HMER: HYPERBOLIC EVENT MEMORY AND REASONING FOR STREAMING TEXT-TO-VIDEO RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

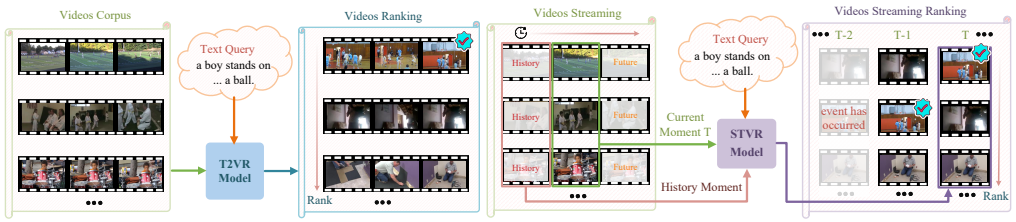
Existing text-to-video retrieval methods usually perform offline cross-modal matching on complete videos, which limits their applicability to streaming scenarios where videos arrive continuously and future content is unavailable. To address this issue, we study Streaming Text-to-Video Retrieval (STVR), which requires a model to identify relevant video streams while determining whether the target event is occurring at the current moment. This setting poses three main challenges: incomplete visual evidence, interference from past events, and unstable moment-level event predictions. To tackle these challenges, we propose Hyperbolic Memory-Based Event Reasoning (HMER), which jointly models event context, cross-modal hierarchy, and moment-level event states. Within HMER, we develop a Radial Valley-Guided Event Memory (RVG-EM) module to estimate event transitions from changes in hyperbolic radial trajectories. It aggregates historical information only within the current event, enriching local visual evidence while reducing cross-event interference. Since a video segment provides only local evidence for the complete event described by the text, we model the query and video segment as a general–specific semantic relation. Based on this view, we propose the Angular Cone Separation (ACS) loss to organize matching and non-matching representations with angular cone constraints, improving hierarchical alignment and cross-modal discrimination. Because semantic relevance alone cannot determine whether the target event is currently occurring, we further propose a Streaming Query-Aware Event Reasoning (SQ-AER) module. It jointly predicts query-matching confidence and event boundaries and combines them with historical high-confidence predictions for online event assessment. This enables the model to determine when retrieval results should be output. Experiments on DiDeMo, TVR, and Charades-STA show that HMER effectively models event context, cross-modal hierarchical relations, and moment-level event states, leading to consistent improvements on the STVR task.

1 INTRODUCTION

With the rapid growth of internet platforms (Kriz et al., 2025), urban sensing systems (Ki et al., 2025), and embodied AI devices (Fan et al., 2025), retrieving relevant video content from large-scale collections using natural language queries has become an important problem in computer vision (Bolya et al., 2026; Chen et al., 2025) and multimodal learning (Wang et al., 2025b; Zhao et al., 2026; Wang et al., 2025a; Yu et al., 2026). Existing text-to-video retrieval (T2VR) methods (Lan et al., 2025; Yang et al., 2026; Zhang et al., 2026a) usually assume that candidate videos are fully available and perform offline cross-modal matching on complete video content. As shown in Figure 1(a), this setting is suitable for static video libraries but less effective in streaming scenarios. In applications such as multi-camera surveillance (Le et al., 2025), live-stream moderation (Yew et al., 2026), and online robotic perception (Lee & Tian, 2026), video content arrives continuously and may not have a known end time. The system must determine whether the queried event is currently occurring using only the content observed so far, rather than waiting for the complete video. This creates a clear gap between conventional offline retrieval and streaming video retrieval.

To address this issue, we study Streaming Text-to-Video Retrieval (STVR). As shown in Figure 1(b), STVR takes a text query and multiple synchronously arriving video streams as input. At each time step, the model must determine whether the queried event is currently occurring and retrieve the

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.



(a) Conventional Offline Retrieval (T2VR) (b) Proposed Streaming Retrieval (STVR)

Figure 1: Comparison between conventional T2VR and the proposed STVR settings. T2VR performs offline matching on complete videos, whereas STVR performs online event assessment and video stream retrieval using current and historical observations.

corresponding video stream using only the current observation and historical information. Unlike one-shot matching on complete videos, STVR is an online decision process that updates as new video content arrives. The model must identify not only which stream is relevant to the query, but also whether the target event is occurring at the current moment.

This setting introduces three main challenges. First, the text query usually describes a complete event, while the current video segment provides only partial visual evidence. The model must identify query-relevant visual evidence from limited observations. Second, the accumulated video history may contain multiple events. Directly aggregating all past observations can introduce irrelevant or outdated information, so the model needs to retain context related to the current event. Third, semantic relevance does not necessarily mean that the target event is happening now. Even after identifying the correct stream, the model must still estimate the event state and temporal range. Changes in local observations can also make moment-level predictions unstable. Therefore, effective STVR requires joint modeling of cross-stream relevance, event context, and moment-level event states under incomplete observations.

Motivated by these challenges, we propose Hyperbolic Memory-Based Event Reasoning (HMER), which jointly models event context, cross-modal hierarchy, and moment-level event states. To reduce interference from unrelated historical events, we develop a Radial Valley-Guided Event Memory (RVG-EM) module. In hyperbolic space, radial positions are related to the semantic levels of features, while changes in event content may induce noticeable changes in radial trends (Desai et al., 2023). RVG-EM captures decrease-to-increase reversals in the radial trajectory as local valleys to estimate event transitions. It then aggregates only historical segments from the current event. This provides useful context for the current segment while reducing interference from other events.

Even with intra-event context, a video representation still provides only local visual evidence for the complete event described by the query. Therefore, the query–video relation is simply semantic similarity, but a general–specific hierarchical correspondence between abstract event semantics and concrete visual evidence. To model this relation, we propose the Angular Cone Separation (ACS) loss. ACS treats the query representation as the cone apex, constrains matching video representations to lie inside the cone, and suppresses non-matching representations from entering it. This geometric constraint preserves the hierarchical relation between the full query and local video evidence while improving the separation between matching and non-matching representations.

Semantic matching alone is insufficient for streaming retrieval. A video stream may be relevant to the query, while the target event is not occurring at the current moment. We therefore design a Streaming Query-Aware Event Reasoning (SQ-AER) module. Based on the event context provided by RVG-EM, SQ-AER performs fine-grained text–video interaction to predict query-matching confidence and target event boundaries for the current segment. It also retains historical high-confidence predictions and uses their predicted event intervals to determine whether retrieval results should be triggered at the current moment. In this way, HMER jointly models event context, hierarchical cross-modal matching, and moment-level event states to handle incomplete observations, cross-event interference, and online event occurrence estimation. Experiments on DiDeMo, TVR, and Charades-STA demonstrate the effectiveness of HMER. Our main contributions are as follows:

- We investigate Streaming Text-to-Video Retrieval (STVR), extending conventional one-shot retrieval on complete videos to online event retrieval under continuous observation.

This setting requires models to handle incomplete visual evidence, interference from past events, and moment-level event occurrence estimation.

- We propose a hyperbolic event modeling mechanism built on RVG-EM and ACS. It uses hyperbolic radial changes to capture event transitions and models global text semantics and local video evidence as a general–specific hierarchy. This enables unified modeling of event evolution and cross-modal hierarchical relations.
- We propose SQ-AER, which reformulates cross-stream similarity ranking as joint reasoning over stream relevance and event occurrence. It predicts matching confidence and event boundaries and uses historical high-confidence predictions to control when retrieval results are triggered, reducing reliance on instantaneous similarity.

2 Related Work

2.1 Text-to-Video Retrieval

Text-to-Video Retrieval (T2VR) aims to retrieve videos that are semantically relevant to natural language queries. Existing methods mainly improve cross-modal matching through semantic augmentation (Wu et al., 2023; Hur et al., 2025), uncertainty modeling (Fang et al., 2023; Li et al., 2023), and representation refinement (Liu et al., 2022; Wang et al., 2023; Jin et al., 2024). For untrimmed videos, Partially Relevant Video Retrieval (PRVR) (Dong et al., 2022) further focuses on identifying query-relevant segments and reducing interference from irrelevant content through robust semantic representation learning (Moon et al., 2026; Li et al., 2026b), implicit segment modeling (Wang et al., 2024; Li et al., 2025), temporal span modeling (Song et al., 2025), and context enhancement (Li et al., 2026a;c). These methods improve local query–video matching in complex videos, but they usually assume that the complete video is available during retrieval. In contrast, Streaming Text-to-Video Retrieval (STVR) operates on continuously arriving video streams, where the target event may not yet be fully observed. The model must therefore use current and historical observations to assess query–stream relevance and determine whether the target event is currently occurring.

2.2 Streaming Video Understanding

Streaming video understanding requires models to continuously process newly arriving video content without access to future observations. A key challenge is to make effective use of historical information while updating predictions over time (Gan et al., 2023; Di et al., 2025). Existing studies mainly focus on online event localization, streaming question answering, and proactive response. Online event localization methods often use teacher guidance (Gan et al., 2023) or combine hierarchical memory with future prediction (Zheng et al., 2025) to improve temporal modeling and reduce boundary prediction delay. Streaming question answering methods rely on mechanisms such as KV-Cache retrieval (Di et al., 2025) and hierarchical event memory (Liang et al., 2026; Zhang et al., 2026c; Jiang et al., 2026) to access long-term context efficiently. Recent studies further model response timing (Qian et al., 2025; Zhang et al., 2026b) and use readiness signals (Azad et al., 2026) or temporal activation predictions (Kim et al., 2026) to determine when enough evidence is available for an answer. These methods improve event localization, historical reasoning, and timely response in a single video stream. In contrast, STVR considers multiple video streams arriving simultaneously. It requires the model to continuously compare their relevance to the query and determine whether the target event is occurring at the current moment.

3 Methodology

3.1 Problem Statement and Framework Overview

Problem Statement Given a natural language query Q and a set of N_v untrimmed video streams $\mathcal{V} = \{V_n\}_{n=1}^{N_v}$, the n-th stream is represented as$V_n = \{v_{n,i}\}_{i=1}^{T}$, where $v_{n,i}$ denotes the video segment arriving at time i and T is the total number of time steps. For each query Q, there is a target video stream $V^\star \in \mathcal{V}$. The queried event occurs within $[\tau_\star^s, \tau_\star^e]$, where $\tau_\star^s$ and $\tau_\star^e$ denote its start and end times, respectively. In STVR, all candidate video streams arrive synchronously. At time t, the model can only access the observations received so far: $V_n^t = \{v_{n,i}\}_{i=1}^{t}$. The observable

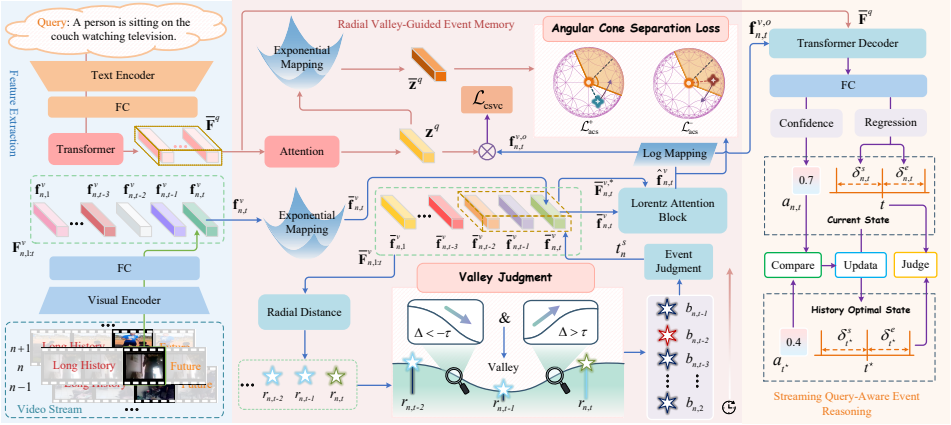


Figure 2: Overview of the proposed HMER framework. HMER unifies hyperbolic event memory, hierarchical cross-modal representation learning, and query-aware event reasoning for streaming text-to-video retrieval.

video set is therefore $\mathcal{V}^t = \{V_n^t\}_{n=1}^{N_v}$. Given Q and $\mathcal{V}^t$, the online retrieval model $\mathcal{M}$ produces the retrieval output $\mathbf{s}^t = \mathcal{M}(Q, \mathcal{V}^t)$. Let $\mathcal{V}_K^t \subseteq \mathcal{V}^t$ denote the Top-K candidate video streams ranked by their matching scores at time t. If the model determines that the queried event is occurring at the current moment, it outputs the Top-K candidate set: $\mathbf{s}^t = \mathcal{V}_K^t$. Otherwise, it outputs an empty set: $\mathbf{s}^t = \varnothing$. Over the full streaming process, the model produces an ordered sequence of retrieval outputs: $\mathcal{S} = (\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^T)$.

Framework Overview As shown in Figure 2, HMER mainly consists of two core modules in addition to the feature extraction module: RVG-EM and SQ-AER. RVG-EM uses local valleys in the hyperbolic radial trajectory to estimate event transitions and constructs a memory-enhanced representation from the current event context. This provides useful historical information for the current video segment while reducing interference from other events. The cosine similarity between each memory-enhanced video representation and the sentence-level query representation is then used to rank and select candidate video streams. SQ-AER performs fine-grained cross-modal interaction on the selected candidates to predict query-matching confidence and event boundaries. It also maintains historical high-confidence predictions and uses the corresponding event interval to determine whether retrieval results should be output at the current moment. During training, ACS is used to model the hierarchical relation between the abstract semantics of the query and the concrete visual evidence. This constraint encourages structured cross-modal representations and improves both video stream retrieval and event localization.

3.2 Feature Extraction

Given a natural language query Q, we use a pre-trained RoBERTa encoder (Liu et al., 2019) to extract token-level features and project them into a d-dimensional joint embedding space through a fully connected (FC) layer, yielding $\mathbf{F}^q = [\mathbf{f}_1^q, \mathbf{f}_2^q, \ldots, \mathbf{f}_{N_q}^q]^\top \in \mathbb{R}^{N_q \times d}$, where N_q is the number of query tokens and $\mathbf{f}_i^q \in \mathbb{R}^d$ denotes the feature of the i-th token. We then apply a Transformer encoder (Vaswani et al., 2017) to obtain the context-enhanced query representation $\bar{\mathbf{F}}^q = [\bar{\mathbf{f}}_1^q, \bar{\mathbf{f}}_2^q, \ldots, \bar{\mathbf{f}}_{N_q}^q]^\top \in \mathbb{R}^{N_q \times d}$. For the n-th video stream V_n, we use a pre-trained ResNet-152 (He et al., 2016) to extract visual features from incoming video segments and project them into the same d-dimensional space through a separate FC layer. Let $\mathbf{f}_{n,t}^v \in \mathbb{R}^d$ denote the feature of the video segment arriving at time t. The video feature sequence observed up to time t is $\mathbf{F}_{n,1:t}^v = [\mathbf{f}_{n,1}^v, \mathbf{f}_{n,2}^v, \ldots, \mathbf{f}_{n,t}^v]^\top \in \mathbb{R}^{t \times d}$.

3.3 Radial Valley-Guided Event Memory

Video segments usually provide only partial, concrete visual evidence of a target event, while text queries often describe the abstract semantics of the entire event. This creates a general–specific hierarchical correspondence. In streaming scenarios, video segments arrive continuously and event content changes over time. Directly aggregating all historical segments may introduce information

from previous events and interfere with the current representation. Compared with Euclidean space, hyperbolic space has exponentially growing capacity with respect to radius, making it suitable for representing complex hierarchical structures (Krioukov et al., 2010; Nickel & Kiela, 2017). Moreover, the radial position of a feature can reflect its semantic level (Desai et al., 2023; Pal et al., 2025). Based on this property, we construct RVG-EM. It estimates event boundaries from temporal changes in radial distance and limits feature interaction to the current event, thereby reducing cross-event interference.

We adopt the Lorentz model for its numerical stability and fix the curvature to -1 (Nickel & Kiela, 2018). Specifically, given a segment feature $\mathbf{f}^v_{n,t} \in \mathbb{R}^d$ arriving at time t from the n-th video stream, we map it to the Lorentz manifold $\mathbb{L}^m$ via the exponential map at the origin, as defined in Eq. equation 15, yielding the hyperbolic video representation $\bar{\mathbf{f}}^v_{n,t}$. At time t, the hyperbolic feature sequence of the video stream is represented as $\bar{\mathbf{F}}^v_{n,1:t} = [\bar{\mathbf{f}}^v_{n,1}, \bar{\mathbf{f}}^v_{n,2}, \ldots, \bar{\mathbf{f}}^v_{n,t}]^\top \in \mathbb{R}^{t\times(m+1)}$. or the hyperbolic representation $\bar{\mathbf{f}}^v_{n,t}$, the radial distance to the reference origin $\boldsymbol{o}$ is defined as $r_{n,t} = \operatorname{arcosh}(-\langle \boldsymbol{o}, \bar{\mathbf{f}}^v_{n,t}\rangle_{\mathcal{L}})$, where $\langle\cdot,\cdot\rangle_{\mathcal{L}}$ denotes the Lorentz inner product defined in Eq. equation 12. The radial trajectory up to time t is denoted as $(r_{n,1}, r_{n,2}, \ldots, r_{n,t})$.

Existing studies on hyperbolic vision-language representations suggest that the radial position of a feature is related to its semantic level (Desai et al., 2023; Pal et al., 2025). Video segments within the same event usually exhibit continuous semantic changes, leading to relatively smooth radial trajectories. When an event transition occurs, the semantic state changes and the radial trend may also change noticeably. We therefore focus on local valleys in the radial trajectory. A clear change from decreasing to increasing radial distance indicates a reversal in the radial trend. We use such a valley as a cue for event boundaries and estimate the start of the current event online. For $t \geq 3$, we calculate the first-order differences of the radial distance between consecutive time steps:

$$\Delta r_{n,t,t-1} = r_{n,t} - r_{n,t-1}, \qquad \Delta r_{n,t-1,t-2} = r_{n,t-1} - r_{n,t-2}. \tag{1}$$

If the radial distance drops significantly from $t-2$ to $t-1$ and rises significantly from $t-1$ to t, then $t-1$ forms a local valley. We treat the interval between $t-1$ and t as an event boundary. The corresponding boundary indicator is defined as

$$b_{n,t-1} = \mathbb{I}\left[\Delta r_{n,t-1,t-2} < -\tau \ \wedge \ \Delta r_{n,t,t-1} > \tau\right], \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function, τ is the radial change threshold, and $\wedge$ denotes logical "AND". When $b_{n,t-1} = 1$, the segment at time t is regarded as the start of a new event. For $t \geq 3$, the starting point of the event memory for the n-th video stream at time t is defined as

$$t^s_{n,t} = \max\left(\{1\} \cup \{j+1 \mid j \in \{2,\ldots,t-1\},\ b_{n,j} = 1\}\right). \tag{3}$$

For $t < 3$, we set $t^s_{n,t} = 1$. The event memory therefore starts after the most recently detected boundary. If no boundary has been detected, it starts from the first segment. The hyperbolic representation of the current event window is given by

$$\bar{\mathbf{F}}^{v,*}_{n,t} = [\bar{\mathbf{f}}^v_{n,t^s_{n,t}}, \bar{\mathbf{f}}^v_{n,t^s_{n,t}+1}, \ldots, \bar{\mathbf{f}}^v_{n,t}]^\top \in \mathbb{R}^{(t-t^s_{n,t}+1)\times(m+1)}. \tag{4}$$

We use the current segment representation $\bar{\mathbf{f}}^v_{n,t}$ as the query and the current event window $\bar{\mathbf{F}}^{v,*}_{n,t}$ as the keys and values. A Lorentz attention module (Li et al., 2025) then produces the event-enhanced representation $\hat{\mathbf{f}}^v_{n,t}$. Since the event window contains both the current segment and historical segments from the same event, this operation preserves current information while aggregating relevant intra-event context. Restricting feature interaction to the current event window also reduces interference from other events. Through a logarithmic map at the origin, defined in Eq. equation 16, we map $\hat{\mathbf{f}}^v_{n,t}$ back to the tangent space and obtain the Euclidean video representation $\mathbf{f}^{v,o}_{n,t}$ for cross-modal similarity computation.

3.4 Angular Cone Separation Loss

Sentence-level queries describe the overall semantics of a target event, while relevant video segments usually provide only local visual evidence from specific scenes and action states. We therefore view the query as a general semantic description and the relevant video segment as a specific visual instance. To model this general–specific hierarchical relation, we propose the ACS, building

on Entailment Loss. ACS uses the query representation as the cone apex, constrains matching video representations to lie inside the query cone, and suppresses non-matching representations from entering it, as illustrated in Fig. 3.

We obtain the sentence-level text representation $\mathbf{z}^q$ via attention pooling (Li et al., 2026a), and map it to the same Lorentz space as the video representations using the exponential map, yielding $\bar{\mathbf{z}}^q$. A hyperbolic entailment cone is constructed with $\bar{\mathbf{z}}^q$ as the apex, and its half-aperture angle is defined as$\omega\left(\bar{\mathbf{z}}^q\right) = \arcsin(2c/\left\|\bar{\mathbf{z}}_s^q\right\|)$, where $\bar{\mathbf{z}}_s^q$ denotes the spatial component of the hyperbolic text representation, and c controls the cone aperture. We set $c = 0.1$. For the event-augmented representation $\hat{\mathbf{f}}_{n,i}^v$ of the n-th video stream at time i, the angle with respect to the query cone axis is defined as

$$\phi\left(\bar{\mathbf{z}}^q, \hat{\mathbf{f}}_{n,i}^v\right) = \arccos\left(\frac{\hat{\mathbf{f}}_{n,i,0}^v + \bar{\mathbf{z}}_0^q \langle \bar{\mathbf{z}}^q, \hat{\mathbf{f}}_{n,i}^v \rangle_{\mathcal{L}}}{\left\|\bar{\mathbf{z}}_s^q\right\| \sqrt{\langle \bar{\mathbf{z}}^q, \hat{\mathbf{f}}_{n,i}^v \rangle_{\mathcal{L}}^2 - 1}}\right), \tag{5}$$

where $\bar{\mathbf{z}}_0^q$ and $\hat{\mathbf{f}}_{n,i,0}^v$ denote the time components of the hyperbolic text and video representations, respectively.

For video segments that match the query, we require their representations to satisfy $\phi(\bar{\mathbf{z}}^q, \hat{\mathbf{f}}_{n,i}^v) \leq \omega(\bar{\mathbf{z}}^q)$. This constraint places matching video content within the semantic region defined by the query, thereby preserving the general–specific relation between sentence-level event semantics and local visual evidence. For non-matching segments, we suppress their entry into the query cone to reduce confusion with the query semantics. For a given query, let $\mathcal{P}_+$ and $\mathcal{P}_-$ denote the index sets of positive and negative video segments in the current batch, respectively, where (n, i) denotes the segment from the n-th video stream at time i. The positive sample loss is defined as

$$\mathcal{L}_{\text{acs}}^+ = \frac{1}{|\mathcal{P}_+|} \sum_{(n,i)\in\mathcal{P}_+} \max\left(0,\, \phi\left(\bar{\mathbf{z}}^q, \hat{\mathbf{f}}_{n,i}^v\right) - \omega\left(\bar{\mathbf{z}}^q\right)\right). \tag{6}$$

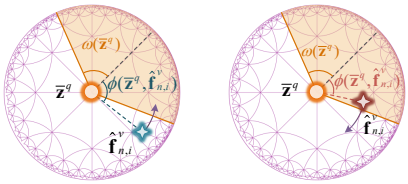
This loss penalizes positive samples outside the query cone and encourages them to move into the corresponding semantic region. The negative sample loss is defined as

$$\mathcal{L}_{\text{acs}}^- = \frac{1}{|\mathcal{P}_-|} \sum_{(n,i)\in\mathcal{P}_-} \max\left(0,\, \omega\left(\bar{\mathbf{z}}^q\right) - \phi\left(\bar{\mathbf{z}}^q, \hat{\mathbf{f}}_{n,i}^v\right)\right). \tag{7}$$

This loss penalizes negative samples inside the query cone, preventing irrelevant segments from entering the query-defined semantic region. The overall ACS loss is defined as $\mathcal{L}_{\text{acs}} = \mathcal{L}_{\text{acs}}^+ + \mathcal{L}_{\text{acs}}^-$. Compared with the traditional Entailment Loss, ACS explicitly constrains how video representations are distributed with respect to the query cone in angular space. Matching segments are kept within the query-defined semantic region to preserve the general–specific relation between global query semantics and local visual evidence. Non-matching segments are kept outside this region, improving the separation between matching and non-matching representations.

3.5 Streaming Query-Aware Event Reasoning

Although RVG-EM provides relevant context within the current visual event, coarse-grained video stream retrieval alone cannot determine whether the target event is occurring at the current moment. We therefore design the Streaming Query-Aware Event Reasoning (SQ-AER) module. SQ-AER performs fine-grained cross-modal interaction to predict query-matching confidence and event boundaries for the current segment. It also combines current predictions with historical high-confidence information for online decision-making.



(a) $\mathcal{L}_{\text{acs}}^+$ (b) $\mathcal{L}_{\text{acs}}^-$

Figure 3: Geometric illustration of ACS. Positive samples are encouraged to lie inside the query entailment cone, while negative samples are encouraged to remain outside the cone.

Given the event-augmented representation $\mathbf{f}_{n,t}^{v,o}$ generated by RVG-EM for the n-th video stream at time t, together with the context-enhanced text token features $\bar{\mathbf{F}}^q$, we use a Transformer decoder (Vaswani et al., 2017) to model the fine-grained relation between the current video segment and the query. We use $\mathbf{f}_{n,t}^{v,o}$ as the query and $\bar{\mathbf{F}}^q$ as the key

and value, producing the query-aware representation $\mathbf{h}_{n,t}$. A fully connected layer further produces the time-step representation $\bar{\mathbf{h}}_{n,t}$, which is fed into the confidence head $\mathcal{H}_{\text{con}}$ and the boundary regression head $\mathcal{H}_{\text{reg}}$ (Zheng et al., 2025):

$$a_{n,t} = \mathcal{H}_{\text{con}}\left(\bar{\mathbf{h}}_{n,t}\right), \qquad \left(\delta^s_{n,t}, \delta^e_{n,t}\right) = \mathcal{H}_{\text{reg}}\left(\bar{\mathbf{h}}_{n,t}\right). \tag{8}$$

Here, $a_{n,t}$ denotes the matching confidence between the current video segment and the target event. $\delta^s_{n,t}$ and $\delta^e_{n,t}$ denote the temporal distances from the current anchor t to the predicted event start and end boundaries, respectively. The prediction state of the n-th video stream at time t is represented as $\mathcal{S}_{n,t} = \left(t, a_{n,t}, \delta^s_{n,t}, \delta^e_{n,t}\right)$. This state jointly describes the prediction time, matching confidence, and event span, allowing predictions from different candidates to be compared in a unified form.

Let $\mathcal{C}_t$ denote the set of candidate video streams at time t. The candidate with the highest confidence is defined as $\hat{n}_t = \arg\max_{n \in \mathcal{C}_t} a_{n,t}$. To reduce the effect of short-term confidence fluctuations, the model maintains the highest-confidence prediction observed up to the current time: $\mathcal{S}^\star_t = (t^\star, a_{t^\star}, \delta^s_{t^\star}, \delta^e_{t^\star})$, where $t^\star$ denotes the anchor time of the retained prediction, $a_{t^\star}$ is the highest matching confidence observed up to time t, and $\delta^{s,\star}_t$ and $\delta^{e,\star}_t$ are the predicted distances from this anchor to the event start and end boundaries.

The current highest-confidence candidate is compared with the retained historical state. If its confidence $a_{\hat{n}_t,t}$ is higher than the confidence $a_{t^\star}$ of the retained state, we update $\mathcal{S}^\star_t$ with $\mathcal{S}_{\hat{n}_t,t}$; otherwise, $\mathcal{S}^\star_{t-1}$ is retained. The event interval corresponding to the retained state is defined as $\mathcal{I}_{t^\star} = \left[t^\star - \delta^s_{t^\star},\ t^\star + \delta^e_{t^\star}\right]$. If $t \in \mathcal{I}_{t^\star}$, the model considers the target event to be occurring at the current moment and triggers the retrieval output. Otherwise, no retrieval result is produced. By retaining historical high-confidence predictions, this mechanism provides more stable event intervals and reduces frequent decision changes caused by short-term confidence fluctuations.

3.6 Model Training

To improve cross-stream discrimination when the target event occurs, we propose the CSVC. CSVC aligns the query representation with target video segments inside the ground-truth event interval, while using segments from other video streams at the same time step as negative samples. This cross-stream contrast reduces semantic confusion across video streams and helps the query match the target stream more accurately. Given a sentence-level text representation $\mathbf{z}^q$, let $V^\star$ denote the target video stream with index $n^\star$. Let $\mathbf{f}^{v,o}_{n,i}$ denote the video representation of stream n at time i, $\mathcal{T}_n$ the set of candidate timestamps for that stream, and $\mathcal{T}_{n^\star,+} \subseteq \mathcal{T}_{n^\star}$ the set of positive timestamps within the ground-truth event interval. For temporally synchronized candidate video streams, the CSVC loss is defined as

$$\mathcal{L}_{\text{csvc}} = -\frac{1}{|\mathcal{T}_{n^\star,+}|} \sum_{i \in \mathcal{T}_{n^\star,+}} \log\left(\frac{\exp\left(\cos\left(\mathbf{z}^q, \mathbf{f}^{v,o}_{n^\star,i}\right)/\gamma\right)}{\sum_{j=1}^{N_v} \exp\left(\cos\left(\mathbf{z}^q, \mathbf{f}^{v,o}_{j,i}\right)/\gamma\right)}\right), \tag{9}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity, γ is the temperature parameter, and N_v is the total number of candidate video streams. This loss encourages the query to be closer to the target event segment while remaining distinct from irrelevant content in other streams.

We use Focal Loss (Lin et al., 2017) to supervise the query-matching confidence in SQ-AER. To improve discrimination against hard cross-stream negatives, we select the segment with the highest matching confidence from the other video streams at the same time step: $n^-_i = \arg\max_{n \neq n^\star} a_{n,i}$, $a^-_i = a_{n^-_i,i}$. The confidence supervision loss is defined as

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{T}_{n^\star}|} \sum_{i \in \mathcal{T}_{n^\star}} \left[\mathcal{L}_{\text{focal}}\left(a_{n^\star,i}, y_{n^\star,i}\right) + \mathcal{L}_{\text{focal}}\left(a^-_i, 0\right)\right]. \tag{10}$$

Here, $a_{n^\star,i}$ denotes the query-matching confidence for the target stream at time i, and a^-_i denotes the matching confidence of the hard negative sample at the same time step. The label $y_{n^\star,i}$ is set to 1 if the corresponding segment falls within the ground-truth event interval and 0 otherwise. This loss learns the occurrence state of the target event while suppressing high false-matching responses from other streams.

Event boundaries are supervised using Mean Absolute Error. For a positive sample at time i within the ground-truth event interval, the predicted start and end positions are $i - \delta^s_{n^\star,i}$ and $i + \delta^e_{n^\star,i}$,

Table 1: Comparison of STVR performance on the TVR and DiDeMo datasets . The best and second-best results in each column, including ties, are boldfaced and underlined, respectively.

Method	TVR						DiDeMo					
	IoU=0.3		IoU=0.5		IoU=0.7		IoU=0.3		IoU=0.5		IoU=0.7	
	R@10	R@100	R@10	R@100	R@10	R@100	R@10	R@100	R@10	R@100	R@10	R@100
HLFormer	3.31	6.82	1.18	2.53	<u>0.48</u>	1.28	<u>4.44</u>	7.70	0.62	1.10	0.07	0.10
MSC	0.57	2.55	0.16	0.82	0.06	0.36	3.39	7.53	0.50	1.15	0.07	<u>0.12</u>
BOA	3.37	6.78	<u>1.21</u>	2.58	0.47	<u>1.29</u>	3.96	<u>7.97</u>	0.60	<u>1.30</u>	0.07	0.10
Dream	<u>3.44</u>	<u>6.88</u>	1.19	<u>2.62</u>	<u>0.48</u>	1.25	4.21	7.90	<u>0.67</u>	1.17	0.07	0.10
Holmes	2.81	6.40	0.96	2.35	0.45	1.05	3.99	7.40	0.50	1.07	<u>0.10</u>	<u>0.12</u>
Ours	**7.83**	**17.40**	**4.41**	**11.88**	**1.47**	**5.20**	**19.54**	**31.25**	**11.64**	**21.60**	**5.86**	**11.94**

respectively. The boundary regression loss is defined as

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{T}_{n^\star,+}|} \sum_{i \in \mathcal{T}_{n^\star,+}} \frac{1}{2} \left(\left| \left(i - \delta^s_{n^\star,i}\right) - \tau^s_\star \right| + \left| \left(i + \delta^e_{n^\star,i}\right) - \tau^e_\star \right| \right), \tag{11}$$

where $\delta^s_{n^\star,i}$ and $\delta^e_{n^\star,i}$ denote the temporal distances from the current anchor i to the predicted event start and end boundaries, respectively, and $\tau^s_\star$ and $\tau^e_\star$ denote the ground-truth start and end times. This loss is applied only to positive samples and encourages the predicted event boundaries to match the ground truth. The overall training objective is $\mathcal{L} = \mathcal{L}_{\text{acs}} + \mathcal{L}_{\text{csvc}} + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{reg}}$. Additionally, the model inference procedure is provided in the Appendix.

4 Experiments

4.1 Datasets and Metrics

Datasets To evaluate the proposed method on the STVR task, we conduct experiments on three widely used benchmarks: DiDeMo (Anne Hendricks et al., 2017), TVR (Lei et al., 2020), and Charades-STA (Gao et al., 2017). DiDeMo contains about 10,600 untrimmed Flickr videos with an average duration of 30 seconds and 3.9 text queries per video. TVR contains about 21,800 video clips from six TV series, with an average duration of 76.2 seconds and five text queries per video. Charades-STA contains 6,672 videos of daily indoor activities, with an average duration of about 30 seconds and 2.4 text queries per video. For DiDeMo and Charades-STA, we use the fine-grained text annotations provided by VERIFIED (Chen et al., 2024). These annotations are used to evaluate the model's ability to understand fine-grained event semantics and distinguish different video streams.

Metrics To evaluate performance on the STVR task, we use the joint metric $\text{Recall@}K, \text{IoU} = m$, which measures both video stream retrieval and event occurrence assessment. It reports the percentage of test queries whose joint temporal Intersection-over-Union (IoU) between the online prediction and the ground-truth event interval is no lower than m under the Top-K retrieval setting. Here, $K \in \{10, 100\}$ denotes the number of candidate video streams returned at each time step, and $m \in \{0.3, 0.5, 0.7\}$ is the joint temporal IoU threshold. The joint temporal IoU is computed on a discrete time axis and considers both retrieval results and event state predictions at each time step. Outside the ground-truth event interval, any retrieval output is treated as a false positive and contributes to the union. Within the ground-truth event interval, if the model does not determine that the event is occurring and produces no output, the time step is treated as a miss and contributes only to the union. A time step contributes to the intersection only when the model determines that the event is occurring, outputs retrieval results, and the target video stream is included in the current Top-K candidates.

4.2 Comparison With State-of-the-Art Methods

To evaluate the proposed method on the STVR task, we adapt several representative PRVR methods, including HLFormer (Li et al., 2025), MSC (Moon et al., 2026), BOA (Li et al., 2026a), Dream (Li et al., 2026c), and Holmes (Li et al., 2026b), to the STVR setting. For a fair comparison, all methods use their official implementations and the same video and text features. We also use the same STVR evaluation metrics, apply causal masks to modules that exchange information across time, and introduce threshold-based event state prediction to determine whether the queried event is occurring.

Tables 1 and 2 report the results on TVR, DiDeMo, and Charades-STA. HMER achieved the best results across all 18 evaluation settings and showed consistent advantages under different Top-K settings and joint temporal IoU thresholds. On TVR, when $K = 100$, HMER achieved recall rates of 11.88% and 5.20% at $m = 0.5$ and $m = 0.7$, respectively. These results exceeded the strongest baselines by 9.26 and 3.91 percentage points. On DiDeMo, at $m = 0.5$, HMER achieved Recall@10 of 11.64% and Recall@100 of 21.60%, improving over the strongest baselines by 10.97 and 20.30 percentage points. When the threshold increased to $m = 0.7$, HMER still achieved a Recall@100 of 11.94%, exceeding the strongest baseline by 11.82 percentage points.

On Charades-STA, the performance of the compared methods dropped sharply as the IoU threshold increased. With $K = 100$ and $m = 0.5$, the strongest baseline achieved a recall of only 0.22%, while HMER reached 13.49%, an improvement of 13.27 percentage points. At $m = 0.7$, all compared methods obtained 0.00% for both Recall@10 and Recall@100, whereas HMER achieved 0.89% and 5.81%, respectively. These results demonstrate the effectiveness of HMER for joint streaming retrieval and event occurrence assessment. They also show that STVR becomes more challenging under stricter temporal overlap thresholds. Since the joint metric considers target-stream hits, missed detections within the event interval, and false positives outside it, the gains of HMER support the joint modeling of video representations and query-aware event states in streaming scenarios.

Table 2: STVR performance comparison on Charades-STA.

Method	IoU=0.3		IoU=0.5		IoU=0.7	
	R@10	R@100	R@10	R@100	R@10	R@100
HLFormer	2.80	14.11	0.11	0.22	0.00	0.00
MSC	0.16	1.34	0.05	0.11	0.00	0.00
BOA	2.28	14.76	0.03	0.03	0.00	0.00
Dream	3.47	14.95	0.05	0.22	0.00	0.00
Holmes	2.88	13.41	0.08	0.22	0.00	0.00
Ours	**5.78**	**21.91**	**2.63**	**13.49**	**0.89**	**5.81**

4.3 Ablation Study

To analyze the contribution of each component, we conduct ablation experiments on the DiDeMo dataset, as shown in Table 3. The "Base" model removes RVG-EM, ACS, and SQ-AER. Adding SQ-AER improves all six metrics. At IoU=0.3 and IoU=0.5, R@100 increases from 9.00% and 1.32% to 25.09% and 12.98%, respectively, showing the benefit of jointly modeling video stream relevance and target event occurrence. Adding RVG-EM to Base+SQ-AER further improves performance. At IoU=0.3, R@10 and R@100 increase from 13.53% and 25.09% to 18.22% and 31.02%, while at IoU=0.7 they increase from 1.32% and 4.44% to 2.87% and 6.68%. This shows that radial valley-based event partitioning helps reduce cross-event interference. Combining RVG-EM and ACS with SQ-AER yields the best results across all six metrics. At IoU=0.3, R@10 and R@100 reach 19.54% and 31.25%, respectively, while at IoU=0.5 they achieve 11.64% and 21.60%. At the stricter IoU=0.7 threshold, R@10 and R@100 still reach 5.86% and 11.94%, respectively. These results show that the three components are complementary and jointly improve STVR performance, especially under stricter IoU thresholds. **Additionally, more experimental analyses are provided in the Appendix**.

Table 3: Ablation study on the DiDeMo dataset. The best results in each column are boldfaced.

Base	SQ-AER	RVG-EM	ACS	DiDeMo					
				IoU=0.3		IoU=0.5		IoU=0.7	
				R@10	R@100	R@10	R@100	R@10	R@100
✓				3.76	9.00	0.55	1.32	0.07	0.10
✓	✓			13.53	25.09	4.96	12.98	1.32	4.44
✓	✓	✓		18.22	31.02	7.25	14.31	2.87	6.68
✓	✓	✓	✓	**19.54**	**31.25**	**11.64**	**21.60**	**5.86**	**11.94**

5 CONCLUSION

This paper studies STVR, which extends offline retrieval on complete videos to online retrieval over continuously arriving video streams. STVR requires the model to identify relevant video streams and determine whether the target event is occurring using only current and historical observations. To address this task, we propose HMER, a hyperbolic memory-based event reasoning framework. HMER builds event-aware visual context from historical observations and combines query-aware reasoning with historical high-confidence predictions to determine when retrieval results should be output. Experiments show that HMER consistently improves STVR performance, demonstrating its effectiveness for cross-modal retrieval under incomplete observations.

AI USE STATEMENT

In this work, we used large language models to assist with identifying and correcting errors in the manuscript and code, as well as with translation, literature search, reference checking, and improving the clarity, grammar, and readability of the manuscript. All AI-assisted content was carefully reviewed by the authors. Specifically, we manually checked the revised text, mathematical expressions, and references, and verified all suggested code modifications for correctness and consistency with the proposed method. The authors take full responsibility for the final content of this work.

REFERENCES

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017.

Shehreen Azad, Vibhav Vineet, and Yogesh S Rawat. Streamready: Learning what to answer and when in long streaming videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 2026.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

Ines Chami, Albert Gu, Dat P Nguyen, and Christopher Ré. Horopca: Hyperbolic dimensionality reduction via horospherical projections. In *International Conference on Machine Learning*, pp. 1419–1429. PMLR, 2021.

Houlun Chen, Xin Wang, Hong Chen, Zeyang Zhang, Wei Feng, Bin Huang, Jia Jia, and Wenwu Zhu. Verified: A video corpus moment retrieval benchmark for fine-grained video understanding. *Advances in Neural Information Processing Systems*, 37:40393–40406, 2024.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5672–5686, 2022.

Xianke Chen, Daizong Liu, Xun Yang, Xirong Li, Jianfeng Dong, Meng Wang, and Xun Wang. Prvr: Partially relevant video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, ZhongTao, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 42115–42127, 2025.

Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 246–257, 2022.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6342–6352. IEEE, 2025.

Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. Uatvr: Uncertainty-adaptive text-video retrieval. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13677–13687. IEEE, 2023.

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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285. IEEE, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Chan Hur, Jeong-hun Hong, Dong-hun Lee, Dabin Kang, Semin Myeong, Sang-hyo Park, and Hyeyoung Park. Narrating the video: Boosting text-video retrieval via comprehensive utilization of frame-level captions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24077–24086. IEEE, 2025.

Hongxiang Jiang, Zengrui Ge, Guo Chen, Qixiong Wang, Jile Jiao, Xuetao Feng, Yuan Wang, and Yan Wang. Memento: Toward an all-day proactive assistant for ultra-long streaming video. In *International Conference on Learning Representations*, volume 2026, pp. 35267–35281, 2026.

Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, Xueqing Deng, Peng Wang, Zhao Zhang, Xiaohui Shen, and Jiashi Feng. Mv-adapter: Multimodal video transfer learning for video text retrieval. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27134–27143. IEEE, 2024.

Minjoo Ki, Daejung Kim, Kisung Kim, Seon Joo Kim, and Jinhan Lee. Carim: Caption-based autonomous driving scene retrieval via inclusive text matching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22036–22045. IEEE, 2025.

Junho Kim, Hosu Lee, James M. Rehg, Minsu Kim, and Yong Man Ro. Stride: When to speak meets sequence denoising for streaming video understanding. In *European Conference on Computer Vision*, pp. 38–56. Springer, 2026.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 82(3):036106, 2010.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Eugene Yang, and Benjamin Van Durme. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24149–24158. IEEE, 2025.

Bangxiang Lan, Ruobing Xie, Ruixiang Zhao, Xingwu Sun, Zhanhui Kang, Gang Yang, and Xirong Li. Hybrid-tower: Fine-grained pseudo-query interaction and generation for text-to-video retrieval. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24497–24506. IEEE, 2025.

Vu-Minh Le, Thao-Anh Tran, Duc Huy Do, Xuan Canh Do, Huong Ninh, and Hai Tran. Online 3d multi-camera perception through robust 2d tracking and depth-based late aggregation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 5449–5459. IEEE, 2025.

Junwoon Lee and Yulun Tian. Latentam: Real-time, large-scale latent gaussian attention mapping via online dictionary learning. *IEEE Robotics and Automation Letters*, 2026.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pp. 447–463. Springer, 2020.

Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. *Advances in Neural Information Processing Systems*, 36:24564–24585, 2023.

Huafeng Li, Jialong Zhao, Yafei Zhang, and Jie Wen. Bidirectional cross-modal collaborative alignment via semantic-guided visual embeddings for partially relevant video retrieval. *IEEE Transactions on Image Processing*, 2026a.

Jun Li, Jinpeng Wang, Chaolei Tan, Niu Lian, Long Chen, Yaowei Wang, Min Zhang, Shu-Tao Xia, and Bin Chen. Enhancing partially relevant video retrieval with hyperbolic learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23074–23084. IEEE, 2025.

Jun Li, Peifeng Lai, Xuhang Lou, Jinpeng Wang, Yuting Wang, Ke Chen, Yaowei Wang, and Shu-Tao Xia. Revisiting uncertainty: On evidential learning for partially relevant video retrieval. In *Forty-third International Conference on Machine Learning*, 2026b.

Jun Li, Xuhang Lou, Jinpeng Wang, Yuting Wang, Yaowei Wang, Shu-Tao Xia, and Bin Chen. Imagine before concentration: Diffusion-guided registers enhance partially relevant video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9710–9721, June 2026c.

Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. Oasis: On-demand hierarchical event memory for streaming video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, pp. 319–335. Springer, 2022.

WonJun Moon, MinSeok Jung, Gilhan Park, Tae-Young Kim, Cheol-Ho Cho, Woojin Jun, and Jae-Pil Heo. Mitigating semantic collapse in partially relevant video retrieval. *Advances in Neural Information Processing Systems*, 38:23196–23217, 2026.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pp. 3779–3788. PMLR, 2018.

Avik Pal, Max Van Spengler, Guido D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 87371–87399, 2025.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055. IEEE, 2025.

Peipei Song, Long Zhang, Long Lan, Weidong Chen, Dan Guo, Xun Yang, and Meng Wang. Towards efficient partially relevant video retrieval with active moment discovering. *IEEE Transactions on Multimedia*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Mengzhao Wang, Huafeng Li, Yafei Zhang, Jinxing Li, Dapeng Tao, and Zhengtao Yu. Disentangling inter- and intra-video relations for multi-event video-text retrieval and grounding. *IEEE Transactions on Image Processing*, 34:7558–7571, 2025a.

Mengzhao Wang, Huafeng Li, Yafei Zhang, Jinxing Li, Minghong Xie, and Dapeng Tao. Dual-task mutual reinforcing embedded joint video paragraph retrieval and grounding. *IEEE Transactions on Multimedia*, 27:5879–5894, 2025b.

Yuting Wang, Jinpeng Wang, Bin Chen, Ziyun Zeng, and Shu-Tao Xia. Gmmformer: Gaussian-mixture-model based transformer for efficient partially relevant video retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 5767–5775, 2024.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2804–2815. IEEE, 2023.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10704–10713. IEEE, 2023.

Yi Yang, Zheng Wang, Xing Xu, Jingkuan Song, and Heng Tao Shen. Gravitation-driven semantic alignment for text video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14946–14956, 2026.

Wei Chee Yew, Hailun Xu, Sanjay Saha, Xiaotian Fan, Hiok Hian Ong, David Yuchen Wang, Kanchan Sarkar, Zhenheng Yang, and Danhui Guan. Dynamic content moderation in livestreams: Combining supervised classification with mllm-boosted similarity matching. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2528–2538, 2026.

Xiaoyan Yu, Mengzhao Wang, Huafeng Li, Yafei Zhang, and Dapeng Tao. Dual-branch collaborative implicit-explicit mutual learning for weakly supervised video-sentence retrieval and grounding. *IEEE Transactions on Multimedia*, 2026.

Donglin Zhang, Zheng Rao, Xing Xu, Xiao-Jun Wu, and Josef Kittler. Probabilistic embeddings with evidence learning and refinement for text-video retrieval. *IEEE Transactions on Image Processing*, 2026a.

Kairui Zhang, Zhenyu Yang, Bing Wang, Shengsheng Qian, and Changsheng Xu. Querystream: Advancing streaming video understanding with query-aware pruning and proactive response. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Yulin Zhang, Cheng Shi, and Sibei Yang. Weavetime: Streaming from earlier frames into emergent memory in videollms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16920–16932, June 2026c.

Jialong Zhao, Huafeng Li, Yafei Zhang, Wen Wang, and Changchun Hua. Training-free video corpus moment retrieval via synergistic collaboration and adaptive calibration. *IEEE Transactions on Image Processing*, 2026.

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

A Appendix

A.1 Definition of Hyperbolic Space

Hyperbolic space $\mathbb{H}^m$ is an m-dimensional Riemannian manifold with constant negative sectional curvature and admits several isometric models. We adopt the Lorentz model for its numerical stability in hyperbolic representation learning (Nickel & Kiela, 2018) and, following (Chen et al., 2022), fix the curvature to -1. Formally, the Lorentz model is denoted by $\mathbb{L}^m = (\mathcal{L}^m, \mathfrak{g})$, where $\mathcal{L}^m$ is a hyperboloid embedded in the $(m+1)$-dimensional Minkowski space, and $\mathfrak{g}$ is the Riemannian metric induced by restricting the ambient Lorentzian metric to each tangent space.

The metric tensor of the Minkowski space is $\mathrm{diag}(-1, 1, \ldots, 1)$. Accordingly, the Lorentz inner product is defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}} := -x_0 y_0 + \boldsymbol{x}_s^\top \boldsymbol{y}_s, \tag{12}$$

where $x_0, y_0 \in \mathbb{R}$ denote the time-like components, and $\boldsymbol{x}_s, \boldsymbol{y}_s \in \mathbb{R}^m$ denote the spatial components. The hyperboloid $\mathcal{L}^m$ is defined as

$$\mathcal{L}^m := \left\{ \boldsymbol{x} \in \mathbb{R}^{m+1} \;\middle|\; \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = -1,\ x_0 > 0 \right\}. \tag{13}$$

Thus, $\mathcal{L}^m$ corresponds to the upper sheet of a two-sheeted hyperboloid, and any point on $\mathcal{L}^m$ can be expressed as

$$\boldsymbol{x} = (x_0, \boldsymbol{x}_s^\top)^\top \in \mathbb{R}^{m+1}, \qquad x_0 = \sqrt{1 + \|\boldsymbol{x}_s\|_2^2}. \tag{14}$$

We choose $\boldsymbol{o} = (1, 0, \ldots, 0)^\top$ as the reference origin for subsequent mappings and radial distance computations.

For a point $\boldsymbol{x} \in \mathcal{L}^m$, the tangent space is defined as $T_{\boldsymbol{x}}\mathbb{L}^m = \{\boldsymbol{v} \in \mathbb{R}^{m+1} \mid \langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathcal{L}} = 0\}$. The exponential and logarithmic maps provide transformations between this tangent space and the hyperboloid. Specifically, the exponential map sends a tangent vector $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathbb{L}^m$ to a point on the hyperboloid and is given by

$$\exp_{\boldsymbol{x}}(\boldsymbol{v}) = \cosh(\|\boldsymbol{v}\|_{\mathcal{L}})\boldsymbol{x} + \sinh(\|\boldsymbol{v}\|_{\mathcal{L}}) \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_{\mathcal{L}}}, \tag{15}$$

where $\|\boldsymbol{v}\|_{\mathcal{L}} = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle_{\mathcal{L}}}$ denotes the norm induced on the tangent space. Conversely, the logarithmic map sends a point $\boldsymbol{y} \in \mathcal{L}^m$ to a tangent vector in $T_{\boldsymbol{x}}\mathbb{L}^m$ and is given by

$$\log_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{\mathrm{arcosh}(-\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}})}{\sqrt{\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}}^2 - 1}} \left(\boldsymbol{y} + \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}} \boldsymbol{x} \right). \tag{16}$$

These two maps are inverses of each other, with $\exp_{\boldsymbol{x}}(\boldsymbol{0}) = \boldsymbol{x}$ and $\log_{\boldsymbol{x}}(\boldsymbol{x}) = \boldsymbol{0}$ defined by continuity.

A.2 Model Inference

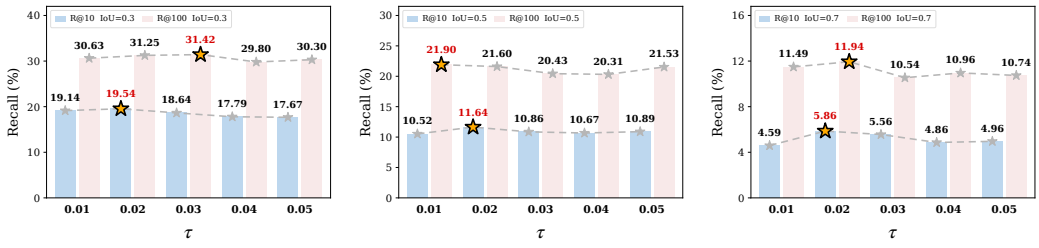
For each candidate video stream, the model obtains a memory-augmented representation through RVG-EM and computes its cosine similarity with the sentence-level query representation. These similarities are used to rank the video streams at the current time step. SQ-AER predicts query-matching confidence and event boundaries for the top-K candidates. The model selects the prediction with the highest query-matching confidence among these candidates and compares its confidence with that stored in the historical best state. This state, including its temporal anchor and boundary offsets, is updated only when the current prediction has higher confidence. If the current time falls within the event interval predicted by this state, the model outputs the top-K streams from the current retrieval ranking. Otherwise, it outputs an empty set.

A.3 Implementation Details

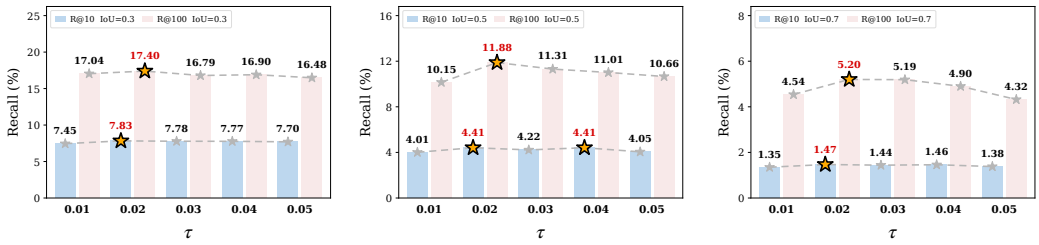
For all datasets, we use ResNet-152 (He et al., 2016) to extract video features. RoBERTa (Liu et al., 2019) is used to encode text queries. To maintain temporally synchronized multi-stream inputs, shorter videos are cyclically repeated to match the length of the longest video stream. For each query, evaluation is limited to the original duration of its target video. The model is trained with the BertAdam optimizer using a learning rate of 2.5×10^{-4} and a batch size of 128 for 100 epochs. The temperature parameter γ is set to 0.1, and the radial variation threshold τ is set to 0.02.

A.4 Analysis of the Radial Threshold τ

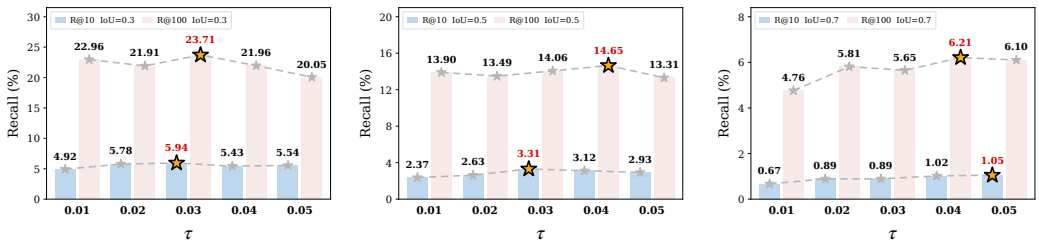
The parameter τ in Eq. 2 controls the sensitivity of radial valley detection. A smaller τ allows subtle radial changes to be detected, while a larger τ requires more pronounced decreases and increases. To study its effect, we evaluate $\tau \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$ on DiDeMo, TVR, and Charades-STA, as shown in Figure 4. On DiDeMo, $\tau = 0.02$ achieves the best results on four of the six metrics. In particular, at $\mathrm{IoU} = 0.7$, R@10 and R@100 reach 5.86% and 11.94%, respectively.



(a) DiDeMo



(b) TVR



(c) Charades-STA

Figure 4: Effect of the radial valley detection threshold τ on STVR performance on (a) DiDeMo, (b) TVR, and (c) Charades-STA. Within each subfigure, the three panels correspond to IoU $= 0.3$, 0.5, and 0.7, from left to right. Blue and pink bars denote R@10 and R@100, respectively. Red values and yellow stars indicate the best results for each metric at each IoU threshold, including ties.

On TVR, $\tau = 0.02$ also achieves the best or tied-best results across all six metrics. In contrast, Charades-STA favors larger thresholds, with the best results obtained at $\tau = 0.03$, 0.04, or 0.05, depending on the metric. These results suggest that stricter valley detection does not consistently improve retrieval performance. Given its strong performance on DiDeMo and TVR, we retain $\tau = 0.02$ as the default setting.

A.5 Analysis of Memory and Prediction Strategies

To further examine the design choices in RVG-EM and SQ-AER, we compare different memory and prediction strategies on DiDeMo, as shown in Table 4. "No memory" removes historical context aggregation from RVG-EM and uses only the current segment. "Long-term memory" aggregates all segments observed up to the current time without radial valley-based memory resets. "Per-step prediction" keeps RVG-EM unchanged but uses the highest-confidence prediction among the current Top-K candidates at each time step, instead of retaining the highest-confidence prediction observed so far. The remaining components are kept unchanged for each variant.

Long-term memory outperforms No memory across all six metrics, showing the benefit of using historical context to supplement incomplete visual evidence. However, aggregating the entire observed history remains less effective than the event-aware memory used by HMER. At IoU=0.7, R@100

Table 4: Comparison of memory and prediction strategies on DiDeMo. All results are reported in percentages. The best results in each column are boldfaced.

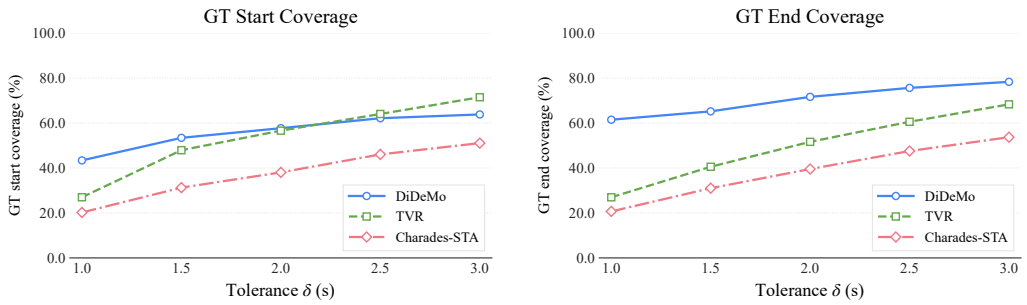
Method	IoU=0.3		IoU=0.5		IoU=0.7	
	R@10	R@100	R@10	R@100	R@10	R@100
No memory	14.48	26.29	6.15	16.27	1.00	3.59
Long-term memory	17.27	29.88	8.70	17.70	2.84	7.36
Per-step prediction	14.18	25.99	5.68	13.78	1.97	6.20
HMER (Ours)	**19.54**	**31.25**	**11.64**	**21.60**	**5.86**	**11.94**

increases from 3.59% with No memory to 7.36% with Long-term memory, and further to 11.94% with HMER. These results support the use of radial valleys to restrict context aggregation to the current event, preserving useful historical information while reducing cross-event interference.

Replacing historical prediction retention with Per-step prediction also reduces performance across all six metrics. At IoU=0.7, R@10 and R@100 decrease from 5.86% and 11.94% to 1.97% and 6.20%, respectively. This shows that retaining historical high-confidence predictions is more effective than relying only on current-time predictions for online event occurrence assessment. The results are consistent with the motivation of SQ-AER to reduce the influence of short-term prediction fluctuations by using previously established high-confidence event intervals.

A.6 Analysis of Radial Valleys near Event Boundaries

To examine whether radial valleys occur near event boundaries, we evaluate their coverage of ground-truth start and end boundaries on DiDeMo, TVR, and Charades-STA. A boundary is considered covered if at least one detected valley occurs within $\pm\delta$ seconds of it, where $\delta \in \{1, 1.5, 2, 2.5, 3\}$. We report the percentage of covered start and end boundaries separately across all test queries, including those from videos without detected valleys.

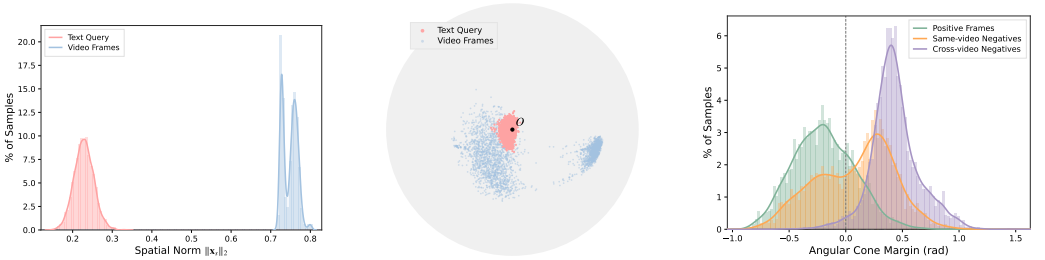


(a) GT start coverage

(b) GT end coverage

Figure 5: Coverage of ground-truth event boundaries by detected radial valleys on DiDeMo, TVR, and Charades-STA. (a) Start-boundary coverage. (b) End-boundary coverage. The horizontal axis denotes the tolerance δ in seconds. The vertical axis shows the percentage of boundaries with at least one detected valley within $\pm\delta$ seconds.

Figure 5 shows that coverage increases with the tolerance on all three datasets. DiDeMo consistently achieves the highest end-boundary coverage, while TVR shows the highest start-boundary coverage at larger tolerances. At $\delta = 3$ seconds, start- and end-boundary coverage both exceed 50% on all three datasets. These results suggest that radial valleys can provide useful cues near annotated event boundaries by capturing significant semantic transitions in the video stream. Although radial valleys do not exactly correspond to all event boundaries, their consistent occurrence around the start and end boundaries indicates that radial variations in hyperbolic space contain informative temporal signals for event transition detection.



(a) Spatial norm distribution (b) 2D HoroPCA projection (c) Angular cone margin distribution

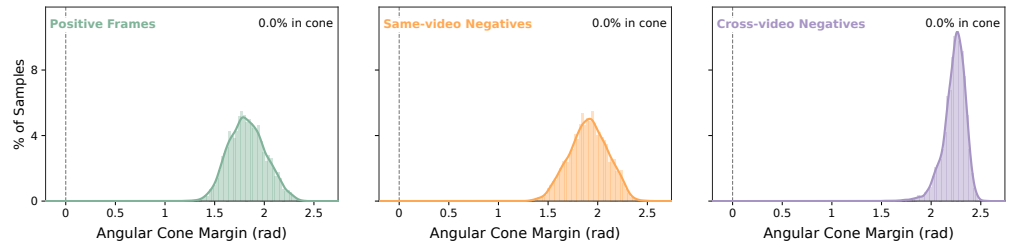
Figure 6: Geometric analysis of the learned hyperbolic representations. (a) Spatial norm distributions of text queries and video representations, where the horizontal axis denotes $\|\boldsymbol{x}_s\|_2$. (b) Two-dimensional HoroPCA projection of the same representations, where O denotes the projected location of the original hyperbolic origin. (c) Angular cone margin distributions for three types of query–video pairs, where the horizontal axis denotes the angular cone margin m in radians and the dashed line corresponds to $m = 0$. The vertical axes in (a) and (c) show the percentage of samples falling into each interval.

A.7 Visualization of the Learned Hyperbolic Space

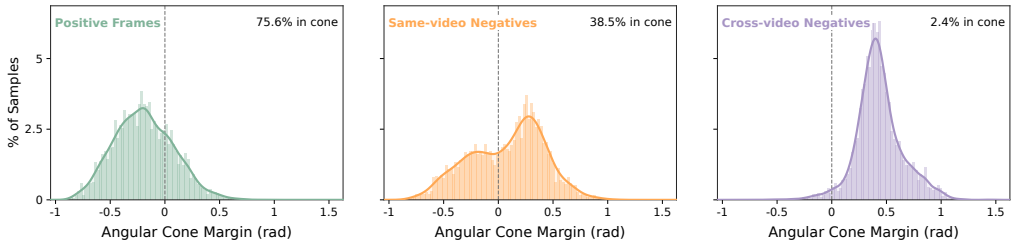
To examine the cross-modal geometry learned by the model, we visualize the representations from three aspects: radial distribution, 2D projection, and angular cone margin distribution. Specifically, we extract the sentence-level query representations $\bar{\mathbf{z}}^q$ and event-augmented video representations $\hat{\mathbf{f}}^v_{n,i}$. We compute the spatial norms of these representations in Lorentz space and project the representations into 2D using HoroPCA (Chami et al., 2021). Equal numbers of text and video representations are sampled to fit HoroPCA, after which all visualization samples are projected. We also compute the angular cone margin $m = \phi(\bar{\mathbf{z}}^q, \hat{\mathbf{f}}^v_{n,i}) - \omega(\bar{\mathbf{z}}^q)$ for positive samples, intra-video negative samples, and inter-video negative samples to examine whether the learned geometry follows the ACS objective.

Figure 6(a) shows the spatial norm distributions of text queries and video representations. Text queries generally have smaller spatial norms than video representations, forming a clear radial ordering between the two modalities. This is consistent with our hierarchical modeling view, in which sentence-level queries capture abstract event semantics and video segments provide concrete visual evidence, forming a general–specific hierarchical correspondence. Figure 6(b) shows the 2D HoroPCA projection of the same representations. Query representations are relatively compact, while video representations are more widely distributed and form several clusters. Their different distribution patterns further illustrate the cross-modal structure learned in hyperbolic space.

We further analyze the query–video relation using the angular cone margin. As shown in Figure 6(c), positive samples are mainly distributed to the left of the zero line, while inter-video negative samples are mostly located to the right. Intra-video negative samples appear on both sides. This separation is consistent with the ACS objective, which encourages positive samples to lie inside the query cone and negative samples to remain outside it. These results provide further evidence that ACS learns meaningful geometric relations between queries and video segments. To further examine the effect of ACS, Figure 7 compares the angular cone margin distributions of models trained with and without ACS. Without ACS, all three distributions are concentrated above zero, including that of positive pairs. With ACS, the proportion of positive pairs inside the query cones increases from 0.0% to 75.6%, while only 2.4% of inter-video negative pairs lie inside. This 73.2-percentage-point gap demonstrates clear geometric discrimination between positive and inter-video negative pairs under the cone criterion. Intra-video negatives are more challenging, with 38.5% lying inside the cones. This overlap may partly reflect shared, query-related visual content outside the annotated interval, which can remain semantically compatible with the query despite being temporally misaligned. Overall, ACS promotes the intended cone inclusion for positive pairs while preserving strong exclusion of inter-video negatives. The remaining intra-video overlap highlights the complementary roles of ACS in organizing semantic relations and SQ-AER in determining whether the queried event is currently occurring.



(a) Without ACS



(b) With ACS

Figure 7: Effect of ACS on angular cone margin distributions. (a) Without ACS. (b) With ACS. From left to right, each row shows positive pairs, intra-video negative pairs, and inter-video negative pairs. Histograms show the percentage of samples in each interval. The dashed line marks $m = 0$, and the annotated percentages indicate the proportions of pairs inside the query cones ($m \leq 0$).

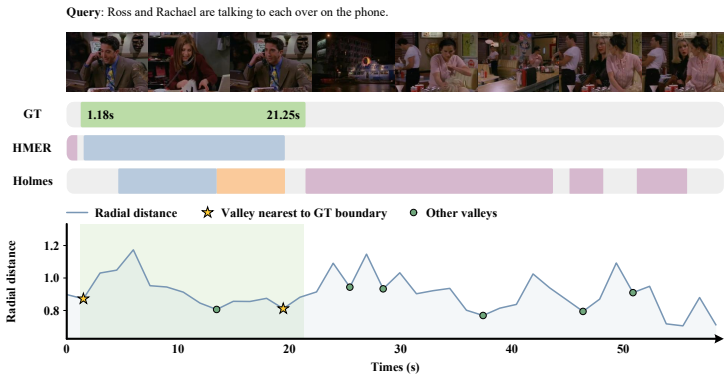
A.8 Qualitative Results

Figure 8 compares the qualitative STVR results of HMER and Holmes on TVR under the Top-10 setting. Taking Figure 8(c) as an example, the query is "Sky, Castle and Beckett were having a conversation." The video shows the conversation through alternating shots of different characters. Thus, the model needs to continuously identify query-relevant content as the visual observations change. HMER correctly retrieves the target video stream during most of the event interval. In contrast, although Holmes continues to output retrieval results, the target stream is often absent from its Top-10 results. After the conversation ends, Holmes also produces several false outputs outside the ground-truth interval, while HMER shows a shorter period of false triggering. These results illustrate the advantage of HMER in both target stream retrieval and output timing. However, HMER still exhibits short-term retrieval errors and occasional false outputs after the event.

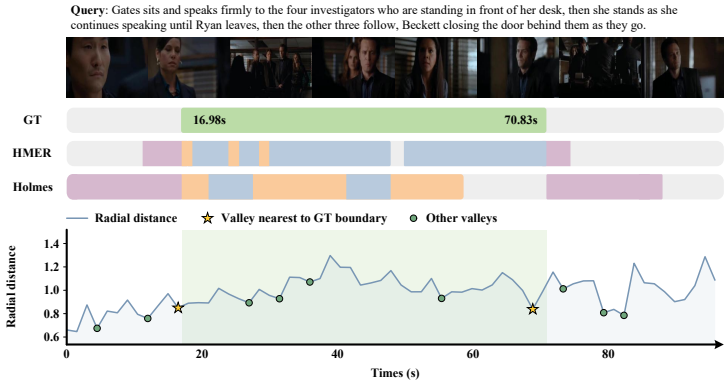
We further visualize the radial-distance curves to examine the temporal cues used by RVG-EM. The available annotations specify query-relevant event intervals but do not exhaustively label all event transitions within each video. We therefore compare detected valleys with the annotated start and end boundaries. In all three examples, valleys occur near both boundaries, illustrating the temporal correspondence between radial reversals and the query-relevant event intervals. Together with the dataset-level coverage results in Figure 5, these examples provide complementary quantitative and qualitative evidence for using radial valleys to guide online event-memory updates and restrict context aggregation across estimated event transitions.

B Limitations and Future Work

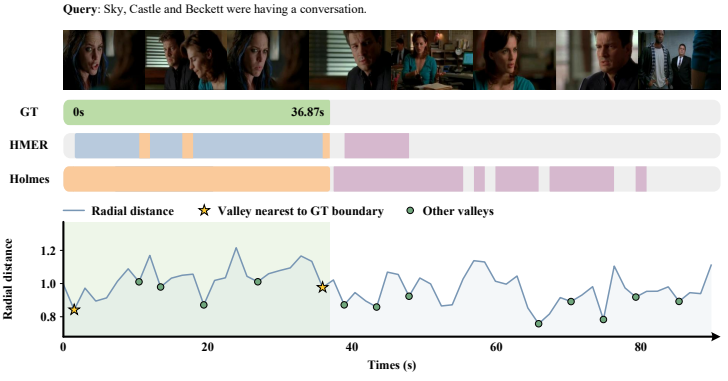
The current model considers only visual and textual information, while useful cues from other modalities, such as audio and subtitles, remain unexplored. Future work will investigate adaptive sampling and multimodal information fusion to improve processing efficiency and extend the framework to more streaming scenarios.



(a)



(b)



(c)

Figure 8: Qualitative results on TVR. The GT row shows the ground-truth event interval in green. The HMER and Holmes rows show the retrieval outputs of our method and the baseline, respectively. Blue denotes correct outputs within the event interval, where the target video stream is included in the Top-10 results. Orange denotes outputs within the event interval whose Top-10 results do not include the target video stream. Purple denotes outputs triggered outside the event interval. The lower plot in each example shows the radial distance over time, with light-green shading indicating the ground-truth interval. Yellow stars highlight the detected valleys nearest in time to the two ground-truth boundaries, while green circles denote the remaining detected valleys.