has progressed through three broad architectural generations. Early system based on ²³Convolutional Neural Networks(CNNs) coupled with Long Short-Term Memory (LSTM) decoders demonstrated the feasibility of caption generation but also were ¹⁶limited in their ability to capture global visual context [1]. The introductions of attention based mechanisms improved region-level focus during decoding stages [2], while transformer-based architectures subsequently enabled global dependency modeling across both the visual and textual domains. More recently, large-scale multimodals models such as BLIP [3] and its successors have further advanced performance through unified pretraining and instruction tuning.

Despite these advance, most studies evaluate complete, end-to-end captioning systems

without isolating the contribution of the individual architectural components. As a result, it is often unclear whether an observed performance gain arises from a stronger visual encoder, a more capable language decoder, or simply from larger-scale pretrainings. This lack of component-level interpretability limits the principled designing of efficient systems, particularly for resource-constrained deployment where every parameter and every unit of inference cost matters.

1.2 Problem Statement

The traditional models those are based on CNNs and LSTM based networks provide reasonable result but often fails to capture global context and generate detailed captions. The Transformer-based architectures such as the Vision Transformer (ViT) and the large language models have improved performances by modeling long-range dependency. However because these models are typically studied as monolithic system, the relative importance of the visual encoder vs. the language decoder remains poorly understood. In this This thesis addresses this gap by systematically comparing different encoder-decoder combinations under a unified experimental framework, in order to analyze the role of the visual encoder and the language decoder in image captioning.

1.3 Research Questions

In this work it is guided by three research questions mainly:

- **RQ1 (Encoders):** How exactly do CNN-based and transformer-based visual encoders differ in their impact on captioning performance?
- **RQ2 (Decoders):** How do the recurrent(LSTM) and pretrained language model (GPT-2) decoders compare in caption generation quality and linguistic fluency?
- **RQ3 (Interaction):** Which component, the encoder or the decoder contributes more to overall captioning quality?

1.4 Contributions

The contributions of this thesis are as follows:

1. A unified taxonomy of visual encoders(CNN-base and transformer-based) and language decoders (recurrent and pretrained language model-based) for image captioning.
2. A controlled, component-wise experimental comparison of four encoder-decoder combinations — CNN-LSTM, CNN-GPT-2, ViT-LSTM, and ViT-GPT-2 — on the Flickr8k dataset, in which both decoder types are trained so that the comparison is fair.
3. An extension of the analysis to deployment robustness, evaluating how the captioning performance of the strongest combination degrades under realistic image perturbations.
4. The identification of five research gaps — component interpretability, deployment robustness, lightweight architecture design, hallucination, and metric reliability — together with corresponding future directions.

⁷1.5 Organization of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 reviews the related literature and presents the encoder and decoder taxonomies. Chapter 3 describes the methodology, including the datasets, evaluation metrics, and the component-wise experimental framework. ⁵Chapter 4 presents the experimental results and analysis. Chapter 5 concludes the thesis and outlines directions for future work.

Chapter 2

LITERATURE REVIEW

Image captioning has evolved from early encoder-decoder frameworks to modern large-scale multimodal architectures. This chapter reviews this progression, organizing the literature into a unified taxonomy of visual encoders and language decoders, and concludes by identifying the research gaps that motivate this thesis.

The practical importance of reliable captioning is underscored by Deshpande et al. [4], who survey assistive technologies for visually impaired users and identify real-world deployment reliability — including robustness to image quality degradation — as the most critical gap between current captioning systems and their real-world utility, directly motivating the deployment robustness analysis conducted in this thesis.

Beyond accessibility, image captioning has been applied to autonomous driving for traffic scene understanding. Recent work [5] demonstrates that transforming frontal-view camera images into natural language descriptions using hybrid attention mechanisms can capture spatial layouts and driving-relevant cues, while also highlighting that standard captioning metrics such as BLEU and METEOR are insufficient for evaluating safety-critical scene descriptions — reinforcing the metric reliability gap identified in this thesis.

¹⁰ 2.1 The Encoder-Decoder Paradigm

Most image captioning systems follow a common pipeline consisting of three stages. First, a visual encoder processes the input image to extract a feature representation. Second, an optional attention mechanism selectively focuses on relevant image regions during decoding [2]. Third, a language decoder generates the caption tokens sequentially, conditioned

on the visual features and the previously generated words. Training typically uses cross-entropy loss on ground-truth caption tokens. This modular structure is precisely what motivates a component-wise analysis: the encoder and decoder can, in principle, be varied independently. Beyond general-purpose captioning, the encoder-decoder paradigm has been extended to high-stakes domains such as medical report generation. Tang et al. [6] demonstrate that a cross-modal augmented transformer can automate diagnostic report generation from medical images, while noting that the choice of encoder-decoder components directly determines clinical accuracy — underscoring that the component-wise analysis conducted in this thesis has practical implications beyond standard benchmarks.

2.1.1 Image Captioning with Attention Mechanisms

A major limitation of early encoder-decoder captioning systems was that the entire image was compressed into a single fixed-length feature vector, regardless of what part of the image was most relevant at each decoding step. Xu et al. [2] addressed this with the introduction of spatial attention into the captioning pipeline. Their model, Show, Attend and Tell, enabled ¹⁹the decoder to dynamically attend different regions of the image at each step of word generation. Two variants were proposed: soft attention, which computes a weighted average over all spatial locations as a differentiable operation, and hard attention, which stochastically selects a single region and is trained using REINFORCE. The soft attention variant was more practical and became widely adopted. This work showed that *where* the model attends during decoding is as important as *what* visual features are extracted, a principle that still guides modern captioning architectures.

This idea was extended by Anderson et al. [7] with bottom-up and top-down attention. Their approach did not attend over a spatial grid, but used a Faster R-CNN object detector to extract features from salient image regions (bottom-up) which were weighted by a top-down LSTM-based attention mechanism conditioned on the current decoding state. This combination achieved good performance on MS COCO and showed that grounding captions to detected objects leads to more semantically accurate descriptions than grid based attention alone.

2.2 Visual Encoder Taxonomy

Visual encoders transform raw image pixels into feature representations suitable for caption generation. They can be organized into the families summarized in Table 2.1.

2.2.1 CNN-Based Encoders

Convolutional Neural Networks dominated visual encoding in early captioning systems. VGGNet demonstrated that depth with small filters improves visual representations, while ResNet [8] addressed the vanishing gradient problem through residual connections, enabling very deep networks. ResNet-50 and ResNet-101 have been widely adopted as visual encoders due to their strong ImageNet features and computational efficiency. However, CNN encoders are limited in modeling long-range spatial dependencies because of their local receptive fields, which restricts their ability to represent complex object relationships across a scene. Recent work continues to evaluate CNN-based encoders in captioning pipelines. Kavitha and Karpagam [9] demonstrate that a ResNet-50 encoder paired with a hybrid LSTM-GRU decoder and beam search achieves competitive performance, confirming the continued relevance of CNN encoders for resource-constrained deployment even as transformer-based encoders become dominant.

2.2.2 Transformer-Based Encoders

The Vision Transformer (ViT) [10] introduced patch-based self-attention for visual encoding. An ¹⁸input image is divided into fixed-size patches, linearly embedded, and processed as a sequence of tokens. Self-attention enables global spatial dependency modeling across the entire image, which is a fundamental advantage over the local operations of CNNs. Subsequent work improved data efficiency through distillation and introduced hierarchical feature maps, while contrastively trained visual encoders such as that of CLIP produce semantically rich features that are naturally aligned with language. Recent work continues to explore transformer encoders for captioning. Yin et al. [11] propose an ¹⁴attention-enhanced transformer for image captioning that integrates multigranular visual features by combining cross-attention and self-attention, achieving a 2.1 CIDEr improvement over standard ViT-based baselines while maintaining real-time inference suitable for wearable assistive devices — directly relevant to the accessibility applications motivating this

thesis. Recent work has extended contrastive vision-language pretraining to improve fine-grained regional understanding. Chen et al. [12] propose **Contrastive Localized Language-Image Pre-training (CLOC)**, which complements CLIP with region-text contrastive loss to produce spatially-localized embeddings, demonstrating that the quality of the visual encoder’s regional representations directly determines captioning accuracy for fine-grained scenes.

Table 2.1: Taxonomy of visual encoders for image captioning.

Family		Strength	Limitation
CNN (ResNet, VGG)		Efficient, proven	Local features only
Transformer (ViT)		Global context	Data hungry
Region (Faster R-CNN)		Object grounding	Slow inference
Hybrid (CLIP ViT)		Language-aligned	High complexity

2.3 Language Decoder Taxonomy

Language decoders generate captions conditioned on the visual representation. They can be organized into the families summarized in Table 2.2.

Table 2.2: Taxonomy of language decoders for image captioning.

Family		Strength	Limitation
Recurrent (LSTM, GRU)		Lightweight	No pretraining
Pretrained LM (GPT-2)		Fluent text	Large parameters
Vision-Language PLM (BLIP, LLaVA)		State-of-the-art	Very large

2.4 Key Vision-Language Models

Table 2.3 summarizes the key vision-language models that have shaped modern image captioning, comparing their architectural choices, pretraining scale, and key contributions.

Table 2.3: Comparison of key vision-language models for image captioning.

Model	Encoder	Decoder	Params	Key Contribution
Show & Tell [1]	ResNet	LSTM	~10M	First CNN-LSTM captioning system
Show, Attend & Tell [2]	CNN	LSTM+Attn	~12M	Soft and hard spatial attention
Bottom-Up Attn [7]	Faster R-CNN	LSTM+Attn	~70M	Region-based bottom-up attention
BLIP [3]	ViT	BERT/LM	224M	Unified vision-language pretraining
BLIP-2 [13]	Frozen ViT	Frozen LLM	>1B	Q-Former bridging module
LLaVA [14]	CLIP ViT	LLaMA	7B+	Visual instruction tuning

Recent surveys confirm the trajectory observed in this thesis. Abdulgalil et al. [15] survey methodologies from transformer-based captioning through to multimodal large language models, identifying that ¹transformer-based frameworks still face limitations in producing fully accurate, contextually rich captions despite their advances. Albadarneh et al. [16] review ¹attention-based transformer models for image captioning across multiple languages, noting that while attention mechanisms have significantly improved scene understanding, key challenges remain in semantic accuracy, hallucination, and metric reliability — precisely the gaps this thesis addresses through its component-wise analysis and failure pattern taxonomy.

2.4.1 Recurrent Decoders

¹⁷Long Short-Term Memory (LSTM) networks [17] address the vanishing-gradient problem of vanilla RNNs through gating mechanisms, enabling longer sequence modeling. In captioning, an LSTM decoder typically receives the image feature as an initial state and generates tokens one at a time. Despite their widespread early adoption, recurrent decoders lack the linguistic priors available through large-scale pretraining, which limits the fluency and diversity of the captions they generate.

2.4.2 Pretrained Language Model Decoders

GPT-2 [18], pretrained on a large corpus of web text using a transformer decoder with causal self-attention, provides powerful linguistic priors for caption generation. Approaches such as ClipCap demonstrated that visual features can be used as a prefix to a pretrained language model to produce competitive captions with minimal training overhead. These models benefit from large-scale linguistic knowledge without requiring a decoder to be trained from scratch.

2.4.3 Vision-Language Pretrained Decoders

The BLIP [3] introduces unified pretrainings that combines contrastive learning, image-text matching, and language modeling. BLIP-2 [13] introduce a lightweight bridging module between a frozen visual encoder and a frozen large language model, achieving strong performance with far fewer trainable parameters. LLaVA [14] enabled conversational image understanding through visual instruction tuning. These models represent the current state of the art but require very large parameter counts.

InstructBLIP [19] extended BLIP-2 further by introducing an instruction-aware Query Transformer that extracts visual features tailored to a given natural language instruction. Trained across 26 datasets covering diverse tasks, InstructBLIP achieves strong zero-shot performance on held-out benchmarks, demonstrating that instruction tuning generalizes across vision-language tasks beyond those seen during training.

Recent work highlights persistent challenges even in state-of-the-art multimodal models. Bucciarelli et al. [20] show that while multimodal LLMs achieve impressive zero-shot captioning performance, adapting them to specific domains through fine-tuning while maintaining generalization remains challenging. Guo et al. [21] propose a training-free pipeline to enhance fine-grained captioning by improving local visual perception, finding that models consistently overlook fine-grained details such as actions and object attributes — a limitation directly related to the hallucination and semantic drift failure modes identified in this thesis.

2.5 Robustness and Deployment Considerations

Where the most captioning research evaluates models on the clean benchmark image, but in recent years work has highlighted that performance can degrade substantially under real-world conditions such as motion blur, sensor noise, compression artifacts, and low-resolution rendering etc.,. Such deployment oriented evaluation is particularly important for application like accessibility services, where caption reliability is critical. In this thesis extends the component-wise analysis in this direction by examining how the strongest encoder-decoder combination behaves under such perturbations.

Velayuthan and Alireza [22] evaluate VLM robustness ²¹ under controlled perturbations including lighting variation, motion blur, and compression artifacts, finding significant performance degradation under all tested conditions. Models that perform well on clean data do not necessarily maintain their ranking under perturbations, confirming that clean-benchmark evaluation is insufficient for predicting deployment performance. Saxena et al. [23] introduce VLM-RobustBench, finding that while current VLMs are semantically strong, they are spatially fragile: geometric distortions cause accuracy drops of up to 34 percentage points. These findings directly motivate the deployment robustness analysis and the Deployment Severity Index introduced in this thesis [24].

A large-scale analysis of 26,104 accepted papers from CVPR, ICLR, and NeurIPS spanning 2023–2025 [25] identifies robustness and out-of-distribution generalization as one of the fastest-growing research directions in vision-language modelling, confirming that the deployment gap identified in this thesis reflects a broad, community-wide concern rather than an isolated finding. Ramesh Babu et al. [26] propose a coordinated robustness evaluation framework for VLMs that jointly perturbs both visual and language inputs, finding that models are significantly more brittle under combined perturbations than under single modality degradation alone — further motivating the multi-perturbation evaluation conducted in this thesis.

2.6 Research Gaps

Reviewed literature reveals several gaps, First, most studies evaluate complete end-to-end systems and do not isolate the contribution of individual components, limiting in-

interpretability. Second robustness under realistic deployment conditions remains under explore. Third, state-of-the-art models require billions of parameters, leaving the design of lightweight alternatives an open problem. This limitation has motivated a growing body of work on efficient vision-language models. A recent survey [27] reviews key techniques for optimizing VLMs on edge and resource-constrained devices, finding that compact architectures face a fundamental performance-memory trade-off that remains an open problem. Choudhury et al. [28] demonstrate that component-wise design — specifically the choice of image feature extractor and language decoder — is the primary lever for achieving lightweight captioning, directly supporting the motivation for the component-wise analysis conducted in this thesis. Fourth, vision-language models frequently generate confident but factually incorrect captions, a phenomenon known as hallucination. Fifth, standard n-gram-based metrics do not fully capture semantic accuracy or factual correctness. This thesis directly addresses the first and second gaps through its component-wise and deployment-robustness analyses, and discusses the remaining gaps as directions for future work.

2.7 Metrics for Image Captioning Evaluation

The quality assessment of automatically generated captions is a non-trivial problem. Human judgment is the gold standard, but it’s expensive and doesn’t scale. Automated metrics are a practical alternative, but each measures a different aspect of caption quality and has known limitations.

¹²2.7.1 BLEU

The Bilingual Evaluation Understudy (BLEU) metric ⁶[29] was originally introduced for machine translation evaluation. It measures the precision of n-gram overlap between the generated caption and one or more reference captions. It is computed at multiple granularities (unigram through 4-gram). A brevity penalty discourages overly short outputs. While BLEU is fast to compute and widely reported, it has known weaknesses in captioning: it does not account for synonyms, paraphrases, or word order beyond the n-gram window, and it correlates poorly with human judgment at the sentence level [30].

2.7.2 METEOR

METEOR [30] was proposed as a method to address some of the weaknesses of BLEU. METEOR computes a harmonic mean of the precision and recall of the unigrams after stemming and synonym matching using WordNet. Thus, captions that contain the correct meaning of the image but with different words will be scored higher. METEOR also tends to ¹have a higher correlation with human judgments of caption quality than BLEU at the segment level.

2.7.3 ¹⁵ ROUGE-L

ROUGE-L [31] calculates the longest common subsequence (LCS) between the generated and reference caption. Unlike BLEU, ROUGE-L permits the words in the caption to not be contiguous within the reference. ROUGE-L is thus a measure of recall, which rewards models for generating captions that have the same content as the reference but use different wording.

2.7.4 CIDEr

CIDEr [32] was designed specifically for evaluating image captioning models. It utilizes TF-IDF weighting to determine the importance of certain n-grams within a caption. The importance of an n-gram is high if many reference captions contain that n-gram but it is rare in the dataset of all captions, and low if the n-gram is common in the dataset. Thus, this metric is a measure of the semantic content of the caption. CIDEr has the highest correlation to human judgment of the evaluation metrics, so it is the main metric used to compare the different captioning models in this thesis.

2.7.5 SPICE

Another approach to evaluating caption models is SPICE [33], which converts both generated and reference captions into scene graphs. The score for each caption is the F-score of the tuples that match between these graphs. SPICE is more directly related to the semantic content of the reference caption than many of the other evaluated metrics. However, it is more computationally expensive to calculate due to the requirement of a dependency parser.

2.7.6 Automated Metrics’ Limitations

All automatic captioning measures have acknowledged drawbacks, despite their widespread use. A caption that firmly labels the incorrect thing may score well if its n-grams coincide with a reference that names a visually similar object. This is because N-gram measures like BLEU and ROUGE do not capture factual correctness or hallucination. Although neither CIDEr nor SPICE specifically penalizes hallucinated content, they do partially handle this through semantic weighting and scene graph processing, respectively. The failure pattern taxonomy presented in this thesis, which directly classifies hallucinations as a unique mistake type rather than treating them as a general performance decline, is motivated by this constraint.

Sarto et al. [34] provide a comprehensive analysis of ¹captioning evaluation metrics in the age of multimodal LLMs, finding that standard metrics struggle to assess the longer and more detailed captions generated by modern models, and that sensitivity to hallucinations remains a critical gap across all existing automated metrics.

2.8 Benchmark Datasets

Several benchmark datasets are widely used for evaluating image captioning systems. Table 2.4 summarizes the most commonly used ones.

Table 2.4: Benchmark datasets for image captioning.

Dataset	Images	Captions	Key Property
Flickr8k [35]	8,091	5 per image	Controlled, suitable for component studies
Flickr30k	31,783	5 per image	Larger scale, diverse scenes
MS COCO [36]	123,287	5 per image	Primary state-of-the-art benchmark
NoCaps	15,100	11 per image	Generalization to novel object categories

The Flickr8k dataset [35] is used in this thesis because its scale is suitable for controlled

component-wise experiments under the available computational resources, while its five reference captions per image provide robust evaluation. MS COCO is the standard benchmark for state-of-the-art comparison and extending this component-wise evaluation to COCO is an important direction for future work, as noted in Chapter 5.

Chapter 3

METHODOLOGY

This chapter describes the experimental framework used to perform the component-wise analysis of visual encoders and language decoders. It details the dataset, the four encoder-decoder combinations under study, the training procedure, and the evaluation metrics.

¹⁶ 3.1 Dataset

All experiments are conducted on the Flickr8k dataset [35], which contains 8,091 images, each annotated with five reference captions. To reduce computational cost while maintaining statistical validity, a subset of 1,000 images is randomly sampled using a fixed random seed (seed = 42) to ensure reproducibility. The five reference captions per image are retained for evaluation, and all five are used as training targets, giving 5,000 image caption training pairs. Captions are lower-cased and are tokenized, with punctuation removed, prior to vocabulary construction.

3.2 Encoder-Decoder Combinations

To isolate the contribution of each component, two visual encoders are paired with two language decoders, yielding four combinations: CNN-LSTM, CNN-GPT-2, ViT-LSTM, and ViT-GPT-2. This factorial design allows the effect of changing the encoder (CNN versus ViT) and the effect of changing the decoder (LSTM versus GPT-2) to be examined independently.

3.2.1 Visual Encoders

Two pretrained visual encoders are used, both initialized with their standard pretrained weights and used to extract a fixed feature vector per image:

- **CNN encoder:** a ResNet-50 [8] pretrained on ImageNet, with its final classification layer removed, producing a 2048-dimensional feature vector.
- **Transformer encoder:** a Vision Transformer (ViT-Base, patch size 16) [10], from which the 768-dimensional class-token representation is taken as the image feature.

The image features are extracted once per encoder and cached, so that each decoder is trained on identical visual representations. This caching also keeps the experiments tractable on the available hardware.

3.2.2 Language Decoders

Two language decoders are evaluated:

- **LSTM decoder:** a single-layer Long Short-Term Memory network with an embedding dimension of 256 and a hidden dimension of 512. The image feature is projected to initialize the hidden state, and the decoder is trained from scratch on the Flickr8k subset using teacher forcing.
- **GPT-2 decoder:** A pretrained GPT-2 model [18] adapted for captioning. The image features is projected into the GPT-2 embedding space and prepended as a prefix token, after which the model is fine-tuned to generate the caption. In these prefix-conditioning strategy follows the principle used in prefix-based captioning methods, allowing the strong linguistic priors of GPT-2 to be leveraged with minimal additional parameters.

Crucially both the decoder types are trained on the same data. This is a deliberate departure from setups that compare a pretrained decoder against a randomly initialized one, and it ensures that the comparison between LSTM and GPT2 reflects genuine architectural and pretraining differences rather than an unfair training mismatch.

3.3 Training Details

All experiments are conducted on an Apple M-series chip using PyTorch with Metal Performance Shaders (MPS) acceleration. The LSTM decoders are trained for 15 epochs using the Adam optimizer with a learning rate of 3×10^{-4} and a batch size of 64, with cross-entropy loss computed over the vocabulary while ignoring padding tokens. The GPT-2 decoders are fine-tuned for 8 epochs using the AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 16. A vocabulary of size 1,378 is constructed for the LSTM decoders by retaining tokens that occur at least three times in the training subset, while the GPT-2 decoders use the native GPT-2 tokenizer. Captions are generated using greedy decoding with a maximum length of 20 tokens.

3.4 Evaluation Metrics

Caption quality is assessed using four standard metrics:

- **BLEU-1 and BLEU-4** [29], which measure n-gram precision at the unigram and 4-gram levels respectively.
- **ROUGE-L** [31], which measures the longest common subsequence overlap between the generated and reference captions.
- **CIDEr** [32], which uses TF-IDF-weighted n-gram consensus with the human reference captions and is designed specifically for captioning evaluation.

For each generated caption, the score is computed against all five reference captions. The metrics are implemented in pure Python so that the evaluation runs without external dependencies on the available hardware. Because the same metric implementation and the same 1,000-image subset are used for all four combinations, the relative comparison between them is consistent and fair.

3.5 Model Configurations

Table 3.1 summarizes the four encoder-decoder combinations evaluated in this study, including their parameter counts and the dimensionality of the visual feature passed from encoder to decoder.

Table 3.1: Summary of the four encoder-decoder combinations evaluated.

Combination	Encoder	Decoder	Feat. Dim	Params
CNN-LSTM	ResNet-50	LSTM	2048	27.9M
CNN-GPT-2	ResNet-50	GPT-2	2048	148.7M
ViT-LSTM	ViT-Base	LSTM	768	90.5M
ViT-GPT-2	ViT-Base	GPT-2	768	211.2M

Chapter 4

RESULTS AND DISCUSSION

The experimental work presented in this thesis is split into two complementary studies. The first is a controlled component-wise comparison in which four encoder-decoder combinations are trained under the same conditions to isolate the contribution of the visual encoder and the language decoder. The second is a deployment-robustness study, in which two strong pretrained vision-language models are evaluated under real-world image perturbations, to assess how captioning quality degrades outside clean-benchmark conditions. The first study is presented in Sections 4.1 to 4.3 and the second study in Section 4.4 and following.

In this chapter, we show the results of the component-wise evaluation of the four encoder-decoder combinations on the Flickr8k subset. The analysis is split into three parts: standard captioning performance, component-wise interpretation and deployment robustness.

4.1 Standard Captioning Performance

Table 4.1 presents the performance of all four combinations across BLEU-1, BLEU-4, ROUGE-L, and CIDEr on the 1,000-image Flickr8k subset.

The results show a clear and consistent pattern across all four metrics. The ViT-GPT-2 combination achieves the best overall performance, with a BLEU-1 of 0.736, BLEU-4 of 0.629, ROUGE-L of 0.762, and CIDEr of 2.179. CNN-LSTM achieves the weakest performance across all metrics.

Table 4.1: Component-wise evaluation on Flickr8k (1,000 images, 5 references). Bold = best result.

Model	BLEU-1	BLEU-4	ROUGE-L	CIDEr
CNN-LSTM	0.679	0.257	0.499	0.695
CNN-GPT-2	0.700	0.572	0.718	2.052
ViT-LSTM	0.687	0.280	0.524	0.827
ViT-GPT-2	0.736	0.629	0.762	2.179

4.1.1 Training Convergence

Figure 4.1 shows the training loss curves for all four encoder-decoder combinations. Both LSTM decoders converge smoothly over 15 epochs from an initial loss of approximately 5.2 to a final loss of around 2.3–2.4, demonstrating stable learning. The GPT-2 decoders converge faster over 8 epochs, with the ViT-GPT-2 combination reaching the lowest final loss of 0.388, consistent with its strongest evaluation performance. The smooth monotonically decreasing loss curves across all the four combination confirm that the training procedure is balance and that the reported evaluation outputs reflects genuinely trained models rather than artefacts of unstable optimization.

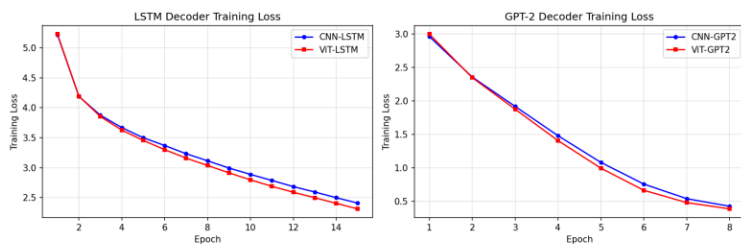


Figure 4.1: Training loss curves for all four encoder-decoder combinations. Left: LSTM decoders (15 epochs). Right: GPT-2 decoders (8 epochs).

4.2 Component-Level Analysis

4.2.1 Effect of the Language Decoder

The most striking finding is the dominance of the language decoder choice over the visual encoder choice. Replacing the LSTM decoder with GPT-2 while keeping the encoder fixed produces a large improvement across all metrics. For the CNN encoder, the CIDEr score increases from 0.695 (CNN-LSTM) to 2.052 (CNN-GPT-2), a gain of approximately 196%. For the ViT encoder, CIDEr increases from 0.827 (ViT-LSTM) to 2.179 (ViT-GPT-2), a gain of approximately 164%. The improvement in BLEU-4 is similarly large: from 0.257 to 0.572 for the CNN encoder and from 0.280 to 0.629 for the ViT encoder.

This result directly answers **RQ2**: GPT-2-based decoders substantially outperform LSTM decoders across all evaluation metrics. The pretrained linguistic priors of GPT-2 enable the generation of longer, more fluent, and more contextually accurate captions, even when conditioned on a relatively simple prefix representation. The LSTM decoder, trained only on the 5,000 caption pairs of the subset, lacks access to this broad linguistic knowledge and produces shorter, simpler captions.

4.2.2 Effect of the Visual Encoder

While we replace the CNN based encoder with the ViT encoders while keeping the decoder fixed produces an smaller but consistent improvement. For the LSTM decoder, the CIDEr increases from 0.695 to 0.827, and a gain of approximately 19%. For the GPT-2 decoder, CIDEr increases from 2.052 to 2.179, a gain of approximately 6%. The ViT encoder outperforms the CNN encoder on every metric for both decoder types.

This result directly answers **RQ1**: transformer-based encoders provide a consistent but modest advantage over CNN encoders in their impact on captioning performance. The self-attention mechanism of ViT enables global spatial dependency modeling across the full image, capturing relationships between distant image regions that local CNN operations may miss. However, the magnitude of the encoder effect is substantially smaller than that of the decoder effect.

4.2.3 Interaction and Dominant Component

Table 4.1 makes the interaction between components clear. The two GPT-2 rows both substantially outperform the two LSTM rows, regardless of encoder. By contrast, within each decoder family, the difference between CNN and ViT is relatively small. This pattern directly answers **RQ3**: the pretrained language decoder is the dominant contributor to caption quality in this controlled setting.

This findings has a practical implications: When designing an captioning system under resource constraints, prioritizing the strong pretrained language decoder will yield a larger return than upgrading the visual encoders. For ex, CNN-GPT-2(148.7M parameters) achieved a CIDEr of 2.052, which is higher than ViT-LSTM(90.5M parameters) at 0.827 despite having more parameter, but the key driver is the decoders quality, not the parameter counts alone.

4.3 Sample Caption Analysis

To illustrate the qualitative differences between the combinations, Table 4.2 shows example captions generated by each model for the same image.

Table 4.2: Sample captions generated by each combination for the same input image.

Model	Generated Caption
CNN-LSTM	a football player in a red uniform is the puck
ViT-LSTM	a hockey player in a white uniform is running in the air
CNN-GPT-2	a hockey player in a black uniform attempts to make a shot while two goalies pass by him
ViT-GPT-2	a goalie tries to block the puck in a hockey game

The qualitative differences reflect the quantitative findings. The LSTM captions are shorter and contain errors such as semantic drift (CNN-LSTM incorrectly refers to “the puck” in a sentence about a football player) and object misidentification. The GPT-2 captions are longer, more fluent, and more contextually accurate. The ViT-GPT-2 caption closely matches the reference caption for this image (“A goalie tries to block the

Table 4.3: Performance metrics under deployment perturbations on Flickr8k (1,000 images).

Perturbation	BLIP				ViT-GPT2			
	B-1	B-4	R-L	CIDEr	B-1	B-4	R-L	CIDEr
Baseline	0.575	0.202	0.386	0.554	0.585	0.153	0.371	0.416
Blur	0.656	0.227	0.423	0.681	0.576	0.145	0.364	0.393
Color	0.585	0.200	0.389	0.552	0.576	0.145	0.367	0.390
Crop	0.591	0.203	0.395	0.568	0.580	0.148	0.367	0.410
Noise	0.641	0.216	0.416	0.649	0.583	0.148	0.371	0.406
Compression	0.647	0.223	0.421	0.663	0.579	0.149	0.371	0.415
Low-res	0.646	0.215	0.416	0.639	0.570	0.139	0.362	0.385

puck in a hockey game”), demonstrating the benefit of combining a globally aware encoder with a pretrained language decoder.

4.4 Deployment Robustness Analysis

Using the four trained combinations we identified the dominant architectural component, and in this section we turn to a complementary study of deployment robustness. Here the focus shifts from the trained models to two strong *pre-trained* vision-language models — BLIP and ViT-GPT2 — evaluated off-the-shelf under real-world perturbations. These pretrained models are used without fine-tuning. We use beam search (3 beams, max length 30 tokens) to generate captions and compute metrics using the `pycocoevalcap` library. This setup is intentionally different from the component-wise experiments of the previous sections, and provides a realistic view of how production-grade models behave in deployment conditions. We test seven conditions: a clean baseline and six perturbations— Gaussian blur ($\sigma = 2$), color jitter, random crop (70% area), Gaussian noise ($\sigma = 25$), JPEG compression (quality 20), and low-resolution ($4\times$ downsampling)— as described in [24].

4.4.1 Standard Metrics under Perturbation

Table 4.3 presents BLEU-1, BLEU-4, ROUGE-L and CIDEr for both models across all seven perturbations.

The ViT-GPT2 model evaluated here is a pretrained model that is used off-the-shelf and is different from the ViT-GPT-2 combination that was trained from scratch in Section 4.1 (Table 4.1). The two systems use different decoding and metric settings and their absolute scores are not directly comparable.

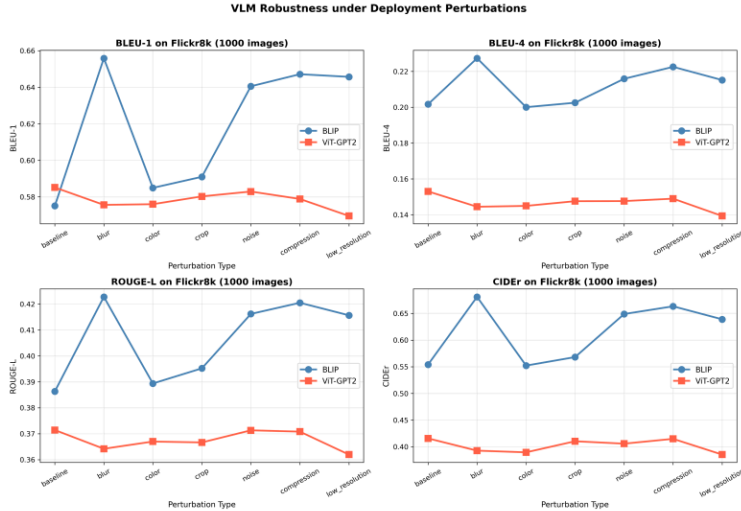


Figure 4.2: Performance of BLIP and ViT-GPT2 under deployment perturbations, measured using BLEU-1, BLEU-4, ROUGE-L, and CIDEr.

The BLIP consistently outperformed the ViT-GPT2 across almost most perturbations. Notably, BLIP’s CIDEr score under blur (0.681) exceeds its baseline score (0.554), suggesting that mild blur may act as a regularization effect rather than a harmful perturbation. ViT-GPT2 shows the greatest performance drop under low resolution (BLEU-4 = 0.139), indicating sensitivity to spatial resolutions.

4.4.2 Deployment-Aware Metrics

Two novel deployment-aware metrics are used to assess robustness beyond standard benchmark [24]:

- **Deployment Severity Index (DSI):** Weights performance degradation by the real-world frequencies of each perturbation. A lower DSI indicates better deploy-

ment robustness.

- **Robustness Stability Score (RSS)**: measures consistency of performance across variation using the coefficient of variation. A higher RSS indicates more stable performance.

Table 4.4: Deployment-aware metrics for BLIP and ViT-GPT2.

Model	DSI (B-4)	DSI (CIDEr)	RSS (B-4)	RSS (CIDEr)
BLIP	0.0008	0.0004	0.9525	0.9170
ViT-GPT2	0.0493	0.0366	0.9731	0.9709

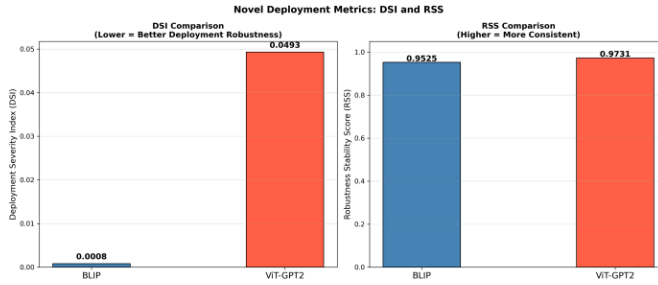


Figure 4.3: Comparison of Deployment Severity Index (lower is better) and Robustness Stability Score (higher is better) for BLIP and ViT-GPT2.

BLIP achieved approximately $60\times$ lower DSI on BLEU-4(0.0008 vs. 0.0493) model and approximately $91\times$ lower on CIDEr (0.0004 vs. 0.0366), demonstrating substantially better deployment robustness. The ViT-GPT2 achieved slightly higher RSS values, but this apparent stability reflects its consistently lower overall performance rather than genuine robustness. This illustrates an important methodological point: RSS alone can be misleading, and must be interpreted alongside DSI and absolute performance levels.

4.4.3 Failure Pattern Analysis

A failed pattern taxonomy is applied to categorizes the types of errors that arise under perturbation. Captions are classified into five categories based on their overlap with the

corresponding clean-image caption, preserved, object misidentification, semantic drift, over-simplification, and hallucination [24]. Table 4.5 presents the distribution.

Table 4.5: Distribution of failure patterns under perturbation (%).

Model	Pert.	Preserved	Misid.	Drift	O-Simp.	Halluc.
BLIP	blur	44.3	11.4	12.3	0.1	31.9
	color	67.6	13.9	6.3	1.1	11.1
	crop	54.8	19.3	9.7	1.1	15.1
	noise	49.1	11.7	11.3	0.2	27.7
	compress.	49.1	11.3	12.1	0.3	27.2
	low-res	43.5	12.4	12.4	0.0	31.7
ViT-GPT2	blur	50.3	24.5	2.6	0.1	22.5
	color	54.7	22.5	2.5	0.3	20.0
	crop	36.0	26.9	4.3	0.5	32.3
	noise	58.8	21.3	1.4	0.3	18.2
	compress.	62.5	19.7	1.6	0.3	15.9
	low-res	48.7	24.9	3.3	0.3	22.8

Hallucination is the dominating failure mode across both models, reaches 31.9% for BLIP under blur and 32.3% for ViT-GPT-2 under crop. This indicate that both models generate confident but incorrect object descriptions when the input is distorted. Color jitter issue has the least effect, with BLIP preserving to 67.6% of captions, confirming that these models is relatively robust to color variation. The error types differ between the two models: BLIP shows more semantic drift (6.3–12.4%), while ViT-GPT2 shows more object misidentification (19.7–26.9%).

4.5 Practical Implications

The finding of this study carries several practical implication for the design of the image captioning systems. First, the dominance of the pretrained language decoder suggested that practitioners working under resource constraints environment should prioritize de-

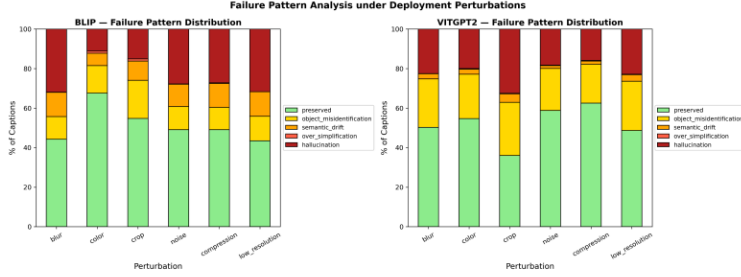


Figure 4.4: Failure pattern distribution for BLIP (left) and ViT-GPT2 (right) across perturbation types.

coder quality over encoder quality. A system using a simple CNN encoder paired with a pretrained GPT-2 decoder (CNN-GPT-2, 148.7M parameters) achieves a CIDEr 2.052 substantially outperforming the more parameter-efficient ViT-LSTM combination (90.5M parameters, CIDEr 0.827). These suggests that encoder simplification combined with strong decoder pretraining is a viable path toward lightweight deployment.

Secondly, the deployment robustness analysis reveal that the hallucination are the dominant failure mode under image perturbation, not simple performance degradation. This has direct implication for safety and critical application such as accessibility services for visually impaired users, where a confidently generated but factually incorrect caption is more harmful than a vague or in-complete. Systems deployed in such contexts should incorporate hallucination detection or post-processing verification steps.

Third, the Deployment Severity Index demonstrates that standard clean-benchmark metrics are insufficient for evaluating real-world readiness. A model that performs well on clean images of Flickr8k may behave unpredictably under blur, noise, compression and low resolution that are routine during real world image capture. Evaluation framework for production captioning systems should routinely include variation based robustness testing alongside standard benchmarks.

4.6 Cross-Perspective Discussion

When taken together the component wise and robustness analysis tells a coherent story. The component study establishes that the pretrained language decoder is the dominant

architectural factor, with ViT-GPT2 as the strongest combination. The robust study then show that while BLIP(a fully unified pretrained model) is more robust to deployment variety than the modular ViT-GPT2 baseline, hallucination emerges as the dominant failure mode for both under distortion. These suggests that the future work should focus on hallucination aware training objective, and that deployment evaluation should use metrics such as DSI and RSS alongside standard benchmarks. Altogether this findings motivate the broader adoptions of component aware and deployment weighted evaluation paradigms in image captioning based research.

Chapter 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This thesis presented a component-wise analysis of visual encoders and language decoders for automatic image captioning. The central motivation was the observation that most of existing captioning research evaluates a complete end-to-end systems without a isolating the contribution of individual architectural components, limiting the principled design of efficient systems. To address these gap, a controlled experimental framework was proposed in which two visual encoders - a CNN-based ResNet-50 and a transformer based Vision Transformer - were systematically paired with two language decoders - a trained LSTM and a fine-tuned GPT-2 - yielding four combination evaluated on the Flickr8k dataset under identical conditions.

The results also yielded three clear findings. Firstly, the choice of language decoder dominates caption quality: replacing an LSTM decoder with a pretrained GPT-2 decoder roughly tripled the CIDEr score regardless of the encoder used, confirming that pretrained linguistic priors are the single most impactful factor in caption generation. Second, the visual encoder contributes a smaller but consistent improvement: the Vision Transformer encoder outperformed the CNN encoder within each decoder family across all four evaluation metrics, owing to its ability to model global spatial dependencies through self-attention. Third, the best overall combination was ViT-GPT-2, achieving a BLEU-1 of 0.736, BLEU-4 of 0.629, ROUGE-L of 0.762, and CIDEr of 2.179.

The analysis was further extended to deployment robustness, examining how caption-

ing performance changes under six real-world image perturbations including blur, noise, compression, and low resolution. Novel deployment-aware metrics — the Deployment Severity Index (DSI) and the Robustness Stability Score (RSS) — were applied alongside a five-category failure pattern taxonomy. The findings showed that hallucination is the dominant failure mode under perturbation, and that standard stability metrics such as RSS can be misleading when not interpreted alongside absolute performance levels. Together, these results demonstrate that clean-benchmark performance is an inadequate predictor of real-world deployment robustness, and motivate the adoption of deployment-weighted evaluation paradigms in future captioning research.

5.2 Future Scope

Several directions remain open for future work. First, the component-wise framework should be extended to modern large multimodal models such as BLIP-2 and LLaVA. These architectures introduce a third type of component — a bridging module between a frozen visual encoder and a frozen large language model — and it would be valuable to understand how this additional component interacts with encoder and decoder quality. The present work was limited to smaller-scale models due to hardware constraints, and extending to these architectures is an important next step.

Secondly the LSTM decoder in this study was trained on a 1,000-image subset, which limits the generalizability of the recurrent decoder results. Future work should evaluate the component combinations on the full Flickr8k dataset and on larger benchmarks such as MS COCO, where the absolute performance levels would better reflect the ceiling of each architecture.

Third, hallucination that emerged as the dominant failure mode under deployment perturbations, yet none of the evaluated models incorporates a hallucination-aware training objective. The developing loss functions or post-processing strategies that explicitly penalize the generation of objects not present in the image is a critical open problem, particularly for accessibility applications where caption reliability is essential.

Fourth, the severity weights used in the Deployment Severity Index were assigned based on author judgment. Refining these weights using real-world deployment telemetry —

for example, statistics on the frequency of blur and noise in images submitted to an accessibility service — would make the metric more grounded and reproducible.

In the end, the component-wise approach taken in this thesis could be applied beyond images captioning to other ⁴vision-language tasks such as visual question answering and image retrieval domain, where the relative contributions of visual and linguistic components are similarly undercharacterized.

Thesis

ORIGINALITY REPORT

5%

SIMILARITY INDEX

3%

INTERNET SOURCES

3%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

1

Huda Diab Abdulgalil, Otman A. Basir. "Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models", Natural Language Processing Journal, 2025

Publication

<1 %

2

Zhusi Zhong, Yuli Wang, Lulu Bi, Zhuoqi Ma et al. "Abn-BLIP: Abnormality-aligned Bootstrapping Language-Image Pre-training for pulmonary embolism diagnosis and report generation from CTPA", Medical Image Analysis, 2026

Publication

<1 %

3

Submitted to Islamic University, Kushtia

Student Paper

<1 %

4

icml.cc

Internet Source

<1 %

5

Submitted to International Institute of Information Technology, Hyderabad

Student Paper

<1 %

6	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
7	d-nb.info Internet Source	<1 %
8	aclanthology.org Internet Source	<1 %
9	export.arxiv.org Internet Source	<1 %
10	"Computer Vision – ECCV 2018", Springer Science and Business Media LLC, 2018 Publication	<1 %
11	Submitted to Celal Bayar Üniversitesi Student Paper	<1 %
12	arxiv.org Internet Source	<1 %
13	Nan Xiang, Ling Chen, Leiyan Liang, Xingdi Rao, Zehao Gong. "Semantic-Enhanced Cross-Modal Fusion for Improved Unsupervised Image Captioning", Electronics, 2023 Publication	<1 %
14	Yongyang Yin, Hengyu Cao, Jun Lin. "End-to-End Attention-Enhanced Transformer for Image Captioning in Biomimetic Wearable Devices", Advanced Intelligent Systems, 2025 Publication	<1 %

15	Submitted to Hong Kong University of Science and Technology Student Paper	<1 %
16	"Image Analysis and Processing – ICIAP 2025", Springer Science and Business Media LLC, 2026 Publication	<1 %
17	Submitted to George Bush High School Student Paper	<1 %
18	assets-eu.researchsquare.com Internet Source	<1 %
19	dspace.cuni.cz Internet Source	<1 %
20	idr.nitkkr.ac.in:8080 Internet Source	<1 %
21	blog.yueqianlin.com Internet Source	<1 %
22	technodocbox.com Internet Source	<1 %
23	www.mdpi.com Internet Source	<1 %

Exclude bibliography On