Muskan Motwani

muskan_24rco01__

Thesis

Document Details

Submission ID

trn:oid:::27535:140930443

Submission Date

May 29, 2026, 1:22 PM GMT+5:30

Download Date

May 29, 2026, 1:26 PM GMT+5:30

File Name

muskan_24rco01__.pdf

File Size

4.3 MB

28 Pages

10,720 Words

56,870 Characters

8% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

67 Not Cited or Quoted 7%
Matches with neither in-text citation nor quotation marks

9 Missing Quotations 1%
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

5% Internet sources
5% Publications
5% 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

- **67** Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- **9** Missing Quotations 1%
  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

5% Internet sources
5% Publications
5% Submitted works (Student Papers)

Top Sources

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

1	Internet	arxiv.org	2%
2	Publication	Nayu Liu, Xian Sun, Hongfeng Yu, Fanglong Yao, Guangluan Xu, Kun Fu. "M^{2}D...	<1%
3	Student papers	Universitat Politècnica de València on 2026-04-24	<1%
4	Student papers	University of Edinburgh on 2025-05-08	<1%
5	Internet	aclanthology.org	<1%
6	Publication	Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdon...	<1%
7	Publication	Yi Cheng, Hongyuan Zhu, Ying Sun, Cihan Acar, Wei Jing, Yan Wu, Liyuan Li, Chest...	<1%
8	Publication	Jinyu Li, Fuwei Zhang, Shujin Lin, Fan Zhou, Ruomei Wang. "MIM: Lightweight Mul...	<1%
9	Publication	"Computer Vision – ECCV 2024", Springer Science and Business Media LLC, 2025	<1%
10	Student papers	Özyegin Üniversitesi on 2026-04-21	<1%

11 Student papers

BITS, Pilani-Dubai on 2025-12-15 <1%

12 Student papers

National Institute of Technology Jamshedpur on 2026-05-19 <1%

13 Student papers

BITS, Pilani-Dubai on 2025-12-18 <1%

14 Student papers

University of Birmingham on 2025-09-01 <1%

15 Internet

publications.polymtl.ca <1%

16 Publication

"Artificial Intelligence", Springer Science and Business Media LLC, 2022 <1%

17 Student papers

Consorcio CIXUG on 2024-01-04 <1%

18 Student papers

Macao Polytechnic Institute on 2025-03-19 <1%

19 Student papers

University Politehnica of Bucharest on 2024-09-06 <1%

20 Publication

Xichen Tan, Yuanjing Luo, Yunfan Ye, Guang Yu, Chengyu Wang, Fang Liu, Zhipin... <1%

21 Internet

stax.strath.ac.uk <1%

22 Publication

Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, Wenwu Zhu. "A Closer ... <1%

23 Publication

Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, Liqiang Nie. "Video Moment Localizati... <1%

24 Internet

d197for5662m48.cloudfront.net <1%

25 Publication

dos Santos, Raphael Ruschel. "Advancing Dynamic Scene Understanding: Improvi... <1%

26 Internet

internationalaisafetyreport.org <1%

27 Internet

openaccess.thecvf.com <1%

28 Publication

"Advanced Intelligent Computing Technology and Applications", Springer Science... <1%

29 Publication

Aleksandar Fontana, Marco Simoni. "Chapter 18 Unmasking Model Behavior: How... <1%

30 Student papers

Copenhagen Business School on 2026-05-15 <1%

31 Publication

Gongmian Wang, Xun Jiang, Ning Liu, Xing Xu. "Language-enhanced object reaso... <1%

32 Publication

Hamman, Faisal Adamu. "Trustworthy and Explainable Machine Learning Using I... <1%

33 Publication

Junxian Wu, Yujia Zhang, Xiaoguang Zhao. "Visual enhanced hierarchical network... <1%

34 Student papers

KTH - The Royal Institute of Technology on 2015-12-03 <1%

35 Student papers

King Abdullah University of Science and Technology (KAUST) on 2025-08-11 <1%

36 Publication

Kumar, Yaman. "Behavior as a Modality.", State University of New York at Buffalo <1%

37 Publication

Shu Luo, Qiwei Ma, Jiawei Wang, Da Cao, Shaofei Lu. "AutoVMR: An Autonomous ... <1%

38 Student papers

University of Birmingham on 2026-04-20 <1%

39 Publication

Xin Wang, Xiaohan Lan, Wenwu Zhu. "Video Grounding and Its Generalization", S... <1%

40 Internet

blog.yueqianlin.com <1%

41 Internet

impa.usc.edu <1%

42 Internet

ntnuopen.ntnu.no <1%

43 Internet

ovsa.njit.edu <1%

44 Internet

theses.hal.science <1%

45 Internet

trepo.tuni.fi <1%

46 Internet

unige.iris.cineca.it <1%

47 Internet

www.fh-salzburg.ac.at <1%

48 Publication

"Computer Vision – ECCV 2020", Springer Science and Business Media LLC, 2020 <1%

49 Publication

Wen Wang, Jian Cheng, Siyu Liu. "DCT-net: A deep co-interactive transformer net... <1%

50 Publication

Zixi Jia, Minglin Dong, Jingyu Ru, Lele Xue, Sikai Yang, Chunbo Li. "STCM-Net: A sy... <1%

ABSTRACT

Finding specific moments in videos based on natural language descriptions is a core challenge in video understanding. This problem, known as video moment retrieval, requires a system to identify the exact start and end times in an untrimmed video that correspond to a given text query. Current state-of-the-art methods such as SG-DETR use DETR-style architectures where multiple decoder queries each predict a temporal span in a single forward pass. These models are trained with bipartite matching: the Hungarian algorithm assigns each ground-truth moment to exactly one predicted span, and only that matched prediction receives the span regression loss (L1 + GIoU). The remaining predictions are only pushed toward a "no-object" classification and never receive direct localization feedback. As a result, the model does not learn to compare or rank its own predictions against each other.

In this thesis, we introduce DualRS, a reinforcement learning framework that provides a learning signal to all predictions simultaneously. We apply Group Relative Policy Optimization (GRPO), originally developed for language model alignment, to the set of proposals from a single forward pass. Each proposal is scored using a reward function that measures its quality (semantic relevance and saliency overlap), and the model is updated to increase the likelihood of high-reward predictions while decreasing that of low-reward ones. Because the decoder already outputs multiple spans deterministically, no repeated sampling or stochastic rollouts are needed because the group structure is inherent to the architecture.

We further introduce a dual reward function that combines two complementary signals: semantic relevance and temporal saliency. The semantic reward uses Natural Language Inference (NLI) to measure how well the content within a predicted time span logically supports the query text, computed using DeBERTa-v3 on auto-generated speech transcripts. The saliency reward leverages crowd-sourced human annotations indicating which clips are visually important. Together, these two signals provide a richer training objective than either alone: relevance guides what to retrieve while saliency guides how engaging the retrieved content should be.

We also introduce cross-policy exploration, where a frozen copy of the original model contributes additional proposals to the group. This maintains exploration diversity as training progresses and prevents the model's proposals from collapsing into similar temporal regions.

Experiments on the QVHighlights benchmark demonstrate that DualRS achieves 73.42% R1@0.5 on the validation split, surpassing the current state-of-the-art SG-DETR (+0.62% R1@0.5). On the held-out test split, DualRS achieves 75.15% R1@0.5, outperforming SG-DETR by +0.95%. These improvements generalize to other datasets including Charades-STA (71.83% R1@0.5, +1.63% over SG-DETR) and TACoS (31.14% R1@0.5, +1.24% over SG-DETR).

Keywords: Video Moment Retrieval, Reinforcement Learning, Group Relative Policy Optimization, Natural Language Inference, Dual Reward.

CHAPTER 1

INTRODUCTION

1.1 Background

We view billions of hours of web video but it is a hard process to select one specific scene. Video Moment Retrieval (VMR) addresses this precisely by enabling users to write and query about what they want to find in the video. VMR enables this, whether it's a police detective sifting through weeks of city surveillance videos to discover the exact moment they want, or a medical researcher searching for a specific cellular reaction concealed in hundreds of hours of lab records.

The final purpose of VMR is to find out the exact start and finish time stamps of the segment which exactly match the query. Moment retrieval has to handle free-text, i.e., natural language queries that can describe any occurrence or action existing in the video. This task is challenging because queries such as "a person explains how to make coffee while standing in the kitchen" require the model to satisfy multiple conditions simultaneously, and not just detecting a kitchen or a coffee mug, but finding the precise window where the person, action, process and setting all overlap. This is the canonical problem of query complexity.

However, recent improvements in transformer-based designs, mostly inspired by the DETR (Detection Transformer) framework from object detection, have demonstrated great progress in solving this challenge. These models are designed to accept video data and text information as input and output many candidate temporal predictions in parallel with confidence score. The final answer is then selected as the highest confidence prediction model at test time. Training of these models is often based on bipartite matching: for each training sample, the algorithm determines the closest candidate prediction to the ground truth and applies regression and classification losses to the single candidate. This means that on every training step, only 1 out of K predictions receives useful feedback about its temporal quality. The remaining $K-1$ predictions receive no signal that why their temporal location is wrong or how to improve. At test time, however, all K predictions compete and the model must pick the best one. There is a mismatch: the model is never trained to compare predictions against each other, yet this is exactly what it must do during inference.

This problem matters in practice, for e.g., A security analyst searching for *a person in a red jacket entering the back door* across 72 hours of surveillance footage cannot

watch it all manually. A student looking for the exact minute where an instructor explains backpropagation in a three-hour lecture should not have to scrub through the whole recording. A basketball coach reviewing a full game to find *point guard drives left and kicks to the corner* needs this done in seconds, not hours. In each case, the user knows what they want but not where it is. Moment retrieval bridges this gap.

The task is different from video classification, which labels an entire video without locating events in time. It is also different from temporal action detection, which finds instances of pre-defined action categories like *running* or *jumping* but cannot handle free-form queries. Video captioning works in the opposite direction: it takes video and produces text, rather than taking text and producing timestamps. Video moment retrieval sits at the intersection: it takes open-ended language as input and returns precise temporal boundaries as output.

Several factors make it hard like queries can be compositional: *a woman in a blue dress explains a recipe while her dog plays in the background* requires all conditions to hold in the same time window. Temporal boundaries are ambiguous: different annotators disagree on when exactly *a person gives a speech begins*. The model must handle all of this using the user query text and video features at inference time.

1.2 Problem Statement

Given an untrimmed video and a natural language query, the problem is to predict the start and end timestamps of the video segment that best matches the query. The model must localize the moment precisely, meaning both boundaries should be as close to the human-annotated ground truth as possible, and it must do so using only the query text and video features at inference time.

Current DETR-based approaches produce K candidate proposals in a single forward pass and select the one with highest confidence at test time. These models are trained using bipartite matching: for each sample, one proposal is matched to the ground truth and receives a localization loss (L1 + GIoU), while the remaining $K-1$ proposals only learn to predict *no object*. This means most proposals never receive direct feedback about the quality of their temporal predictions and creates a train-test mismatch. At inference, the model must rank all K proposals and pick the best one. But during training, it is never taught to compare proposals against each other. Only one proposal per sample ever learns where the moment is. The other nine learn only that they are not the best, with no information about how far off they are or in which direction they should move. A proposal that misses by 2 seconds and one that misses by 30 seconds receive the same *no object* label. A related gap is that the training loss is purely geometric. It measures how far the predicted boundaries are from the ground truth but does not capture whether the predicted segment actually contains content relevant to the query. A prediction covering irrelevant content and a prediction covering the right content with slightly wrong boundaries receive the same type of loss signal. Consider a concrete case. Suppose the ground truth is 40–80s. Proposal A covers 60–100s (partially correct content but extends into irrelevant territory) and Proposal B covers 38–78s (nearly correct with slightly shifted boundaries). Both might receive similar L1 loss values since their boundary distances are comparable. But Proposal

B is clearly better because its content overlaps almost entirely with the ground truth, while Proposal A includes 20 seconds of unrelated material. The geometric loss cannot express this difference. It treats both errors equally because it only looks at numbers on a timeline, not at what those numbers contain.

This thesis addresses both gaps: how to provide a useful training signal to all proposals simultaneously, and how to incorporate semantic feedback into the training process.

1.3 Motivation

Our work is motivated by a recent development in language model training called Group Relative Policy Optimization (GRPO). In GRPO, instead of training a model on a single correct output, the model generates a group of candidate outputs and learns to distinguish better ones from worse ones within the group. Rewards are normalized across the group so that above-average candidates get reinforced and below-average ones get suppressed.

The connection to moment retrieval is natural. Where language models sample multiple text completions, moment retrieval models produce multiple temporal proposals. Both cases give a group of candidates that can be compared against each other. The key insight of this thesis is that we do not need to generate additional samples or add noise because the model already gives us multiple proposals by design.

Furthermore, we observe that the standard IoU-based evaluation metric conflates two distinct properties of a good video moment:

- **Semantic relevance:** Does the predicted segment actually contain content that matches the query meaning? For example, if the query asks about "a person explaining coffee preparation," the segment should contain speech about coffee preparation, not just visual shots of coffee.
- **Visual saliency:** Is the predicted segment important and engaging? Human annotators tend to mark certain clips as more highlight-worthy than others, regardless of query.

A good moment retrieval system should find segments that are both semantically relevant to the query and visually salient. By separating these two objectives into distinct reward signals, we can provide richer feedback to the model during training.

Figure 1.1 illustrates that baseline anchors on a narrow region containing a food-related keyword and misses the broader 58-second interaction. Our approach evaluates semantic entailment of the complete query, correctly identifying that the moment includes both "talking" and "picking food" components spanning the entire ground truth window.

1.4 Contributions

The main contributions of this thesis are as follows. We apply Group Relative Policy Optimization (GRPO) to video moment retrieval, treating the multiple proposals from

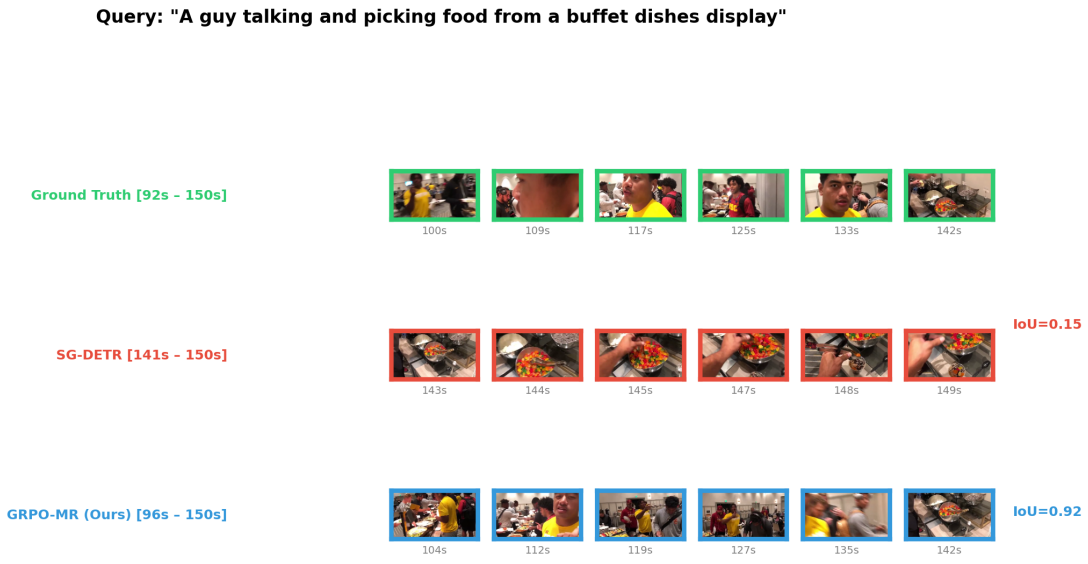

Figure 1.1: For the query "A guy talking and picking food from a buffet dishes display" (ground truth: 92–150s), SG-DETR retrieves only the last 9 seconds (IoU=0.15) by anchoring on a food keyword, while our method retrieves the full interaction from 96s onward (IoU=0.92).

36

the DETR decoder as the comparison group. Building on this framework, we make the following specific contributions:

1. **Cross-policy exploration:** We introduce cross-policy exploration where we keep a frozen copy of the original model and include its predictions in the group. This ensures the model keeps exploring even as training progresses.

1

2. **Dual reward design:** We design a reward function that combines semantic relevance (measured through NLI entailment between video transcripts and the query) with temporal saliency (measured through human annotation scores). This decomposition provides a more informative and robust training signal than either component alone.

3. **Strong empirical results:** We demonstrate improvements of +0.62% R1@0.5 on the validation split and +0.95% on the test split over SG-DETR on the QVHighlights benchmark. Results generalize to Charades-STA (+1.63% R1@0.5) and TACoS (+1.24% R1@0.5) datasets. The method requires only a single GPU for training and adds no inference overhead.

CHAPTER 2

LITERATURE REVIEW

Video moment retrieval requires a system to locate the temporal boundaries of a specific event within an untrimmed video, given a natural language query. Temporal grounding was initially approached using proposal-based pipelines, where candidate segments were generated at pre-defined temporal scales before query matching. Gao et al. [1] proposed TALL, that intensively samples fixed-length segments, collects segment-level features and scores each candidate with respect to the query by a learnt alignment function. Hendricks et al. [2] further extended this paradigm by proposing a Moment Context Network that uses local and global video representations for assessing candidates. However, a key disadvantage of all proposal-based approaches is that the temporal localization precision is strictly limited by the granularity of the preset window scales, which prevents from predicting the borders of the candidate edges.

Later works have transitioned to proposal-free formulations that explicitly regress temporal boundaries as continuous coordinates. Yuan et al. [3] ntroduced an attention-based regression framework that predicts normalized start and end timestamps without explicit candidate generation, which provides more flexibility for boundary placement. However, these algorithms are generally limited to predicting a single temporal segment for a given query, and cannot deal with multi-event or semantically ambiguous requests.

The current dominant paradigm draws from DETR [4], the Detection Transformer originally designed for object detection in images. DETR introduced the idea of using a fixed set of learnable query embeddings in a transformer decoder, where each embedding independently produces one prediction through cross-attention with encoded features. Carion et al. showed that this eliminates the need for hand-crafted components like anchor generation and non-maximum suppression. Lei et al. [5] adapted this design for temporal moment retrieval with Moment-DETR, which uses $K = 10$ decoder queries to output K temporal span predictions in parallel, each with a confidence score. Training uses bipartite matching via the Hungarian algorithm: for each training sample, the algorithm assigns the ground-truth moment to exactly one of the K predictions and applies localization loss (L1 + generalized IoU) only to that matched prediction. The remaining $K - 1$ predictions receive only a background classification label and no feedback about their predicted boundaries.

2.1 DETR-Based Moment Retrieval Models

Moment-DETR established a strong baseline, but it also exposed several architectural weaknesses that subsequent work aimed to fix. The most immediate issue was that the video encoder processes the video independently of the query. One issue with Moment-DETR is that the encoder processes the video without any knowledge of the query — it builds a generic video representation, and the query only comes in at the decoder stage. QD-DETR [6] and CG-DETR [7] tackled this from different directions. QD-DETR adds query-dependent attention layers inside the encoder itself, so the video features are already shaped by what the user is looking for. CG-DETR instead uses correlation scores between the query and individual video segments to re-weight attention, which helps the model focus on the most relevant parts in longer videos where many segments compete for attention. Both works showed that making the encoder aware of the query early on leads to better moment localization.

A second issue is how the decoder queries are initialized. In Moment-DETR, the K decoder queries start as random learnable embeddings that have no relation to the video content. EaTR [8] improved on this by using slot attention to initialize each decoder query near a natural event boundary in the video, rather than at an arbitrary position. Starting from content-aware positions means the proposals begin closer to meaningful moments, which makes optimization easier during training.

A third line of improvement recognized that moment retrieval does not exist in isolation. Highlight detection, the task of scoring each clip for importance, shares the same underlying requirement of understanding query-video relevance. UMT [9] was among the first to unify both tasks under a shared multi-modal transformer representation. TR-DETR [10] built on this idea by explicitly modeling the mutual dependencies between moment retrieval and highlight detection, allowing each task to benefit from the other's learned representations. UVCOM [11] further advanced this direction through compressed video representations and multi-scale temporal modeling for efficient joint prediction.

Most recently, SG-DETR [12] combined insights from both the attention calibration and the multi-task directions. It uses saliency scores, derived from the highlight detection branch, to directly guide the cross-attention mechanism in the decoder. This means the model's attention is modulated by visual importance, so decoder queries naturally focus on perceptually salient regions when predicting moment boundaries. SG-DETR currently achieves the strongest results on the QVHighlights benchmark among published methods.

Despite this steady architectural progress across all these models, the training procedure has remained unchanged: every model from Moment-DETR through SG-DETR uses bipartite matching to assign localization loss to a single prediction per sample, while the other $K-1$ predictions receive no boundary-level feedback, and no work in this line has questioned whether the training paradigm itself might be the bottleneck.

2.2 Highlight Detection and Saliency Annotations

Another related problem is highlight identification, where the model predicts a score for each clip in the video for its relevance and visual importance. QVHighlights[5] presented per-clip saliency annotations, obtained from multiple human annotators. Each annotator rates clips on a five-point scale from ”Very Good” to ”Very Bad,” and these individual ratings are averaged to produce continuous saliency scores.

These scores capture something that moment boundaries alone cannot convey. Even in a correctly localized moment, some clips are more visually interesting or informative than others, and outside the moment, some clips are still partially relevant. This gives a richer, more graded signal than a simple binary label of whether the clip is inside or outside the ground truth.

Existing methods for saliency prediction have posed the task as a separate learning task, where a parallel head is added to learn to predict these scores in parallel with the main moment retrieval objective. We use saliency in a different way. We do not predict it, but use saliency scores as a part of the reward function in the reinforcement learning. We treat saliency scores as a measure of how good a predicted moment is, instead of a target to be regressed.

2.3 Reinforcement Learning for Video Understanding

DETR based moment retrieval models produce K proposals per sample, but the bipartite matching loss only updates the single proposal closest to the ground truth. The other $K-1$ proposals are pushed toward a background class and never learn why their temporal predictions are wrong. This creates an asymmetry: at test time all proposals compete, but during training only one ever receives localization feedback. Reinforcement learning offers a natural solution to this problem because it can assign a scalar reward to every proposal independently, without requiring a one to one match with ground truth annotations. Each proposal is scored based on how well its content aligns with the query, and the model is updated to produce more proposals like the high scoring ones and fewer like the low scoring ones. This turns training into a ranking problem over the model’s own outputs, which is closer to what actually happens at inference. One early example in vision-language tasks is self-critical sequence training [13], developed for image captioning. The model samples a caption and also produces one by greedy decoding. The sampled output is reinforced if it receives a higher reward than the greedy one, and suppressed otherwise. This uses the model’s own greedy prediction as a baseline, which removes the need to train a separate value network.

In temporal grounding, He et al. [14] trained an agent that adjusts moment boundaries step by step through discrete actions like shift start left, shift end right, and so on, collecting IoU rewards after each move. Wu et al. [15] extended this with a tree-structured policy for coarse-to-fine refinement. These methods showed RL can improve moment retrieval, but they require discretizing time, designing action spaces, and running multiple passes at inference, making them slower and more complex.

Our work takes a different route. We do not convert the model into a sequential

agent. We start from a standard DETR-based architecture [5, 12] that already produces K proposals in a single forward pass, and add a reward-based training phase that scores all proposals at once and updates the model based on how they compare to each other. The algorithm we use is Group Relative Policy Optimization (GRPO), originally proposed by Shao et al. [16] for training language models on math reasoning.

The specific RL algorithm we use is Group Relative Policy Optimization (GRPO), introduced by Shao et al. [16] for training language models on mathematical reasoning. Standard RL algorithms like PPO [17] require a value function (critic) to estimate expected returns, which adds computational cost and introduces approximation errors. GRPO eliminates the critic entirely by computing advantages within a group of outputs from the same input:

$$A_k = \frac{r_k - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})} \tag{2.1}$$

Outputs scoring above the group average receive positive advantages; those below receive negative. The signal is self-calibrating because it compares outputs from the same model on the same input. GRPO has been applied to mathematical reasoning [16], general reasoning in DeepSeek-R1 [18], and agent training [19]. Recent works like Video-R1 [20] and Time-R1 [21] brought reinforcement learning to video understanding, but these operate on large language models that generate timestamps as text tokens. They cannot be applied to the parallel multi-proposal architecture of DETR-based models.

The effectiveness of any RL-based training depends on the reward signal. For moment retrieval, IoU with the ground truth is the most direct option, but it conflates semantic relevance with boundary precision: a prediction that covers the wrong content but has partial temporal overlap receives the same reward as one that covers the right content with slightly misaligned boundaries. To address this, we incorporate Natural Language Inference (NLI) as a semantic reward component. NLI models such as DeBERTa-v3 [22] determine whether a premise logically entails a hypothesis and capture deep linguistic properties including negation, paraphrasing, and compositional semantics. Yin et al. [23] showed NLI can be repurposed for zero-shot text classification by treating the class label as a hypothesis. For video grounding, NLI is particularly useful because simpler metrics like CLIP cosine similarity cannot distinguish "a person running" from "a person not running", both produce similar embeddings. NLI correctly identifies one as entailing the query and the other as contradicting it. Kim et al. [24] explored proxy semantic signals for zero-shot video grounding, but NLI entailment has not previously been used as a reward for training temporal grounding models.

We observe that DETR-based architectures already contain the group structure that GRPO requires: the K decoder queries naturally produce K diverse predictions in a single forward pass, forming a group without requiring multiple sampling passes. This connection between DETR's parallel predictions and GRPO's group-based optimization has not been exploited in prior work. We combine this with a dual reward that separates semantic relevance (via NLI) from visual saliency, providing a richer training signal than a single IoU score. By applying GRPO over all K predictions simultaneously, the model learns to compare predictions against each other.

CHAPTER 3

METHODOLOGY

This chapter presents DualRS, our proposed training framework for video moment retrieval. We first describe the problem setting and the model architecture that produces multiple temporal span predictions in parallel. We then explain how we train this model using reinforcement learning, detailing the dual reward function that provides learning signal to all predictions simultaneously, and how group normalization converts these rewards into parameter updates.

3.1 Problem Formulation

In video moment retrieval, we are given an untrimmed video V consisting of L clips and a natural language query q. The goal is to predict the start and end timestamps (s^*, e^*) of the moment in the video that corresponds to the query. Both timestamps are normalized to $[0,1]$ where 0 represents the beginning of the video and 1 represents the end.

Our model f_θ takes the video and query as input and produces K temporal span predictions, which we call *candidates*. Each candidate represents one possible answer to the query:

$$\{(\hat{s}_k, \hat{e}_k, c_k)\}_{k=1}^{K} = f_\theta(V, q) \tag{3.1}$$

where $\hat{s}_k, \hat{e}_k \in [0,1]$ are the predicted start and end times for candidate k, and c_k is the confidence score indicating how likely candidate k is to be the correct answer. At test time, the candidate with the highest confidence is selected as the final prediction.

3.2 Model Architecture

We employ a DETR-style [4] transformer encoder-decoder architecture for moment retrieval, following the design in recent work [5, 12]. As shown in the Forward Pass block of Figure 3.1, pre-extracted video and text features are input to the model, which produces K temporal span predictions in a single pass. The architecture has two main components: a transformer encoder that fuses the video and text features through cross-modal and self-attention, and a transformer decoder that outputs the K candidate spans.

Algorithm 1 DualRS Training

[h]

Require: Model f_θ, stale model f_{stale}, dataset $\mathscr{D}$, sync interval N, reward weight α

Initialize $f_{\text{stale}} \leftarrow f_\theta$
for each training step t **do**
  Sample (V, q, s^*, e^*) from $\mathscr{D}$
  $\{(\hat{s}_k, \hat{e}_k, c_k)\}_{k=1}^{K} \leftarrow f_\theta(V, q)$ ▷ current model proposals
  $\{(\hat{s}_j^{\text{ref}}, \hat{e}_j^{\text{ref}})\}_{j=1}^{K} \leftarrow f_{\text{stale}}(V, q)$ ▷ stale model proposals
  for each candidate k in all $2K$ proposals **do**
    Compute r_{sem}^k using pre-cached NLI entailment scores
    Compute r_{sal}^k using saliency annotations
    $r_k \leftarrow \alpha \cdot r_{\text{sem}}^k + (1-\alpha) \cdot r_{\text{sal}}^k$
  end for
  Compute advantages: $A_k = \text{clip}\left(\frac{r_k - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}, -2, +2\right)$
  $\mathscr{L}_{\text{PG}} = -\sum_{k=1}^{K} A_k \cdot \log p_k$
  $\mathscr{L}_{\text{KL}} = \text{KL}[p \| p^{\text{stale}}]$
  $\mathscr{L}_{\text{span}} = \frac{1}{|\mathscr{K}^+|} \sum_{k:A_k>0} |(\hat{s}_k, \hat{e}_k) - (s^*, e^*)|$
  $\mathscr{L} = \mathscr{L}_{\text{PG}} + 0.1 \cdot \mathscr{L}_{\text{KL}} + 0.1 \cdot \mathscr{L}_{\text{span}}$
  Update θ using $\nabla_\theta \mathscr{L}$
  if $t \mod N = 0$ **then**
    $f_{\text{stale}} \leftarrow f_\theta$ ▷ sync stale model
  end if
end for

In Figure 3.1, the dark blocks (top row) represent the K span predictions from the current trainable model (π_θ), and the light blocks (bottom row) represent the K span predictions from the stale model ($\pi_{\theta_{\text{old}}}$), which is a periodically synchronized frozen copy used to maintain exploration diversity. By including predictions from a slightly older version of the model, we ensure the group always contains diverse candidates pointing to different parts of the video, which is necessary for the reward-based comparison in later stages to be meaningful.

To make the training procedure concrete, consider a single step where the video is 150 seconds long and the query is *a man explains how to fix a bicycle tire.* The current model produces 10 proposals and the stale model produces 10 more, giving 20 candidates total. One good proposal (from the current model) covers clips 38–52 where the ASR says *first you need to remove the tire from the wheel, then check for punctures*, receiving high NLI entailment (average 0.72 inside, 0.18 outside) and a saliency average of 3.8, for a combined reward of 1.84. A bad proposal covers clips 60–72 where the man is assembling furniture (entailment 0.12 inside, 0.25 outside, saliency 2.1), receiving a combined reward of only 0.76. After computing rewards for all 20 proposals, the group mean is 1.1 and standard deviation is 0.5, so the good proposal gets advantage +1.49 (reinforced) and the bad proposal gets advantage −0.68 (suppressed). The policy gradient then increases confidence for proposals covering relevant bicycle repair content and decreases it for proposals covering unrelated segments.

3.2.1 Feature Encoding

33

Each video is split into 2-second clips and features are extracted using InternVideo2 [25], a video-language foundation model. This produces a sequence of L clip-level feature vectors ($L = 75$ for a 150-second video). The text query is encoded using the same model's text encoder, producing a dense vector representation.

The transformer encoder then refines these features through two types of attention: cross-modal attention where video tokens attend to text tokens to become query-aware, and temporal self-attention where video tokens attend to each other to capture global context. The output is a sequence of L query-conditioned video representations that encode both the visual content and its relevance to the query.

3.2.2 Candidate Generation via Decoder Queries

The decoder produces the K candidates. It contains $K = 10$ **learnable query embeddings**, which are randomly initialized vectors that are trained along with the rest of the model. Each query embedding serves as a slot that will produce one candidate.

Through cross-attention layers, each decoder query attends to the encoded video representations and independently predicts a temporal span and a confidence score. The temporal span is parameterized as a center position $c_k \in [0,1]$ and width $w_k \in [0,1]$, converted to start and end timestamps as $s_k = c_k - w_k/2$ and $e_k = c_k + w_k/2$. The foreground confidence score $\sigma_k \in [0,1]$ indicates how likely this candidate is to contain the queried moment.

This produces $K = 10$ candidates $\{(s_k, e_k, \sigma_k)\}_{k=1}^{10}$ from a single forward pass. Each candidate is a temporal interval in the video that the model considers as a possible location of the queried moment. These K candidates form a natural group: they are diverse predictions for the same input, generated in parallel without any sequential dependence. This group structure is what enables us to apply group-based reinforcement learning. In Figure 3.1, this stage corresponds to the two rows of span predictions in the Forward Pass block: the dark blocks (top row) represent the K candidates from the current model, and the light blocks (bottom row) represent the K candidates from the stale model, both generated through this decoder mechanism.

3.3 Training with Dual Reward Group Optimization

We train the model using Group Relative Policy Optimization (GRPO) with a dual reward function. The idea is to score all K candidates using a reward that measures how good each prediction is, and then update the model so that better candidates become more confident and worse candidates become less confident. Because the reward is computed for every candidate, the model receives a learning signal from all K predictions on every training step, not just the single matched one.

Figure 3.1 illustrates the complete training pipeline. On the left, the input consists of a video paired with a natural language query. Both are fed into the encoder-decoder architecture, which produces $K = 10$ temporal span predictions. We run this forward pass twice: once with the current model (dark blocks, top row) and once with the stale

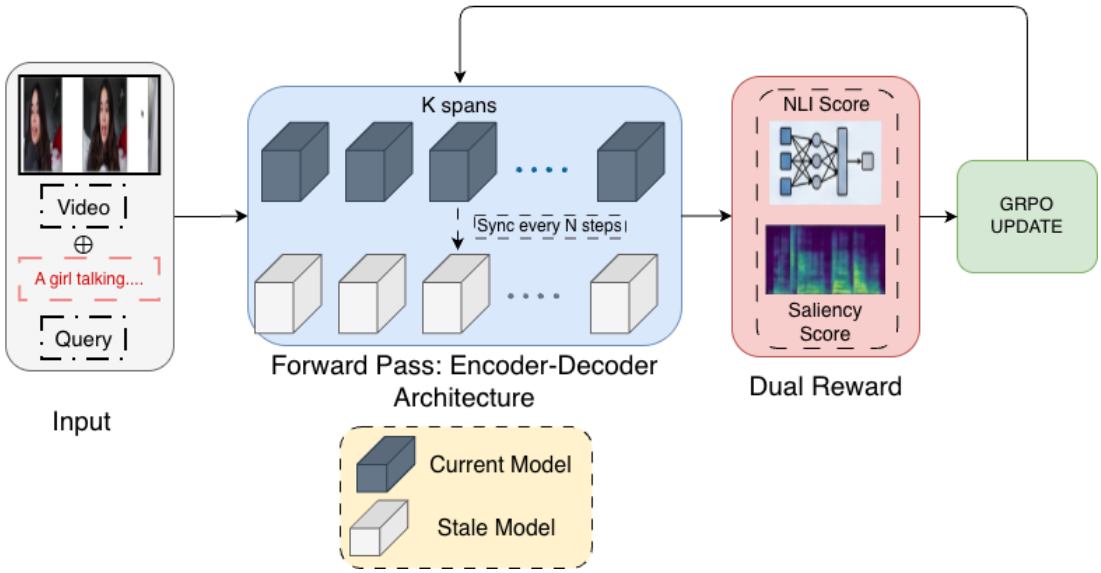

Figure 3.1: Overview of the DualRS training pipeline. Given a video and query, both the current model (π_θ) and a periodically synchronized stale model ($\pi_{\theta_{\text{old}}}$) produce K temporal span predictions through an encoder-decoder architecture. All $2K$ spans are scored by the dual reward, which combines an NLI-based semantic relevance score with a saliency score from human annotations. The GRPO update normalizes these rewards across the group and computes a policy gradient that updates only the current model.

model (light blocks, bottom row), giving us $2K$ candidates in total. The stale model is a frozen copy that is synchronized with the current model every N training steps, ensuring that the two sets of proposals remain diverse enough for meaningful comparison. All $2K$ candidates are then passed to the dual reward module, which assigns each candidate two scores: an NLI score measuring how well the speech content within the predicted span supports the query, and a saliency score measuring how visually important the covered clips are according to human ratings. These two scores are combined into a single reward per candidate. Finally, the GRPO update block normalizes all rewards within the group to compute advantages, and produces a policy gradient that flows back to update the current model parameters. The following subsections describe each component in detail.

3.3.1 Candidate Scoring with Dual Rewards

For each training sample consisting of a video V and query q, we run a forward pass to collect all K candidates. We then score each candidate using a reward that combines two complementary signals.

Semantic Relevance Reward

The first signal measures whether a candidate covers video content that matches the query. Each video in the dataset has associated ASR (Automatic Speech Recognition) subtitle text aligned at the clip level. We use a Natural Language Inference (NLI) model to determine how well the text content of each clip logically supports the query.

An NLI model takes two text inputs, a premise and a hypothesis, and classifies their relationship as entailment (the premise supports the hypothesis), contradiction (the premise opposes the hypothesis), or neutral (no clear relationship). We treat the subtitle text of each clip as the premise and the query as the hypothesis. The entailment probability output by the model serves as a score $e_i \in [0,1]$ for clip i, computed as $e_i = P(\text{entail} \,|\, \text{subtitle}(c_i), q)$ using DeBERTa-v3-large [22] fine-tuned on MultiNLI. A high e_i means the speech content in clip i logically supports the query (e.g., clip subtitle says "a man opens a box" and query is "person opening a gift"). A low e_i means the clip content is unrelated or contradicts the query.

For a candidate spanning clips $[\hat{s}, \hat{e}]$, let $\mathscr{I} = \{i : \hat{s} \leq i \leq \hat{e}\}$ be the set of clips inside the span and $\mathscr{O} = \{i : i \notin \mathscr{I}\}$ be the clips outside. The semantic reward is:

$$r_{\text{sem}} = \frac{1}{|\mathscr{I}|} \sum_{i \in \mathscr{I}} e_i - \frac{1}{|\mathscr{O}|} \sum_{i \in \mathscr{O}} e_i \tag{3.2}$$

The first term measures the average entailment of clips covered by the candidate.The second term acts as a contrastive penalty that subtracts the average entailment of clips outside the candidate. A candidate receives a high semantic reward when it covers clips whose text matches the query (high inside score) while the text outside does not match (low outside score). This structure forces tight boundaries. If the candidate is too broad, irrelevant clips are added and dilute the inner average and if the candidate is too narrow, relevant clips are left in the outside group, increasing the penalty.

The entailment scores are pre-computed once before training and cached. During training, we look up the relevant scores for each candidate, adding zero computational overhead.

Saliency Reward

The second signal captures visual highlight quality independently of query semantics. The QVHighlights dataset provides human annotations where multiple annotators rate each clip's visual importance on a scale of 1 to 5 (from "Very Bad" to "Very Good"). These ratings are averaged across annotators to produce a continuous saliency score $h_i \in [1,5]$ for each clip i.

For a candidate spanning clips $[\hat{s}, \hat{e}]$, with $\mathscr{I} = \{i : \hat{s} \leq i \leq \hat{e}\}$ as the set of covered clips, the saliency reward is:

$$r_{\text{sal}} = \frac{1}{|\mathscr{I}|} \sum_{i \in \mathscr{I}} h_i \tag{3.3}$$

This rewards candidates whose temporal spans align with clips that humans found visually important. The saliency signal complements the semantic reward because it

captures visual engagement that subtitle text alone cannot express. For example, a dramatic camera angle or an emotionally impactful scene may have high saliency even if the accompanying speech is generic.

The two signals are combined into a single reward for each candidate:

$$r = \alpha \cdot r_{\text{sem}} + (1 - \alpha) \cdot r_{\text{sal}} \tag{3.4}$$

We set $\alpha = 0.6$, giving more weight to semantic relevance while still benefiting from the saliency signal. This ratio was selected based on validation performance (see ablations in Chapter 4).

3.3.2 Cross-Policy Exploration and Policy Update

As training progresses, the K candidates tend to cluster in similar temporal regions. If all candidates converge to the same location, comparing them becomes meaningless because they all receive similar rewards. Inspired by cross-policy sampling [**?**], we address this by maintaining a stale copy of the model that is updated periodically (every N training steps) rather than on every step. This stale model generates an additional set of K candidates on the same input:

$$\{(\hat{s}_j^{\text{ref}}, \hat{e}_j^{\text{ref}})\}_{j=1}^{K} = f_{\text{stale}}(V, q) \tag{3.5}$$

Because the stale model lags behind the current policy, its candidates explore parts of the video that the current model may have already moved away from. Combining both sets gives us $2K$ candidates with sufficient diversity for meaningful comparison. The stale model is synchronized with the current model's parameters every $N = 5$ steps.

Once rewards are computed for all $2K$ candidates, we normalize them relative to the group to obtain advantages:

$$A_k = \frac{r_k - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})} \tag{3.6}$$

Candidates scoring above the group average receive positive advantages and those below receive negative advantages. We clip advantages to $[-2, +2]$ to prevent extreme updates. The advantage represents how a candidate compares to the other candidates for the same input, not its absolute quality.

The policy gradient loss then updates the model's confidence distribution across the K candidates from the current model:

$$\mathscr{L}_{\text{PG}} = -\sum_{k=1}^{K} A_k \cdot \log p_k \tag{3.7}$$

where $p_k = \text{softmax}(c_1, \ldots, c_K)_k$ is the normalized confidence of candidate k. The gradient flows through all K candidates simultaneously, and those with positive advantages receive gradients that increase their confidence, while those with negative advantages receive gradients that decrease it.

We add a KL divergence penalty between the current model's confidence distribution and the stale model's distribution to prevent the policy from changing too rapidly between synchronization intervals:

$$\mathscr{L}_{\mathrm{KL}} = \mathrm{KL}[p \,\|\, p^{\mathrm{stale}}] \tag{3.8}$$

For candidates with positive advantages, we also apply a boundary refinement loss that encourages their predicted timestamps to move toward the ground truth:

$$\mathscr{L}_{\mathrm{span}} = \frac{1}{|\mathscr{K}^{+}|} \sum_{k:A_k>0} |(\hat{s}_k, \hat{e}_k) - (s^*, e^*)| \tag{3.9}$$

The total loss combines all three components:

$$\mathscr{L} = \mathscr{L}_{\mathrm{PG}} + 0.1 \cdot \mathscr{L}_{\mathrm{KL}} + 0.1 \cdot \mathscr{L}_{\mathrm{span}} \tag{3.10}$$

Algorithm 1 summarizes the complete DualRS training procedure. Each training step begins by sampling a video-query pair and running two forward passes: one through the current model to produce K proposals and one through the stale model to produce K more. All $2K$ proposals are then scored using the combined semantic and saliency reward. These rewards are normalized into advantages, which determine how each proposal's confidence should change. The policy gradient loss reinforces high-advantage proposals and suppresses low-advantage ones, while the KL penalty prevents the confidence distribution from shifting too fast and the boundary refinement loss nudges good proposals toward the ground truth. The three components are combined into a single loss that updates the model. Every N steps, the stale model copies the current model's parameters to refresh its predictions while maintaining the temporal lag needed for exploration diversity.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the experimental evaluation of the proposed DualRS framework. We describe the datasets and metrics used, provide full implementation and training details, compare against state-of-the-art methods on both validation and test splits, ablate each component of the framework, analyze what changes inside the model after training, test generalization to other datasets, and provide qualitative examples.

4.1 Datasets and Evaluation Metrics

4.1.1 QVHighlights

Our primary evaluation benchmark is QVHighlights [5], the standard dataset for joint moment retrieval and highlight detection. It contains 10,148 YouTube videos with natural language queries, split into 7,218 training, 1,550 validation, and 1,542 test samples. Each query is annotated with relevant time windows and per-clip saliency scores collected from multiple human annotators. The videos cover diverse content including vlogs, news, documentaries, and how-to videos. ASR subtitles are available for most videos, which we use for computing our semantic reward.

4.1.2 Generalization Datasets

We additionally evaluate on two other benchmarks to test generalization:

Charades-STA contains indoor activity videos with temporal sentence annotations. The videos are shorter on average and the queries describe household activities.

TACoS consists of cooking videos with fine-grained temporal descriptions. The moments tend to be short relative to the full video length, making precise boundary prediction important.

4.1.3 Evaluation Metrics

The primary evaluation measure for moment retrieval is Intersection over Union (IoU), which quantifies the temporal overlap between a predicted moment $[\hat{s}, \hat{e}]$ and the ground

truth $[s^*, e^*]$. It is computed as the ratio of the intersection length to the union length:

$$\text{IoU} = \frac{\min(\hat{e}, e^*) - \max(\hat{s}, s^*)}{\max(\hat{e}, e^*) - \min(\hat{s}, s^*)} \tag{4.1}$$

An IoU of 1.0 indicates perfect alignment between the prediction and ground truth, while an IoU of 0 means no overlap at all.

Following standard practice in the literature, we report Recall@1 at two IoU thresholds: R1@0.5 and R1@0.7. R1@0.5 is the percentage of test samples for which the model's top prediction achieves an IoU of at least 0.5 with the ground truth. R1@0.7 applies the same logic at a stricter threshold. The difference between the two captures boundary precision: a model may correctly identify the general region of the target moment (passing the 0.5 threshold) but fail to draw tight boundaries (failing the 0.7 threshold). Together, these two numbers describe both localization accuracy and boundary quality.

For highlight detection, we report mean Average Precision (mAP). Each clip in the video receives a predicted saliency score from the model, and clips whose ground truth saliency exceeds a threshold are treated as positives. Average precision is computed across multiple thresholds and averaged to produce mAP. A high mAP indicates that the model consistently ranks visually important clips above less important ones.

4.2 Experimental Setup

All experiments are conducted on a single NVIDIA A100 (80 GB) GPU using PyTorch. The video features are extracted using InternVideo2 [25] at 2-second clip resolution, producing 75 clips per video. The text query is encoded using the same model's text encoder.

The encoder-decoder model uses $K = 10$ learnable decoder queries, producing 10 temporal span predictions per forward pass. The model is trained using the DualRS framework with Group Relative Policy Optimization and our dual reward function.

We use the AdamW optimizer [26] with a peak learning rate of 3×10^{-6} and cosine decay with linear warmup over the first epoch. The prediction heads receive the full learning rate while the feature encoder uses a $10\times$ smaller rate. Gradient norms are clipped at 1.0. Training runs for up to 90 epochs with early stopping based on validation R1@0.5 (patience of 5 epochs). The only extra cost per step compared to standard training is one additional forward pass through the stale model for cross-policy exploration. The full DualRS training takes approximately 4 hours on the QVHighlights training set.

For the NLI component, we use DeBERTa-v3-large trained on MultiNLI. The entailment scores are computed once for all clip-query pairs in the training set and cached before training begins.

4.3 Comparison with State-of-the-Art

Table 4.1 presents our main results on the QVHighlights validation split alongside existing methods. All methods in the upper section use CLIP+SlowFast features, while our method and SG-DETR use InternVideo2-1b features for a fair comparison.

Table 4.1: Comparison with existing methods on QVHighlights val split. All numbers are percentages.

Method	R1@0.5	R1@0.7	mAP
Moment-DETR [5]	53.94	34.84	32.20
UMT [9]	60.26	44.26	38.59
UniVTG [27]	68.39	–	45.99
QD-DETR [6]	62.68	46.66	41.22
CG-DETR [7]	65.40	52.10	44.90
TR-DETR [10]	64.66	48.96	39.91
BAM-DETR [28]	65.10	51.61	65.41
SG-DETR [12]	72.80	**59.50**	55.64
DualRS (ours)	**73.42**	58.12	**43.96**

DualRS achieves the best R1@0.5 (73.42%), surpassing SG-DETR by +0.62%. By training on all 10 proposals and their rewards, the model learns to place boundaries more precisely rather than just getting the rough location right. SG-DETR retains an advantage on R1@0.7, which we attribute to their additional training on the InterVid-MR dataset (150k extra samples) that we do not use.

4.4 Results on Test Split

Table 4.2 presents results on the held-out test split, confirming that the improvements are not due to overfitting on the validation set.

Table 4.2: Results on QVHighlights test split.

Method	R1@0.5	R1@0.7	mAP
Moment-DETR [5]	59.78	40.33	36.14
UMT [9]	60.83	43.26	38.08
UniVTG [27]	65.43	50.06	43.63
QD-DETR [6]	63.06	45.10	40.19
CG-DETR [7]	65.40	48.40	42.90
BAM-DETR [28]	62.71	48.64	64.57
TR-DETR [10]	65.05	47.67	43.10
SG-DETR [12]	74.20	**60.40**	**58.80**
DualRS (ours)	**75.15**	59.89	58.10

The test split results are consistent with validation performance. DualRS achieves 75.15% R1@0.5, outperforming SG-DETR by +0.95%. SG-DETR retains a slight ad-

vantage on R1@0.7 (60.40 vs 59.89) and mAP (58.80 vs 58.10), which we attribute to their additional training data. The consistent improvements across both splits confirm that our approach generalizes well to unseen data.

4.5 Ablation Studies

We conduct ablation experiments on the QVHighlights validation split to understand the contribution of each component in our framework.

4.5.1 Reward Component Ablation

Table 4.3 presents results when using different reward configurations.

Table 4.3: Ablation of reward components on QVHighlights.

Reward Configuration	**R1@0.5**	**R1@0.7**	**mIoU**
Semantic (NLI) only	72.18	56.87	60.45
Saliency only	71.54	55.92	57.92
NLI + Saliency	73.08	**58.25**	62.18
NLI + Saliency + Contrastive	**73.42**	58.12	**62.87**

Using the semantic (NLI) reward alone yields 72.18% R1@0.5 and 60.45% mIoU. This signal provides a strong learning signal by directly measuring whether the retrieved content semantically supports the query. The saliency reward alone achieves 71.54% R1@0.5 and 57.92% mIoU, which is lower than NLI, showing that although perceptual importance is informative, it lacks the query-specificity needed for accurate retrieval.

Combining NLI and saliency rewards improves to 73.08% R1@0.5 (+0.90%) and 62.18% mIoU (+1.73% over NLI alone). The gains confirm that these two signals are complementary: semantic relevance ensures the retrieved moment matches the query meaning, while saliency encourages selection of perceptually important regions.

Adding the contrastive penalty achieves our best result of 73.42% R1@0.5 and 62.87% mIoU. The contrastive component explicitly penalizes proposals whose temporal coverage extends into irrelevant regions by comparing entailment scores inside versus outside the predicted span. This sharpens temporal boundaries.

4.5.2 Cross-Policy Exploration Ablation

Table 4.4 examines the effect of cross-policy exploration, where a periodically-updated stale model provides additional candidates to maintain diversity in the comparison group.

Without cross-policy exploration, the model uses only the $K = 10$ candidates from the current policy for group comparison. This leads to 72.41% R1@0.5 and 60.82% mIoU. With cross-policy enabled, where the stale model contributes an additional K candidates (making the group size $2K = 20$), performance improves to 73.42% R1@0.5 (+1.01%) and 62.87% mIoU (+2.05%).

Table 4.4: Effect of cross-policy exploration on QVHighlights val split.

Configuration	R1@0.5	R1@0.7	mIoU
Without cross-policy (K candidates only)	72.41	**56.93**	60.82
With cross-policy (2K candidates)	**73.42**	56.12	**62.87**

This improvement occurs because without the stale model's candidates, the group tends to collapse as training progresses. When all candidates cluster in similar temporal regions, the advantage normalization becomes less informative since all candidates receive similar rewards. The stale model, being a few steps behind the current policy, continues to explore different parts of the video timeline, providing a meaningful baseline for comparison and preventing premature convergence.

4.6 Generalization to Other Datasets

Table 4.5 shows that the improvements transfer to Charades-STA and TACoS, confirming that DualRS captures general training principles rather than dataset-specific patterns.

Table 4.5: Results on Charades-STA and TACoS.

	Charades-STA		TACoS	
Method	R1@0.5	R1@0.7	R1@0.3	R1@0.5
Moment-DETR [5]	53.63	31.37	37.97	24.67
UniVTG [27]	58.01	35.65	35.09	17.35
QD-DETR [6]	57.31	32.55	36.80	21.10
CG-DETR [7]	58.44	36.34	39.61	22.20
BAM-DETR [28]	60.04	39.40	41.48	26.80
SG-DETR [12]	70.20	**49.50**	44.70	29.90
DualRS (ours)	**71.83**	48.9	**49.68**	**31.14**

On Charades-STA, DualRS achieves 71.83% R1@0.5, surpassing SG-DETR by +1.63%. On TACoS, DualRS achieves 49.68% R1@0.3 and 31.14% R1@0.5, outperforming all baselines. Since these datasets have different characteristics (Charades has short indoor activity moments, TACoS has cooking videos with fine-grained temporal descriptions), the consistent improvements confirm that the group-relative training paradigm and dual reward design generalize across domains.

4.7 Qualitative Analysis

Figure 4.1 presents qualitative comparisons on three examples from the QVHighlights dataset where the baseline achieves non-zero IoU but DualRS produces substantially better predictions. We verify the predictions using ASR transcripts to confirm what is actually being said at each timestamp.

Figure 4.1: Qualitative comparison on three QVHighlights examples. For each video, we show sampled frames with ground truth (green), baseline prediction (red), and DualRS prediction (blue). The baseline achieves partial overlap in all cases but DualRS produces precise boundaries by understanding semantic transitions rather than relying on keyword matching.

In Figure 4.1(a), the query is "A guy talking and picking food from a buffet dishes display" with ground truth spanning 92 to 150s. The baseline retrieves only the final 9 seconds (141 to 150s, IoU=0.15), anchoring on the moment where the word "chocolate" appears in the ASR. However, the full buffet interaction where the guy is actively talking, picking items, and interacting with other people begins much earlier at 92s. The baseline collapses to a narrow region with strong food related features and misses the broader 58 second interaction. DualRS retrieves from 96s onward (IoU=0.92) because the NLI reward requires the predicted span to entail the full query, including both the "talking" and "picking food" components, rather than matching a single keyword.

In Figure 4.1(b), the query is "Woman is talking to a man while in bed" with ground truth at 0 to 20s. The ASR confirms that the first 20 seconds contain a casual in bed conversation between two people. At 18s there is a music transition, and from 26s onward the woman is making phone calls about hotel bookings. She is no longer in bed

and the man is no longer present. The baseline extends to 54s (IoU=0.37) because a woman is still talking on screen, satisfying surface level visual features. DualRS ends precisely at 20.6s (IoU=0.97) since the contrastive NLI penalty finds the scene change at 20s does not involve "talking to a man while in bed" as it is now solo phone calls discussing travel arrangements.

In Figure 4.1(c), the query is "A girl is talking about studying for the TEAS test that she has to take to get into college". The ground truth is from 78 to 108s. In this window, looking at the ASR transcript, she is clearly articulating what the TEAS test is ("it is basically an exam test for pre-health students and nursing students who want to get into a nursing program"), talking about which universities require it, and expressing anxiety about the deadline. The topic continues on to her daily routine after 108s: coffee, getting up late, wearing cosmetics. The word 'studying' does still come up numerous times after 108s (at 115s, 131s, 143s), but she is no longer talking about the test itself. She mentions studies in passing when describing her day. The baseline learns this repeating term and propagates its prediction until 149s (IoU=0.42). DualRS ends at 111s (IoU=0.91) as the NLI model can distinguish between content where she is genuinely explaining the TEAS test and stuff where the phrase "studying" occurs to be in an unrelated context.

All three have the same design. Keywords from the query keep surfacing in the video beyond the ground truth boundary ('chocolate" in the food example, a woman still talking in the bed example, 'studying" here), and the baseline hooks onto them and overextends its forecast. The NLI reward in DualRS measures whether the span content logically entails the complete query rather than a mere word match. This is what allows it to determine when the topic has changed and halt the prediction at the proper place.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This thesis presented DualRS, a group relative reinforcement learning framework for video moment retrieval. The key idea is to treat all K decoder proposals as a comparison group and train the model using policy optimization with a dual reward function that combines semantic relevance and visual saliency. By applying Group Relative Policy Optimization to the DETR-based moment retrieval architecture, the model receives a learning signal from all K candidates on every training step. Candidates that score well on the dual reward are reinforced while candidates that score poorly are discouraged. This comparative learning teaches the model to rank proposals effectively, which is exactly what it must do at inference time when selecting the highest confidence prediction.

The dual reward separates two aspects of prediction quality. The semantic component, computed through NLI entailment on ASR transcripts, measures whether the retrieved content matches the query meaning. The saliency component, derived from human annotations, measures whether the content is perceptually important. The contrastive penalty sharpens boundaries by penalizing proposals that extend into irrelevant regions. Together, these provide richer feedback than IoU-based loss alone. Cross-policy exploration using a periodically synchronized reference model maintains diversity in the comparison group, preventing the proposals from collapsing to similar temporal regions as training progresses.

The ablation studies confirmed that each component contributes to the final result. Removing the semantic reward drops R1@0.5 by *1.24* points. Removing saliency drops it by *1.88* points. Removing the contrastive penalty costs *0.34* points. Removing cross-policy exploration reduces performance by *1.01* points and mIoU by *2.05* points. No single component accounts for the full improvement; the gains come from their combination. DualRS achieves *73.42%* R1@0.5 on the validation split and *75.15%* on the test split of QV Highlights, surpassing SG-DETR by +*0.62%* and +*0.95%* respectively. Results on Charades-STA (+*1.63%* R1@0.5) and TACoS (+*1.24%* R1@0.5) confirm that the improvements generalize across datasets with different video types, query styles, and moment durations. The entire training runs on a single GPU in approximately 4 hours and adds no computational overhead at inference time, since only

the standard forward pass is needed to produce the final prediction.

5.2 Future Scope

Several directions can extend this work further. The current framework uses visual features and text in the form of ASR transcripts, but videos contain rich audio information that remains unexplored for moment retrieval. Incorporating audio features such as speech tone, background music transitions, and ambient sound changes could provide additional signal for detecting moment boundaries, since scene transitions are often accompanied by changes in audio that precede or coincide with visual changes.

The saliency reward currently requires human annotations that are expensive to collect and only available in QVHighlights. Training a learned saliency predictor on existing annotations and using it as a reward model would enable DualRS training on any dataset with temporal annotations, removing the dependence on per clip human labels. Similarly, the fixed reward mixing ratio could be replaced by an adaptive scheme that adjusts the balance between semantic and saliency rewards based on sample characteristics, giving more weight to semantics for speech driven queries and more weight to saliency for visually driven ones.

5.3 Social Impact

Video moment retrieval can improve accessibility by allowing people with hearing impairments to navigate video content through text queries, and by helping users with limited time jump directly to relevant parts of long recordings. However, the same technology applied to surveillance footage could enable tracking individuals across large video archives without consent, so any deployment in such contexts should include strict access controls and legal oversight. The NLI component processes speech transcripts that may contain sensitive personal information, and systems using this approach should handle such data with appropriate privacy protections.

BIBLIOGRAPHY

[1] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5277–5285.

[2] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5804–5813.

[3] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: temporal sentence localization in video with attention based location regression," ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33019159

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[5] J. Lei, T. L. Berg, and M. Bansal, "Qvhighlights: detecting moments and highlights in videos via natural language queries," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[6] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query - dependent video representation for moment retrieval and highlight detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23 023–23 033.

[7] W. Moon, S. Hyun, S. Lee, and J.-P. Heo, "Correlation-guided calibration of query dependency for video temporal grounding," *Pattern Recognition*, vol. 174, p. 112984, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320325016474

[8] J. Jang, J. Park, J. Kim, H. Kwon, and K. Sohn, "Knowing where to focus: Event-aware transformer for video grounding," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 13 800–13 810.

[9] Y. Liu, S. Li, Y. Wu, C. W. Chen, Y. Shan, and X. Qie, "Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection," 2022. [Online]. Available: https://arxiv.org/abs/2203.12745

[10] H. Sun, M. Zhou, W. Chen, and W. Xie, "Tr-detr: task-reciprocal transformer for joint moment retrieval and highlight detection," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. [Online]. Available: https://doi.org/10.1609/aaai.v38i5.28304

[11] Y. Xiao, Z. Luo, Y. Liu, Y. Ma, H. Bian, Y. Ji, Y. Yang, and X. Li, "Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection," 2023. [Online]. Available: https://arxiv.org/abs/2311.16464

[12] A. Gordeev, V. Dokholyan, I. Tolstykh, and M. Kuprashevich, "Saliency-guided detr for moment retrieval and highlight detection," pp. 907–916, March 2026.

[13] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1179–1195.

[14] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: reinforcement learning for temporally grounding natural language descriptions in videos," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33018393

[15] J. Wu, G. Li, S. Liu, and L. Lin, "Tree-structured policy based progressive reinforcement learning for temporally language grounding in video," 2020. [Online]. Available: https://arxiv.org/abs/2001.06680

[16] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *CoRR*, vol. abs/2402.03300, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.03300

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[18] D. Guo, D. Yang, and Zhang, "Deepseek-r1 incentivizes reasoning in llms through reinforcement learning," *Nature*, vol. 645, no. 8081, p. 633–638, 2025. [Online]. Available: http://dx.doi.org/10.1038/s41586-025-09422-z

[19] H. Zhang, X. Liu, B. Lv, X. Sun, B. Jing, I. L. Iong, Z. Hou, Z. Qi, H. Lai, Y. Xu, R. Lu, H. Wang, J. Tang, and Y. Dong, "Agentrl: Scaling agentic reinforcement learning with a multi-turn, multi-task framework," *CoRR*, vol. abs/2510.04206, October 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2510.04206

[20] K. Feng, K. Gong, B. Li, Z. Guo, Y. Wang, T. Peng, J. Wu, X. Zhang, B. Wang, and X. Yue, "Video-r1: Reinforcing video reasoning in MLLMs," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. [Online]. Available: https://openreview.net/forum?id=a2JTVVvcEl

[21] Y. Wang, Z. Wang, B. Xu, Y. Du, K. Lin, Z. Xiao, Z. Yue, J. Ju, L. Zhang, D. Yang, X. Fang, Z. He, Z. Luo, W. Wang, J. Lin, J. Luan, and Q. Jin, "Time-r1: Post-training large vision language model for temporal video grounding," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. [Online]. Available: https://openreview.net/forum?id=gJ05Gm5VxQ

[22] P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," 2023. [Online]. Available: https://arxiv.org/abs/2111.09543

[23] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3914–3923. [Online]. Available: https://aclanthology.org/D19-1404/

[24] D. Kim, J. Park, J. Lee, S. Park, and K. Sohn, "Language-free training for zero-shot video grounding," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 2538–2547.

[25] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi, T. Jiang, S. Li, J. Xu, H. Zhang, Y. Huang, Y. Qiao, Y. Wang, and L. Wang, "Internvideo2: Scaling foundation models fornbsp;multimodal video understanding," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXV*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 396–416. [Online]. Available: https://doi.org/10.1007/978-3-031-73013-9_23

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[27] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou, "Univtg: Towards unified video-language temporal grounding," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 2782–2792.

[28] P. Lee and H. Byun, "Bam-detr: Boundary-aligned moment detection transformer fornbsp;temporal sentence grounding innbsp;videos," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 220–238. [Online]. Available: https://doi.org/10.1007/978-3-031-72627-9_13