A DUAL-MEMORY FRAMEWORK FOR LANGUAGE-DRIVEN ACTION LOCALIZATION IN STREAMING VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-driven action localization in streaming video aims to temporally ground natural language descriptions in an ongoing video stream, using only past and current frames. In contrast to offline settings with full video access, the streaming scenario lacks future context, leading to substantial uncertainty when recognizing partially observed actions. To address this issue, we draw inspiration from human cognition, where short-term experiences are gradually abstracted and integrated into long-term knowledge that helps interpret or anticipate actions under limited observation. We propose a dual-memory framework, which consists of two complementary memory modules: (1) an intra-memory module that encodes short-term temporal context from historical frames within the current video stream, capturing scene-specific static cues and scene-agnostic action cues, and (2) an inter-memory module that maintains long-term prototypical action representations distilled from previously observed videos. Importantly, the inter-memory is formed by integrating scene-agnostic action representations from intra-memory instances across multiple videos, reflecting how humans consolidate long-term knowledge from repeated short-term observations. During localization, the intra-memory provides the scene-specific reference, while the inter-memory retrieves plausible scene-agnostic action trends. By fusing them, our model synthesizes scene-specific future representations, compensating for the absence of future frames. Comprehensive experiments on several benchmark datasets demonstrate that our method consistently outperforms existing methods under the streaming setting, validating the effectiveness of our cognitively inspired dual-memory design.

1 INTRODUCTION

Understanding and localizing human actions in videos based on natural language queries has gained increasing attention in recent years (Gao et al., 2017; Hendricks et al., 2017a; Mithun et al., 2019; Zhang et al., 2020a; Mun et al., 2020). Language-driven action localization, also known as video moment retrieval (Gao & Xu, 2021; Wu et al., 2021; Zeng et al., 2021; Tang et al., 2021) or temporal sentence grounding (Yuan et al., 2019; Liu et al., 2021a;b), aims to identify the temporal boundaries of video segments that correspond to a given textual description. While prior methods have achieved promising performance under offline settings—where the entire video is fully observed—real-world applications, such as live surveillance, human-robot interaction, and egocentric video understanding, often demand a streaming setting, where only current and past frames are available (Gan et al., 2023; Wen et al., 2023; Liu et al., 2024).

The streaming scenario poses a fundamental challenge: the model must make decisions from past and current frames alone, without future observations. This constraint makes it particularly difficult to determine whether the current moment lies within the temporal segment of a described action, especially when the visual cues are ambiguous or incomplete. As shown in Figure 1(a), the lack of future context introduces significant uncertainty into frame-level reasoning and often leads to unreliable predictions.

To address this challenge, we draw inspiration from human cognition. Humans are capable of recognizing ongoing actions even when only partial observations are made (Kuznetcova et al., 2018;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

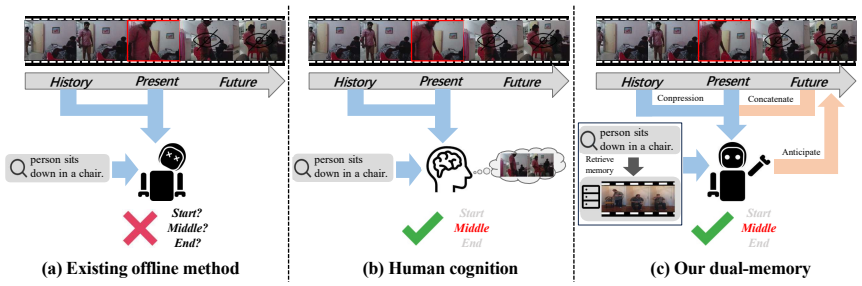

Figure 1: Comparison of different mechanisms for language-driven temporal grounding under partially observed streaming video scenarios.

Seegelke & Schack, 2016; Lu et al., 2022). This ability stems from the interplay between accumulated experience and current observations: through repeated exposure to similar actions across diverse contexts, we abstract scene-agnostic action patterns into *long-term memory*; at the same time, we retain *short-term memory* of the immediate visual stream, capturing scene-specific static references to interpret ongoing actions. Together, these memories enable us to anchor generalized long-term experience onto local short-term cues to anticipate future dynamics. As shown in Figure 1(b), when we see a person reaching toward a chair and pulling it, we may anticipate that he/she is about to sit down—even before the full action unfolds—by aligning the current scene context with the prototypical "person sitting down" action stored in memory.

Motivated by this insight, as shown in Figure 1(c), we propose a dual-memory framework for language-driven action localization in streaming videos. It comprises two interdependent modules: an intra-memory, which encodes short-term context from the current video stream into scene-specific static and scene-agnostic action representations, and a complementary inter-memory, which stores prototypical action representations learned across multiple videos as long-term knowledge.

Specifically, the intra-memory adaptively summarizes recent history into scene-specific static references and scene-agnostic action dynamics. These features are then combined with the original visual features and compactly summarized into a scene-specific static semantic channel and a scene-agnostic dynamic action channel via attention-based pooling. By incorporating saliency and recency biases, this compression ensures the memory prioritizes visually significant events and temporally proximal information, mirroring the human tendency to focus on recent, task-relevant observations.

Based on these short-term cues, the inter-memory maintains a header-slider architecture, where the header anchors linguistic semantics and the slider encodes temporal action knowledge. It employs a hierarchical retrieval process: first identifying the relevant prototypical slider of the action by calculating the semantic similarity between the textual query representations and headers, and then pinpointing the exact temporal stage of the action through correlation between the scene-agnostic action representations and each prototypical segment in the selected slider. During training, the inter-memory updates the identified prototypical slider segment with the current scene-agnostic action representation, ensuring recent experiences are effectively distilled into the corresponding temporal stage of the action. During inference, the inter-memory retrieves the future action dynamics after the identified slider segment, and anchors it back onto the scene-specific static reference to synthesize future representations. Thus, our framework compensates for missing future frames and achieves precise temporal localization under partial observation.

Our main contributions are:

- We propose a novel dual-memory-based framework for language-driven action localization in streaming videos, addressing the lack of future context by mimicking human-like memory usage.
- We design two complementary modules: an intra-memory module to capture short-term temporal context from historical frames, and an inter-memory module to encode prototypical action patterns as long-term knowledge distilled across videos.
- Extensive experiments on benchmark datasets show that our method achieves superior performance in the streaming setting, outperforming existing offline and streaming methods.

2 Related Work

2.1 Language-driven Action Localization

The task of language-driven action localization is first introduced by Gao et al. (Gao et al., 2017) and Hendricks et al. (Hendricks et al., 2017b). Liu et al. (Liu et al., 2023) construct a tree-structured proposal generation mechanism by merging adjacent frames with similar visual-linguistic semantics. Yang et al. (Yang et al., 2024b) introduce a dynamic pathway mechanism for adaptive coarse-to-fine and fine-to-coarse feature modeling based on the input query. Moment-DETR (Lei et al., 2021) adapts DETR framework to moment retrieval, treating moments as set elements to achieve end-to-end moment localization and saliency prediction. TR-DETR (Sun et al., 2024) together with TaskWeave (Yang et al., 2024a) further exploit the complementarity between action localization and highlight detection.

2.2 Streaming Video Understanding

Early research on streaming video understanding focuses on addressing one specific task, such as online action detection (De Geest et al., 2016; Xu et al., 2019; Wang et al., 2021b), online temporal action localization (Shou et al., 2018; Kang et al., 2021), and live highlight detection (Zhao et al., 2022). Recent MLLMs extend streaming understanding to open-ended video interaction. VideoLLM-Online (Chen et al., 2024), StreamChat (Liu et al., 2024), and LiveStar (Yang et al., 2025) focus on streaming interaction, covering multi-turn conversation, retrieval of the latest visuals during generation, and proactive response timing. Flash-VStream (Zhang et al., 2024) and streamBridge (Wang et al., 2025) instead reduce the cost of long streams, by compressing redundant frames or adapting offline Video-LLMs with memory buffering. These models mainly consume the stream history for dialogue, rather than localizing an action at the current frame.

Closer to our setting, TSGSV (Gan et al., 2023) first introduces language-driven action localization into the streaming setting, using a TwinNet architecture with knowledge distillation for future anticipation and a compressor to condense historical features. Zheng et al. (Zheng et al., 2025) further propose a hierarchical event memory for multi-scale event representation within each video. These methods primarily rely on visual information within individual video streams, while our method additionally leverages reusable action knowledge from diverse scenarios across videos, providing broader contextual knowledge that compensates for missing future frames.

3 Our Method

3.1 Overview

Given a video stream $V = \{v_t\}_{t=1}^{T}$ and a natural language query $Q = \{q_n\}_{n=1}^{N}$ describing a target action, the goal of language-driven action localization in streaming videos is to perform frame-wise temporal grounding according to Q. Specifically, at each time step t, the model estimates a triplet of scores $(\hat{y}_t^{\mathrm{s}}, \hat{y}_t^{\mathrm{m}}, \hat{y}_t^{\mathrm{e}})$, representing the probabilities that the current frame v_t belongs to the start, middle, or end of the action segment described by the query Q. Unlike traditional offline settings where the full video is accessible, the streaming scenario imposes a strict temporal constraint: the model can only access the history and current frames $\{v_1, \ldots, v_t\}$ during inference. The future frames $\{v_{t+1}, \ldots, v_T\}$ are unavailable. At each step, the observed video stream is divided into a history segment V_h and a current segment V_p, with lengths T_h and T_p, respectively. The two segments are processed jointly with the query Q by the visual and textual encoders of a base model to extract multimodal features. The base model performs prediction on V_p, estimating frame-wise importance scores and the start/middle/end probabilities of each frame.

To effectively tackle this streaming localization task, as shown in Figure 2, we propose a dual-memory framework that integrates two complementary modules into the base model: an *intra-memory* for summarizing short-term context, and an *inter-memory* for maintaining long-term knowledge across diverse scenarios. The base model serves as the localization backbone, while the dual-memory module augments its visual context under the streaming constraint. To build the intra-memory, we compress the historical features in V_h into a fixed-length representation guided jointly by the historical frame-wise importance scores and the query, thereby capturing important short-

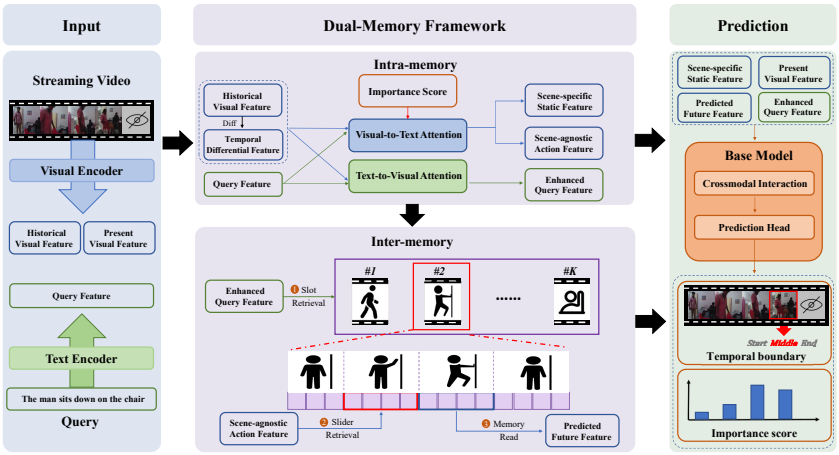

Figure 2: Overview of the proposed dual-memory framework. Given the history, the current stream, and a language query, the intra-memory compresses recent visual context by attention pooling with saliency and recency biases, and the inter-memory keeps these representations as slot-based prototypes of reusable action knowledge across videos. At inference, it retrieves the relevant prototypes and fuses them with the current short-term features to synthesize future representations for temporal grounding.

term context. The inter-memory organizes long-term knowledge with a header-slider structure: the header retrieves a relevant action prototype according to the query, and the slider locates the current action stage using the dynamic intra-memory representation. The matched slider segment is updated during training to consolidate action prototypes, while the following slider segment is read as a scene-agnostic future trend for prediction. This trend is then used to synthesize pseudo-future features, which complement the historical context captured by the intra-memory and together provide additional visual context for the base model to improve frame-wise prediction on V_p.

3.2 Base Model

We first describe the base localization model, which provides the multimodal features and prediction heads used by our dual-memory framework.

Visual and Text Encoder. We use pre-trained encoders to extract video and query features separately. For the video, we apply SlowFast (Feichtenhofer et al., 2019) and the visual encoder of CLIP (Radford et al., 2021) to obtain frame-level features. These two feature streams are L2-normalized and concatenated along the hidden dimension to form the raw video representations $F_h \in \mathbb{R}^{T_h \times d}$ and $F_p \in \mathbb{R}^{T_p \times d}$ for V_h and V_p, respectively. For the textual query, we obtain the token-level features $F_q \in \mathbb{R}^{N \times d}$ using the CLIP text encoder. Then, both video and query features are projected into a shared space using a two-layer MLP with LayerNorm and dropout.

Crossmodal Interaction. Following previous work (Lei et al., 2021; Sun et al., 2024), we adopt a Transformer-based encoder-decoder architecture to model the interaction between video and query features and predict temporal boundaries. The concatenated input features E_{input}, consisting of projected video and query tokens, are fed into a stack of Transformer encoder layers. Each encoder layer consists of a multi-head self-attention mechanism and a feed-forward network (FFN), with fixed positional encodings added to preserve temporal order. The encoder outputs contextualized representations E_{enc}. The decoder also consists of layers, each composed of self-attention, cross-attention over encoder outputs, and an FFN. The decoder is initialized with M learnable moment queries, which are refined through the layers to produce moment-aware embeddings $E_{\text{dec}} \in \mathbb{R}^{M \times d}$.

Prediction Head. To generate the final predictions in the current segment V_p, E_{enc} is used to compute the frame-wise importance scores $\{\hat{s}^p_{\text{imp},i}\}_{i=1}^{T_p}$ and the start/middle/end probabilities of each video frame via a linear layer. Meanwhile, E_{dec} is fed into a three-layer FFN with ReLU to pre-

dict M normalized temporal boundaries (center and width) and the boundary confidence scores $\hat{s}_{\text{loc}} \in \mathbb{R}^M$.

3.3 DUAL-MEMORY

We next detail the two memory modules. The intra-memory compresses the observed history into scene-specific static and scene-agnostic dynamic representations, while the inter-memory organizes long-term action knowledge with a header-slider structure for retrieval, update, and future-trend reading.

3.3.1 INTRA-MEMORY.

As the streaming video unfolds over time, not all visual content is equally important. Humans tend to retain only the most salient parts of recent observations—such as regions with query-relevant content or notable activity—while prioritizing information that is closer to the present. This selective retention helps focus attention and enables efficient reasoning with limited memory. Motivated by this, we design an intra-memory mechanism that compactly preserves informative short-term context from past observations. Specifically, it focuses on extracting scene-specific static references and action dynamics, ensuring that only the most task-relevant and temporally proximal information is maintained to support ongoing prediction. Given the historical visual features $F_h \in \mathbb{R}^{T_h \times d}$, we compute temporal differential features $\Delta F_h \in \mathbb{R}^{T_h \times d}$ to capture instantaneous action dynamics. Specifically, we set $\Delta F_h[1] = \mathbf{0}$ and compute $\Delta F_h[t] = F_h[t] - F_h[t-1]$ for $t = 2, \ldots, T_h$. These two features are concatenated to form an augmented representation $F'_h = \text{LayerNorm}([F_h \parallel \Delta F_h]) \in \mathbb{R}^{T_h \times 2d}$.

To compress F'_h into a fixed-size memory while focusing on visually salient and temporally recent cues, we introduce T_c learnable memory tokens $M_c \in \mathbb{R}^{T_c \times 2d}$. Instead of using only query-key attention, we incorporate a composite bias $B_{\text{total}} \in \mathbb{R}^{T_h}$ to guide the compression process:

$$\begin{cases} B_{\text{imp},i} = \log(\hat{s}^h_{\text{imp},i} + \epsilon), \\ B_{\text{dist},i} = \exp(-|i - T_h|), \quad \text{for } i = 1, \ldots, T_h \\ B_{\text{total},i} = B_{\text{imp},i} + B_{\text{dist},i}, \end{cases} \tag{1}$$

where $\hat{s}^h_{\text{imp},i}$ denotes the importance score of the i-th historical frame, cached from the previous streaming predictions when this frame was processed as part of the current segment; B_{imp} and B_{dist} represent the saliency prior and recency bias, respectively. This ensures the intra-memory focuses on visually significant and temporally proximal observations. The compressed state $\mathcal{M}_{\text{intra}}$ is extracted via an attention-based pooling:

$$A_c = \text{Softmax}\left(\frac{M_c F_h'^{\top}}{\sqrt{2d}} + B_{\text{total}}\right) \in \mathbb{R}^{T_c \times T_h}, \tag{2}$$

$$\mathcal{M}_{\text{intra}} = A_c F'_h \in \mathbb{R}^{T_c \times 2d}. \tag{3}$$

To achieve deep semantic alignment, we also apply a cross-attention mechanism to augment the textual query with visual context, where the query features F_q serve as queries and the visual features F'_h serve as keys and values:

$$F'_q = \text{CrossAttn}(F_q, F'_h, F'_h) \in \mathbb{R}^{N \times d}. \tag{4}$$

Finally, we decouple $\mathcal{M}_{\text{intra}}$ along the feature dimension into a scene-specific static channel $\mathcal{M}^{\text{static}}_{\text{intra}} = \mathcal{M}_{\text{intra}}[:, :\, d] \in \mathbb{R}^{T_c \times d}$ and a scene-agnostic action channel $\mathcal{M}^{\text{delta}}_{\text{intra}} = \mathcal{M}_{\text{intra}}[:, d\, :] \in \mathbb{R}^{T_c \times d}$ for subsequent hierarchical retrieval.

3.3.2 INTER-MEMORY.

While the intra-memory summarizes short-term context from the recent history, it lacks future cues and broader context, which are critical for accurate temporal grounding. To address this limitation, we introduce an inter-memory module that serves as a long-term repository of high-level, reusable action pattern knowledge. The inter-memory is maintained as a learnable slot bank

$\mathcal{M}_{\text{inter}} \in \mathbb{R}^{N_{\text{slot}} \times (L+1) \times d}$. During training, it is updated to consolidate reusable action patterns; during both training and inference, it is queried to read out long-term action trends. Each slot k adopts a header-slider architecture: a semantic header $H_k \in \mathbb{R}^d$ that anchors the slot to specific linguistic semantics, and a prototypical slider $\mathbf{S}_k \in \mathbb{R}^{L \times d}$ that stores the scene-agnostic action dynamics.

Memory Retrieval. First, we perform semantic indexing by computing the cosine similarity between the enhanced textual query F_q' and each slot header H_k to select the most relevant action prototype slider k^*:

$$k^* = \arg\max_k \text{CosineSim}(F_q', H_k). \tag{5}$$

We then perform a sliding-window matching in the selected slider to identify the temporal stage of current action. Taking the dynamic action channel $\mathcal{M}_{\text{intra}}^{\text{delta}}$ as a probe, we compute its correlation with the selected slider $\mathbf{S}_{k^*}$ to find the optimal alignment segment j^*:

$$j^* = \arg\max_j \text{CosineSim}(\mathcal{M}_{\text{intra}}^{\text{delta}}, \mathbf{S}_{k^*}[j : j + T_c]). \tag{6}$$

Memory Update. During training, we incorporate the current action representation into the long-term memory. We perform cross-attention between the localized slider segment $\mathbf{S}_{k^*}[j^* : j^* + T_c]$ and the intra-memory output to produce an updated feature $U_{k^*} \in \mathbb{R}^{T_c \times d}$:

$$U_{k^*} = \text{CrossAttn}(\mathbf{S}_{k^*}[j^* : j^* + T_c], \mathcal{M}_{\text{intra}}^{\text{delta}}, \mathcal{M}_{\text{intra}}^{\text{delta}}). \tag{7}$$

The corresponding slider is then updated via an adaptive gated combination, where the gate is computed from the matching confidence:

$$\lambda = \frac{1 + \text{CosineSim}(\mathcal{M}_{\text{intra}}^{\text{delta}}, \mathbf{S}_{k^*}[j^* : j^* + T_c])}{2}, \tag{8}$$

$$\tilde{\mathbf{S}}_{k^*}[j^* : j^* + T_c] = \lambda \cdot U_{k^*} + (1 - \lambda) \cdot \mathbf{S}_{k^*}[j^* : j^* + T_c], \tag{9}$$

where λ denotes the adaptive gating weight, which down-weights low-similarity updates from noisy or mismatched samples.

Memory Read. We recall future cues from the identified slider segment j^*. We extract the future action trend $\Delta F_f = \mathbf{S}_{k^*}[j^* + T_c : j^* + T_c + T_f]$ from the slider, where T_f denotes the future prediction length. To transform these scene-agnostic dynamics into scene-specific representations, we construct a reference sequence $F_{\text{ref}} = [\mathcal{M}_{\text{intra}}^{\text{static}}; F_p; F_f']$, where $F_f' \in \mathbb{R}^{T_f \times d}$ denotes the pseudo-future learnable tokens. The predicted future representation F_f is then synthesized by fusing the trend with this sequence:

$$F = \text{CrossAttn}(F_{\text{ref}}, \Delta F_f, \Delta F_f), \quad F_f = F[-T_f :]. \tag{10}$$

The estimated F_f is combined with $\mathcal{M}_{\text{intra}}^{\text{static}}$ and F_p to form the memory-enhanced visual context, which is fed to the base model for prediction on V_p, compensating for missing future frames under partial observation.

3.4 Loss Function

At each time step, the model predicts two types of frame-wise outputs within the current segment V_p: an importance score and the start/middle/end probabilities for each frame. It also predicts a confidence score for each temporal boundary. Binary cross-entropy supervises the importance scores with $\mathcal{L}_{\text{imp}}$, the start/middle/end probabilities with $\mathcal{L}_{\text{cls}}$, and the boundary confidence scores with $\mathcal{L}_{\text{loc}}$ (Appendix A.3).

Memory Loss. To ensure the quality of the inter-memory, we use dual constraints on the memory slots. To ensure that each memory slot represents a distinct action prototype, we use an orthogonality regularization on the slot headers to minimize the correlation between different slots:

$$\mathcal{L}_{\text{head}} = \frac{1}{N_{\text{slot}}(N_{\text{slot}} - 1)} \left\| \mathbf{H}\mathbf{H}^\top - \mathbf{I} \right\|_F^2, \tag{11}$$

where $\mathbf{H}$ denotes the matrix of normalized slot headers and $\mathbf{I}$ is the identity matrix. To ensure the sliders effectively capture prototypical action dynamics, we minimize the distance between the slider

segment $\mathrm{sg}(\mathbf{S}_{k^*}[j^* : j^* + T_c])$ and the current action representations $\mathcal{M}_{\text{intra}}^{\text{delta}}$, where $\mathrm{sg}(\cdot)$ stops the gradient, and keep continuity within each slider:

$$\mathcal{L}_{\text{slider}} = \left\|\mathrm{sg}(\mathbf{S}_{k^*}[j^* : j^* + T_c]) - \mathcal{M}_{\text{intra}}^{\text{delta}}\right\|_2^2 + \frac{1}{N_{\text{slot}}(L-1)} \sum_{k=1}^{N_{\text{slot}}} \sum_{j=1}^{L-1} \|\mathbf{S}_k[j+1] - \mathbf{S}_k[j]\|_2^2 . \tag{12}$$

$$\mathcal{L}_{\text{memory}} = \mathcal{L}_{\text{head}} + \mathcal{L}_{\text{slider}}. \tag{13}$$

Overall Loss. We sum the three cross-entropy losses and the two memory losses:

$$\mathcal{L} = \mathcal{L}_{\text{imp}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{head}} + \mathcal{L}_{\text{slider}}. \tag{14}$$

Table 1: Effect of integrating Dual-Memory (DM) on ActivityNet Caption and TACoS. † indicates our reproduced results under the streaming setting.

Model	ActivityNet Caption				TACoS			
	R1		mAP		R1		mAP	
	@0.5	@0.7	@0.5	@0.7	@0.5	@0.7	@0.5	@0.7
Base model	34.17	16.77	37.67	25.74	39.07	22.80	35.67	28.22
+ DM (Ours)	**36.81**	**19.56**	**40.02**	**27.89**	**42.45**	**25.77**	**38.66**	**30.59**
QD-DETR †	35.70	18.01	39.72	27.58	39.13	22.45	36.07	29.03
+ DM (Ours)	**38.11**	**20.26**	**42.48**	**30.73**	**42.01**	**25.58**	**38.19**	**31.62**
TaskWeave †	36.78	18.80	40.34	28.03	40.15	23.04	36.35	28.95
+ DM (Ours)	**39.17**	**20.75**	**42.95**	**31.10**	**42.90**	**25.82**	**38.54**	**31.67**
HEM †	44.83	25.74	48.43	37.82	43.74	29.91	41.83	36.12
+ DM (Ours)	**47.62**	**29.26**	**50.48**	**39.24**	**45.90**	**32.22**	**42.15**	**36.83**

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate our method on QVHighlights (Lei et al., 2021), TACoS (Regneri et al., 2013), and ActivityNet-Caption (Krishna et al., 2017), covering real-world YouTube videos, cooking activities (Rohrbach et al., 2012), and diverse activity videos (Heilbron et al., 2015), respectively. Detailed dataset statistics are provided in Appendix A.1.

Evaluation Metrics. Following previous work (Moon et al., 2023; Nishimura et al., 2024), we adopt Recall1@IoU μ and mean average precision (mAP) as evaluation metrics: (1) R1@μ, which denotes the percentage of top-1 predicted moment having IoU with ground truth larger than threshold μ. (2) mAP@μ, which denotes the mean Average Precision at IoU threshold μ. For fair comparison, we follow the offline evaluation protocol used in prior streaming work (Gan et al., 2023) to compute R1/mAP from the streaming predictions. Specifically, for each candidate video segment, we compute the product of the $\hat{y}^{\text{s}}$ at the first frame and the $\hat{y}^{\text{e}}$ at the last frame as its confidence score, and select high-confidence segments for evaluation. We also report Highlight Detection, which evaluates the frame-level predictions directly with mAP and HIT@1 (Appendix B.1).

Implementation Details. We use the same default hyperparameter configuration across all datasets and backbones. Specifically, we use 16 intra-memory tokens, 64 inter-memory slots, a temporal slider length of 64, a slider-window size T_c of 16, and a prediction length T_f of 8. All five loss terms are jointly optimized end-to-end with a uniform default weight of 1.0. We adopt the AdamW (Loshchilov & Hutter, 2017) optimizer with a learning rate of 0.0004 and a weight decay of 0.0001. We use a random seed of 2025 for all experiments and conduct experiments on a single NVIDIA RTX 4090 GPU.

4.2 COMPARISON WITH EXISTING METHODS

To assess the effectiveness and generalizability of our design across three datasets, we integrate Dual-Memory (DM) into the base model described in Section 3.2, QD-DETR (Moon et al., 2023),

Table 2: Performance comparison on QVHighlights. † indicates our reproduced results under the streaming setting.

Model	R1		mAP	
	@0.5	@0.7	@0.5	@0.7
TSGSV †	26.35	15.27	24.51	20.72
Base model	47.10	30.00	38.50	33.59
+ DM (Ours)	**51.47**	**35.82**	**43.91**	**39.46**
QD-DETR †	50.65	33.10	43.41	41.82
+ DM (Ours)	**53.91**	**37.33**	**47.05**	**45.43**
TaskWeave †	50.23	32.00	40.75	37.35
+ DM (Ours)	**53.41**	**35.77**	**44.12**	**40.65**
HEM †	54.21	36.95	44.79	41.28
+ DM (Ours)	**55.73**	**38.26**	**47.32**	**44.06**

Table 3: Performance comparison with existing methods on ActivityNet Caption and TACoS. Results of existing methods are obtained from the original papers.

Model	ActivityNet		TACoS	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7
Modified offline baselines				
VSLNet	12.89	5.05	25.74	12.60
2DTAN	8.39	2.96	6.82	3.32
SeqPAN	12.57	4.79	25.07	13.67
SMIN	7.47	2.64	6.00	2.92
Online baselines				
OadTR	23.27	10.97	21.12	10.92
LSTR	24.05	11.19	26.02	16.75
GateHUB	23.30	11.31	27.10	17.25
TSGSV	25.48	12.56	29.74	19.07
HEM	45.29	26.25	44.19	30.87
Ours	**47.62**	**29.26**	**45.90**	**32.22**

Table 4: Ablation on intra- and inter-memory modules on QVHighlights.

Model	Memory		R1		mAP	
	Intra	Inter	@0.5	@0.7	@0.5	@0.7
1	×	×	47.10	30.00	38.50	33.59
2	✓	×	49.10	32.53	41.16	36.08
3	✓	✓	51.47	35.82	43.91	39.46

TaskWeave (Yang et al., 2024a), and HEM (Zheng et al., 2025). As shown in Table 1 and Table 2, DM improves different backbones across datasets, with clear gains on high-precision metrics.

We further compare with representative modified offline and online baselines. Results of VSLNet (Zhang et al., 2020a), 2DTAN (Zhang et al., 2020b), SeqPAN (Zhang et al., 2021), SMIN (Wang et al., 2021a), OadTR (Wang et al., 2021b), LSTR (Xu et al., 2021), GateHUB (Chen et al., 2022), and TSGSV (Gan et al., 2023) are obtained from the TSGSV paper (Gan et al., 2023), while the HEM results are from the HEM paper (Zheng et al., 2025). As shown in Table 2 and Table 3, our method consistently outperforms these baselines under the streaming setting. We also evaluate Highlight Detection (HD), based on the frame-level importance scores that guide intra-memory compression, to directly assess prediction performance under partial observation instead of the offline segment-level protocol. HD results and ablations are provided in Appendix B.1.

4.3 Ablation Studies

Evaluation of memory modules. To evaluate the effectiveness of the proposed memory modules, we conduct ablation experiments by gradually enabling the intra-memory and inter-memory, as shown in Table 4. Starting from the base model (Row 1), introducing the intra-memory (Row 2) yields consistent improvements, with R1@0.5 and mAP@0.5 increasing by 2.0% and 2.7%, respectively, by effectively modeling short-term contextual cues. Furthermore, adding the inter-memory (Row 3) provides an additional performance boost, particularly in high-precision metrics where R1@0.7 increases from 32.53% to 35.82%. Whether this gain comes from the future trend after the matched slider segment is evaluated in Appendix B.3. We further evaluate localization performance when the current segment falls into different temporal stages of the ground-truth segment, with details provided in Appendix B.2.

Evaluation of intra-memory design. To evaluate the impact of different designs within the intra-memory module, we conduct ablation experiments on the temporal differential features and the attention bias, as shown in Table 5. Specifically, the absence of ΔF_h and B_{imp} results in a drop of

Table 5: Ablation on different designs of intra-memory on QVHighlights.

Model	Intra-memory			R1		mAP	
	ΔF_h	B_{imp}	B_{dist}	@0.5	@0.7	@0.5	@0.7
1	×	✓	✓	48.51	31.45	40.60	35.28
2	✓	×	✓	49.20	32.42	41.40	36.05
3	✓	✓	×	51.03	35.24	43.37	38.87
4	✓	✓	✓	51.47	35.82	43.91	39.46

2.9% and 2.2% in R1@0.5, respectively. Removing B_{dist} leads to consistent degradation on stricter metrics, reducing R1@0.7 by 0.58% and mAP@0.7 by 0.59%.

Evaluation of inter-memory design. We further investigate the core components of the inter-memory in Table 6. Removing the specialized header tokens (Row 1), by replacing them with mean-pooled slot features for memory retrieval, leads to a drop, proving that learned headers better represent distinct action prototypes. Replacing the slider segment matching (Row 2) with global slot-level interaction and updating also degrades performance, which underscores the necessity of our slider design for fine-grained phase alignment. Lastly, omitting the memory losses (Row 3), including orthogonality and temporal consistency constraints, results in a decline across all metrics. This validates that explicit structural regularization is essential for maintaining the diversity and continuity of the inter-memory. Additional analysis of memory slot capacity and inter-memory robustness is provided in Appendix B.5 and Appendix B.6. Cross-video transfer of the learned prototypes is evaluated in Appendix B.4.

Table 6: Ablation on header, slider, and memory loss on QVHighlights.

Model	Inter-memory			R1		mAP	
	header	slider	$\mathcal{L}_{\text{memory}}$	@0.5	@0.7	@0.5	@0.7
1	×	✓	✓	48.19	31.59	39.85	35.06
2	✓	×	✓	48.64	32.39	40.48	35.92
3	✓	✓	×	49.41	33.98	41.91	37.62
4	✓	✓	✓	51.47	35.82	43.91	39.46

Table 7: Efficiency comparison of TaskWeave with and without Dual-Memory (DM) on TACoS. Time denotes the average inference time per sample.

Model	Params	FPS	Time (ms)	R1@0.7
TaskWeave	7.91M	2663	0.37	23.04
+ DM (Ours)	14.91M	2250	0.46	25.82

4.4 Efficiency Analysis

To evaluate efficiency, we integrate DM into TaskWeave and report runtime and parameter overhead in Table 7. DM reduces speed only from 2663 to 2250 FPS, adds less than 0.1 ms inference time per sample, and improves R1@0.7. These results indicate a favorable accuracy-efficiency trade-off suitable for streaming applications; module-level profiling is provided in Appendix B.7. Qualitative results are provided in Appendix C.2.

5 Conclusion

We have presented a dual-memory framework for language-driven action localization in streaming video, addressing the lack of future context. Inspired by human memory, we leverage an intra-memory to encode short-term scene and action cues, and an inter-memory to capture prototypical action patterns abstracted across videos. By fusing both memories, our method compensates for missing future frames and improves action localization. Extensive experiments validate our design, showing consistent gains across benchmark datasets.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation, providing feedback on research methodology and experiments, and interpreting results. We have not used generative AI tools for developing theoretical models or conceptual frameworks, proposing or refining hypotheses, or implementing methods, and generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work: translated text was checked against the intended meaning, feedback on methodology and experiments was treated as advisory and final design decisions were made by the authors, and interpretations of the reported results were verified against the experimental numbers. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024.

Junwen Chen, Gaurav Mittal, Ye Yu, Yu Kong, and Mei Chen. Gatehub: Gated history unit with background suppression for online action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19925–19934, 2022.

Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *European Conference on Computer Vision*, pp. 269–284. Springer, 2016.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proc. of ICCV*, pp. 6201–6210, 2019.

Tian Gan, Xiao Wang, Yan Sun, Jianlong Wu, Qingpei Guo, and Liqiang Nie. Temporal sentence grounding in streaming videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4637–4646, 2023.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *Proc. of ICCV*, pp. 5277–5285, 2017.

Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proc. of ICCV*, pp. 1503–1512, 2021.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proc. of CVPR*, pp. 961–970, 2015.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *Proc. of ICCV*, pp. 5804–5813, 2017a.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *Proc. of ICCV*, pp. 5804–5813, 2017b.

Hyolim Kang, Kyungmin Kim, Yumin Ko, and Seon Joo Kim. Cag-qil: Context-aware actionness grouping via q imitation learning for online temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13729–13738, 2021.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proc. of ICCV*, pp. 706–715, 2017.

Evgeniya A Kuznetcova, Evgeniya A Kuznetcova, et al. The phenomenon of anticipation in psychology: Theoretical analysis. In *European Proceedings of Social and Behavioural Sciences*, volume 45, pp. 175–184. Future Academy, 2018.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Proc. of NeurIPS*, pp. 11846–11858, 2021.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proc. of CVPR*, pp. 11235–11244, 2021a.

Daizong Liu, Xiaoye Qu, and Pan Zhou. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. In *Proc. of EMNLP*, pp. 9302–9311, 2021b.

Daizong Liu, Xiang Fang, Pan Zhou, Xing Di, Weining Lu, and Yu Cheng. Hypotheses tree building for one-shot temporal sentence localization. In *Proc. of AAAI*, pp. 1640–1648, 2023.

Jihao Liu, Zhiding Yu, Shiyi Lan, Shihao Wang, Rongyao Fang, Jan Kautz, Hongsheng Li, and Jose M Alvare. Streamchat: Chatting with streaming video. *arXiv preprint arXiv:2412.08646*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Qihong Lu, Uri Hasson, and Kenneth A Norman. A neural network model of when to retrieve and encode episodic memories. *elife*, 11:e74445, 2022.

Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proc. of CVPR*, pp. 11592–11601, 2019.

WonJun Moon, Sangeek Hyun, Sanguk Park, Dongchan Park, and Jae-Pil Heo. Query - dependent video representation for moment retrieval and highlight detection. In *Proc. of CVPR*, pp. 23023–23033, 2023.

Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proc. of CVPR*, pp. 10807–10816, 2020.

Taichi Nishimura, Shota Nakada, Hokuto Munakata, and Tatsuya Komatsu. Lighthouse: A user-friendly library for reproducible video moment retrieval and highlight detection. In *Proc. of EMNLP*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, pp. 8748–8763, 2021.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *TACL*, pp. 25–36, 2013.

Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Proc. of ECCV*, pp. 144–157, 2012.

Christian Seegelke and Thomas Schack. Cognitive representation of human action: Theory, applications, and perspectives. *Frontiers in public health*, 4:24, 2016.

Zheng Shou, Junting Pan, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giroi Nieto, and Shih-Fu Chang. Online action detection in untrimmed, streaming videos-modeling and evaluation. In *ECCV*, volume 1, pp. 5, 2018.

Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4998–5007, 2024.

Haoyu Tang, Jihua Zhu, Meng Liu, Zan Gao, and Zhiyong Cheng. Frame-wise cross-modal matching for video moment retrieval. *IEEE Transactions on Multimedia*, pp. 1338–1349, 2021.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. *arXiv preprint arXiv:2505.05467*, 2025.

Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *Proc. of CVPR*, pp. 7026–7035, 2021a.

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7565–7575, 2021b.

Wujun Wen, Yunheng Li, Zhuben Dong, Lin Feng, Wanxiao Yang, and Shenlan Liu. Streaming video temporal action segmentation in real time. In *2023 18th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pp. 316–323. IEEE, 2023.

Ziyue Wu, Junyu Gao, Shucheng Huang, and Changsheng Xu. Diving into the relations: Leveraging semantic and visual structures for video moment retrieval. In *Proc. of ICME*, pp. 1–6, 2021.

Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall. Temporal recurrent networks for online action detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5532–5541, 2019.

Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021.

Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *Proc. of CVPR*, pp. 18308–18318, 2024a.

Shuo Yang, Xinxiao Wu, Zirui Shang, and Jiebo Luo. Dynamic pathway for query-aware feature learning in language-driven action localization. *IEEE Transactions on Multimedia*, 26:7451–7461, 2024b.

Zhenyu Yang, Zhiyuan He, Zeyuan Song, Xiang Li, Yuheng Ji, Wei Wang, Yu Liu, and Limin Wang. Livestar: Live streaming assistant for real-world online video understanding. *arXiv preprint arXiv:2511.05299*, 2025.

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Proc. of NeurIPS*, pp. 534–544, 2019.

Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. Multi-modal relational graph for cross-modal video moment retrieval. In *Proc. of CVPR*, pp. 2215–2224, 2021.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proc. of ACL*, pp. 6543–6554, 2020a.

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Siow Mong Rick Goh. Parallel attention network with sequence matching for video grounding. In *Proc. of ACL Findings*, pp. 776–790, 2021.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proc. of AAAI*, pp. 12870–12877, 2020b.

Yang Zhao, Xuan Lin, Wenqiang Xu, Maozong Zheng, Zhengyong Liu, and Zhou Zhao. Antpivot: Livestream highlight detection via hierarchical attention mechanism. *arXiv preprint arXiv:2206.04888*, 2022.

Minghang Zheng, Yuxin Peng, Benyuan Sun, Yi Yang, and Yang Liu. Hierarchical event memory for accurate and low-latency online video temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21589–21599, 2025.

A Implementation Details

A.1 Dataset Details

QVHighlights (Lei et al., 2021) is a dataset that supports both language-driven action localization and highlight detection. It contains 10,148 videos from YouTube, covering a wide range of real-world content, and includes 10,310 text queries associated with 18,367 annotated moments. The average video duration is 150 seconds, and the average annotated moment lasts 24.6 seconds. Due to the lack of ground-truth annotations on the official test set, evaluation is performed on the provided validation set.

TACoS (Regneri et al., 2013) consists of videos selected from the MPII Cooking Composite Activities video corpus (Rohrbach et al., 2012). It contains 18,818 video-query pairs of cooking activities. Each video contains an average of 148 queries, some of which are annotations of short video clips. The average video length is 286.59 seconds, and the average annotation event length is 27.88 seconds.

ActivityNet-Caption (Krishna et al., 2017) is collected for the dense video captioning task from ActivityNet (Heilbron et al., 2015), where the videos are associated with 200 activity classes. It consists of 19,811 videos with 71,957 queries. The average video length is 117.75 seconds, and each video contains 3.63 annotated moments, with an average duration of 37.14 seconds per moment. The public version of the dataset consists of a training set and two validation sets, containing 10,009, 4,917, and 4,885 videos, respectively.

A.2 Label Construction

Following previous work (Moon et al., 2023; Nishimura et al., 2024), we derive frame-level supervision from the available temporal annotations. For TACoS and ActivityNet-Caption, which provide temporal moment annotations but no frame-level saliency annotations, frames inside any ground-truth segment are assigned saliency label 1, and other frames are assigned 0. For QVHighlights, we use the provided frame-level saliency annotations.

For the start/middle/end supervision, we follow previous work (Gan et al., 2023) to generate Gaussian soft labels centered at the ground-truth start, midpoint, and end positions, respectively. The Gaussian widths are proportional to the ground-truth segment length, with ratios 0.25, 0.21, and 0.25 for the start, middle, and end labels. When multiple ground-truth segments are associated with the same video-query pair, their labels are merged by taking the frame-wise maximum.

A.3 BCE Losses

The frame-level importance score reflects the visual saliency of each frame. We supervise this prediction using frame-level saliency labels $s^p_{\text{imp},i} \in \{0, 1\}$ and apply a binary cross-entropy loss:

$$\mathcal{L}_{\text{imp}} = \frac{1}{T_p} \sum_{i=1}^{T_p} \text{BCE}(s^p_{\text{imp},i}, \hat{s}^p_{\text{imp},i}). \tag{15}$$

We treat the three boundary labels as a multi-label classification problem and apply binary cross-entropy independently for each class:

$$\mathcal{L}_{\text{cls}} = \frac{1}{T_p} \sum_{i=1}^{T_p} \sum_{c \in \{\text{s,m,e}\}} \text{BCE}(y^c_i, \hat{y}^c_i). \tag{16}$$

We use the IoU between each predicted temporal boundary and the ground-truth boundary as a soft label to supervise its confidence score $\hat{s}_{\text{loc},m}$ using the BCE loss:

$$\mathcal{L}_{\text{loc}} = \frac{1}{M} \sum_{m=1}^{M} \text{BCE}(s_{\text{loc},m}, \hat{s}_{\text{loc},m}). \tag{17}$$

B Additional Experiments

B.1 Highlight Detection Results

The Highlight Detection (HD) task aims to identify the most salient or important moments by assigning the frame-wise importance scores $\hat{s}_{\text{imp}}$ in a video sequence, which is crucial for applications such as video summarization and content retrieval. In our framework, HD is a related auxiliary frame-level task because the predicted importance scores guide intra-memory compression for historical feature selection. Moreover, when local observations are incomplete, the temporal trends retrieved from inter-memory can provide complementary cues for judging whether the current segment is important or relevant.

We evaluate the proposed dual-memory mechanism on the HD task across three datasets. As shown in Table 8, integrating DM consistently improves both mAP and HIT@1 across all backbones. Specifically, our design achieves significant mAP gains of 3.1%–6.9% and HIT@1 improvements of more than 3.1% across the three datasets. These results indicate that the dual-memory design not only enhances temporal reasoning but also generalizes well across different HD datasets and model frameworks.

Table 8: Effect of integrating Dual-Memory (DM) on different datasets for the HD task.

Model	QVHighlights		TACoS		ActivityNet	
	mAP	HIT@1	mAP	HIT@1	mAP	HIT@1
Base model	50.62	43.10	47.48	45.84	58.57	38.68
+ DM (Ours)	**57.02**	**50.10**	**54.31**	**48.09**	**61.11**	**49.33**
QD-DETR †	56.43	51.03	62.43	55.07	62.51	48.32
+ DM (Ours)	**59.86**	**54.74**	**65.49**	**57.91**	**64.97**	**51.41**
TaskWeave †	56.97	49.48	58.23	48.97	60.28	48.29
+ DM (Ours)	**60.14**	**52.85**	**61.38**	**56.31**	**64.17**	**53.49**

We further conduct HD ablations on QVHighlights. As shown in Table 9, adding the intra-memory improves mAP from 50.62 to 54.61, and further adding the inter-memory improves mAP to 57.02, confirming that both memory modules contribute to the auxiliary frame-level task.

B.2 Stage-wise Analysis

To evaluate the localization performance under different observation stages, we group each current segment according to the normalized progress of its center within the ground-truth segment: 0–30% as the start stage, 30–70% as the middle stage, and 70–100% as the end stage. Since this analysis focuses on local streaming predictions, we report frame-level mAP using the same binary importance labels as in the Highlight Detection task. As shown in Table 10, our method achieves balanced performance across the start, middle, and end stages. The start-stage result shows that our method remains effective when only partial action evidence is observed, while the end-stage result suggests that the synthesized future features do not destabilize action-ending decisions.

B.3 Future Trend Analysis

To evaluate whether the inter-memory benefit comes from the future trend after the matched slider segment, we keep the base model and the full dual-memory fixed on QVHighlights and change only the slider segment read out as ΔF_f. After match (Ours) reads the segment of length T_f after the matched window. The other rows read the matched window itself, the segment of length T_f before the match, and a random segment of the same length from the selected slider. All four rows use the same parameters. As shown in Table 11, After match (Ours) outperforms the other readouts on every metric, while the other readouts remain close to one another, indicating that the benefit is specific to the future trend after the matched slider segment.

Table 9: Ablation on memory modules for the HD task on QVHighlights.

Model	mAP	HIT@1
Base model	50.62	43.10
+ Intra	54.61	48.38
+ Intra + Inter	57.02	50.10

Table 10: Stage-wise analysis of frame-level performance on QVHighlights.

Stage	mAP	HIT@1
Start	56.73	49.42
Middle	58.36	52.73
End	56.53	48.97

B.4 Cross-video Transfer Analysis

To evaluate whether action prototypes in the inter-memory transfer across videos, we use a cross-dataset setup and evaluate every row on QVHighlights. Ours is the standard setting, with the whole model trained on QVHighlights. The ActivityNet row freezes an inter-memory trained on ActivityNet Caption and trains the remaining parameters on QVHighlights. The random row follows the same training, but replaces that inter-memory with a frozen randomly initialized one. As shown in Table 12, the ActivityNet inter-memory stays within a small gap of Ours on every metric, while both outperform the random inter-memory by a larger margin at the stricter thresholds. This indicates that the inter-memory preserves reusable, scene-agnostic action prototypes that remain effective on videos beyond those it was trained on.

B.5 Inter-memory Size

We evaluate the influence of the total number of memory slots in the inter-memory module on model performance. As illustrated in Figure 3, on both QVHighlights and ActivityNet, increasing the slot number from 64 first brings clear gains, indicating stronger capacity for modeling diverse long-term patterns. At larger sizes on QVHighlights, performance drops beyond 64, while on ActivityNet, 128 slots bring only a slight further improvement and 256 slots become comparable or slightly lower, suggesting diminishing returns as the inter-memory grows. Therefore, we use $N_{\text{slot}} = 64$ as a unified default that balances performance, memory cost, and retrieval efficiency.

B.6 Robustness to Noisy Updates

To evaluate the robustness of inter-memory against noisy updates, we perform a memory stress test by updating the trained inter-memory with mismatched video-query pairs. As shown in Table 13, with the similarity-gated memory update, such noisy updates cause only a small drop of 1.28% in R1@0.5 and 0.98% in R1@0.7. This indicates that the learned prototypes are largely retained and noisy or outlier writes have limited influence on the inter-memory.

B.7 Runtime Profiling

We conduct module-level runtime profiling for each dual-memory component on TACoS with TaskWeave. As shown in Table 14, intra-memory compression, memory retrieval, and future synthesis take 0.061 ms, 0.013 ms, and 0.014 ms per sample, respectively. The total inference time increases from 0.37 ms to 0.46 ms per sample, corresponding to 2250 FPS. These results show that the additional overhead of the dual-memory module is small compared with real-time processing requirements.

Table 11: Future trend analysis on QVHighlights with the base model. After match (Ours) reads the segment after the matched window.

ΔF_f	R1		mAP	
	@0.5	@0.7	@0.5	@0.7
Random	49.35	32.71	41.32	36.20
Matched window	49.88	33.24	41.78	36.65
Before match	49.12	32.48	41.15	36.02
After match (Ours)	**51.47**	**35.82**	**43.91**	**39.46**

Table 12: Cross-video transfer analysis of the inter-memory, evaluated on QVHighlights. ActivityNet freezes an inter-memory trained on ActivityNet Caption. Random replaces it with a frozen random inter-memory. Ours trains the whole model on QVHighlights.

Inter-memory	R1		mAP	
	@0.5	@0.7	@0.5	@0.7
ActivityNet	50.62	34.95	43.12	38.66
Random	49.55	32.90	41.55	36.42
Ours	**51.47**	**35.82**	**43.91**	**39.46**

C Additional Qualitative Results

C.1 Failure Analysis

We further analyze failure cases on QVHighlights. A query-video sample is regarded as failed if the top-1 prediction has a maximum IoU lower than 0.3 with all ground-truth windows. Among 1519 samples, 326 samples fail, accounting for 21.46% of the evaluation set. As shown in Figure 4, the failures mainly fall into five patterns:

- **Insufficient historical observations.** The observed history is too sparse to infer the target action reliably, which may cause missed starts or delayed triggers.
- **Duration distortion / temporal boundary errors.** The model may localize the action but misestimate its duration when the current execution speed differs from the learned temporal pattern.
- **Temporal-stage misalignment.** An incorrect match between the current stage and the slider position shifts the retrieved future trend, leading to boundary offsets.
- **Prototype slider mismatch.** Semantically similar but dynamically different prototype sliders can mislead prediction when the query deviates from common action patterns.
- **Saliency-driven history compression errors.** Over-weighting irrelevant salient frames or under-weighting subtle boundary cues during compression can remove useful history and hurt later retrieval and prediction.

These failures indicate that the main challenges lie in partial observation, temporal alignment, and memory-guided compression/retrieval.

C.2 Visualization

Figure 5 illustrates the dual-memory behavior from three views. In Figure 5(a), the highlighted frames show how the model identifies the query-relevant action segment under partial observation. In Figure 5(b), the intra-memory attention weights show how recent historical frames are selectively compressed into short-term context. In Figure 5(c), the inter-memory retrieval similarities show how relevant prototype slider segments are activated to obtain future trends for streaming localization.

To qualitatively evaluate the effectiveness of our modules, we present several examples of action localization results on the datasets. As shown in Figure 6(a) and (b), our proposed DM can accurately identify the relevance between event and query, and mitigate the impact of partially observed information. However, when the action subject deviates from common experience, DM may provide

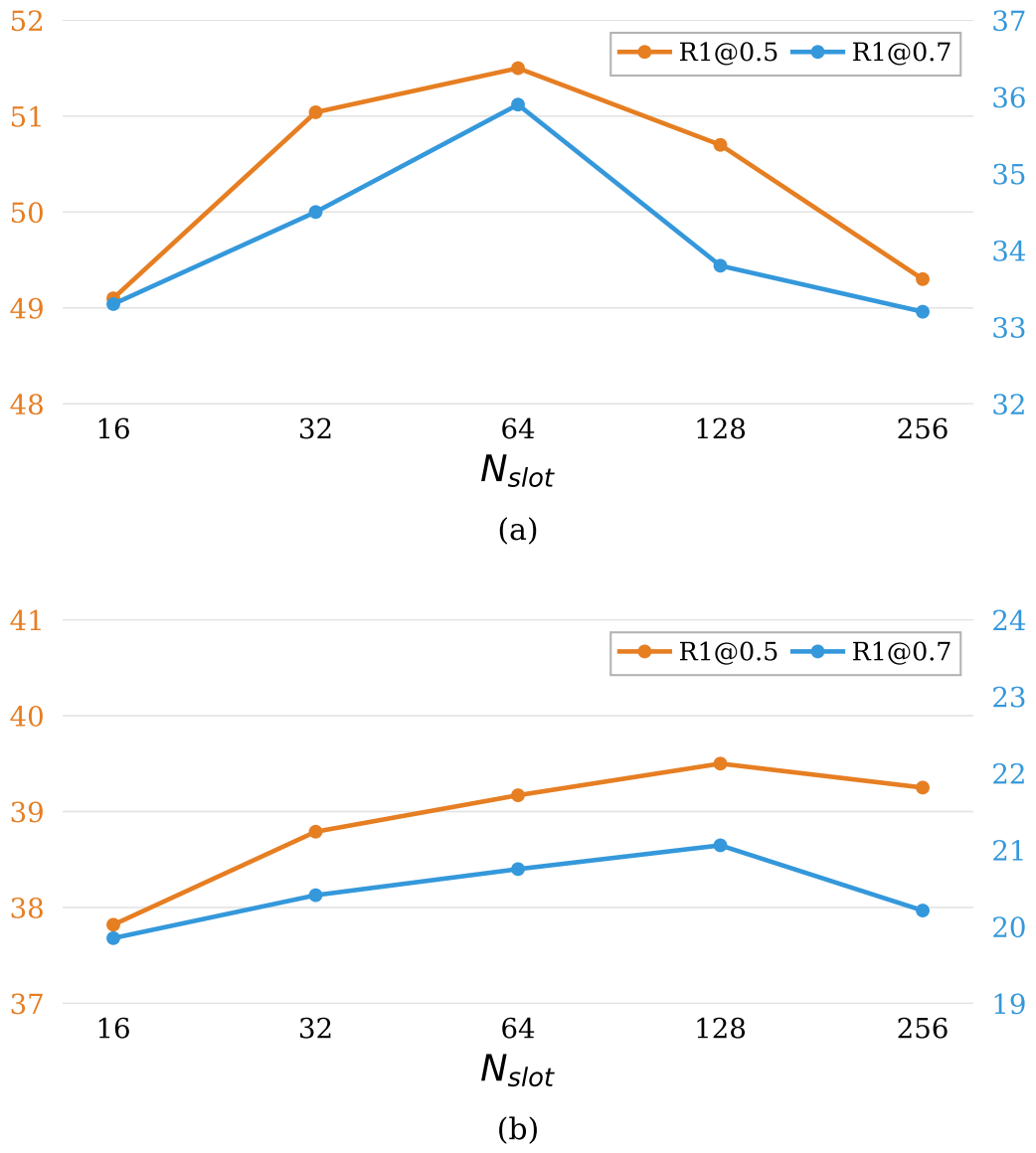

Figure 3: Results with different numbers of memory slots N_{slot}. (a) for base model+DM on QVHighlights Dataset, and (b) for TaskWeave+DM on ActivityNet Caption dataset.

Table 13: Robustness to noisy updates from mismatched video-query pairs.

Setting	R1		mAP	
	@0.5	@0.7	@0.5	@0.7
Noisy updates	52.13	34.79	43.18	39.47
Ours	53.41	35.77	44.12	40.65

misleading guidance. In Figure 6(c), the action of bringing something close to the mouth is typically followed by eating without being "intervened". In this case, DM relies on such prior assumptions and outputs a low start probability even when the event has begun, leading to misjudgment of the start of the event.

Table 14: Module-level runtime profiling on TACoS with TaskWeave.

Module	Runtime (ms)
Total inference time w/o DM	0.37
Intra-memory compression	0.061
Memory retrieval	0.013
Future synthesis	0.014
Total inference time w/ DM	0.46

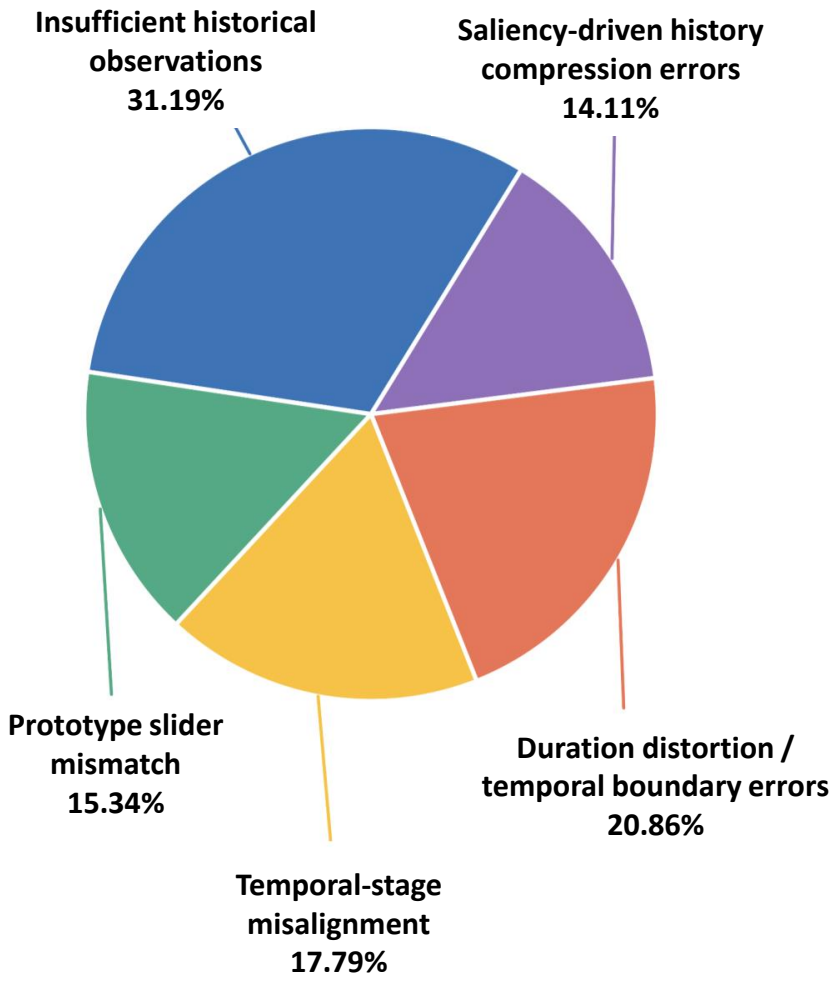

Figure 4: Failure analysis on QVHighlights.

We also provide additional memory visualizations to better understand how the proposed dual-memory framework uses historical and long-term cues. As shown in Figure 7 and 8, the visualizations include intra-memory attention weights and inter-memory retrieval similarities.

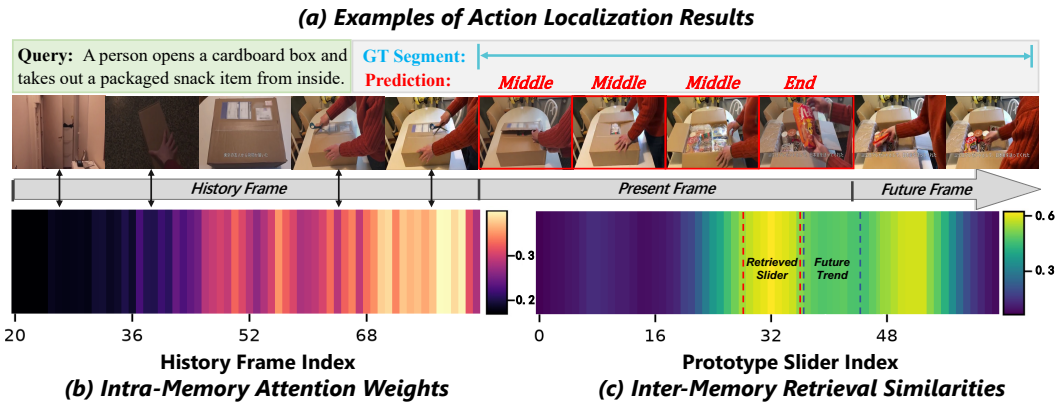

Figure 5: Examples of action localization results.

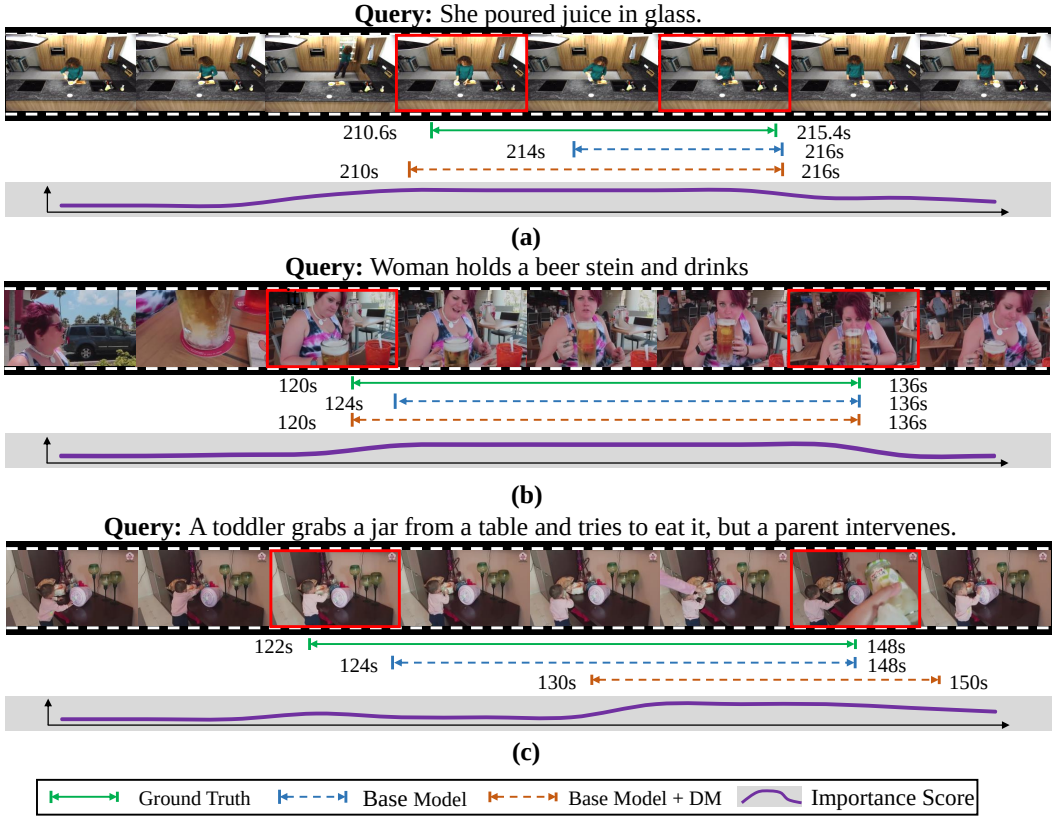

Figure 6: Examples of action localization results.

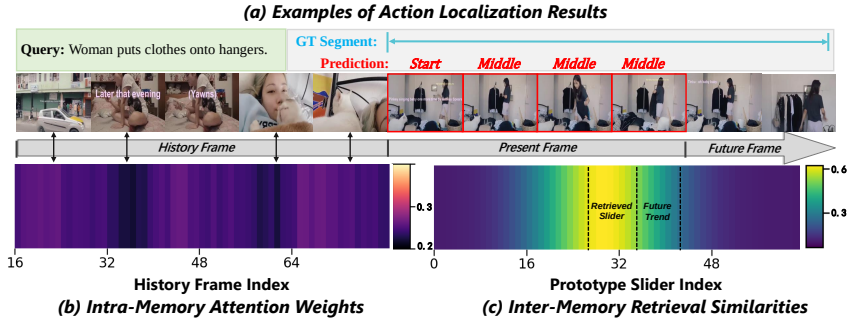

Figure 7: Visualization for intra-memory attention weights and inter-memory retrieval similarities.

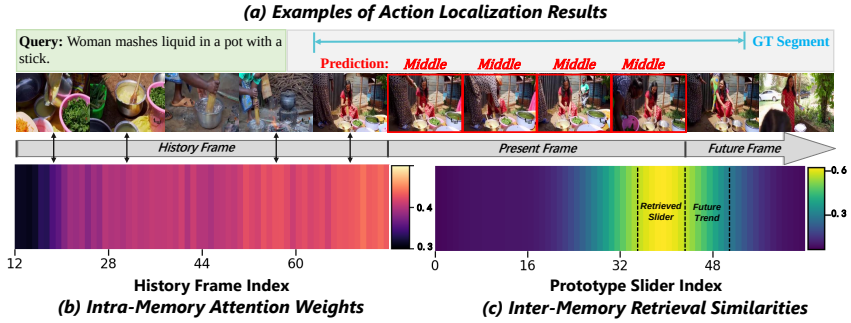

Figure 8: Visualization for intra-memory attention weights and inter-memory retrieval similarities.