UMVR: A UNIFIED AND EFFICIENT MLLM FRAMEWORK FOR SHORT- AND LONG-VIDEO RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-based video retrieval has largely been studied under separate paradigms for short and long videos, despite sharing the same objective of measuring text-video relevance. A key obstacle to unifying these paradigms is that representation granularity is typically *predefined* for each retrieval setting rather than adapted to the temporal structure of individual videos. To address this limitation, we propose Length-Adaptive Event Representation (LAER), which *adaptively* presents each video into temporally coherent events whose number and temporal extent adapt to its content. This event-centric representation supports both localized and holistic relevance modeling, providing a common basis for short- and long-video retrieval. Building on LAER, we introduce UMVR, a unified MLLM-based retrieval framework that uses a shared backbone and scoring formulation across both settings. However, encoding videos with an MLLM incurs substantial computational overhead as the number of visual tokens grows, particularly for long videos. To improve encoding efficiency, we further introduce Event-Aware Token Reduction (EATR), which uses the event structure provided by LAER to allocate visual tokens according to event informativeness, retaining more tokens for informative events while compressing redundant ones more aggressively. Extensive experiments across seven short- and long-video retrieval benchmarks demonstrate the effectiveness and efficiency of UMVR, which maintains strong retrieval performance while reducing the visual-token load for MLLM encoding by 90%.

1 INTRODUCTION

Text-based video retrieval has advanced rapidly with the development of vision-language representation learning, enabling effective semantic matching between natural language queries and videos. However, existing research has largely evolved along two separate directions for short and long videos. For short videos, textual queries are typically associated with the overall video content, and existing methods therefore mainly learn global video representations for video-level matching (Jing et al., 2025; Wang et al., 2023; 2024a; Luo et al., 2022; Ma et al., 2022). In contrast, long videos often contain multiple scenes and activities, while only a limited temporal segment may be relevant to a given query. A single global representation may fail to preserve such localized relevant content, motivating long-video retrieval methods to retain finer-grained temporal representations and match queries against different temporal regions (Dong et al., 2025; Zhang et al., 2025a; Li et al., 2025). As a result, short- and long-video retrieval have developed different representation and matching strategies, with the former emphasizing global matching and the latter localized matching. Although these two settings differ in temporal scale and relevance distribution, they share the same underlying objective of estimating the relevance between a textual query and a candidate video. This common objective raises a natural question: *can short- and long-video retrieval be handled by a single model rather than by separate retrieval paradigms?*

A key challenge in unifying short- and long-video retrieval is accommodating their different representation requirements within a single model. Existing methods typically tailor the number and temporal scope of representation units to a particular retrieval setting. Short-video retrieval methods often encode each video as a single global representation to capture its overall semantics, whereas long-video retrieval methods typically use multiple temporally localized representations to retain finer-grained temporal information. However, temporal structure varies across individual videos, and a granularity prescribed for one setting may not suit this variation. A unified retriever there-

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.

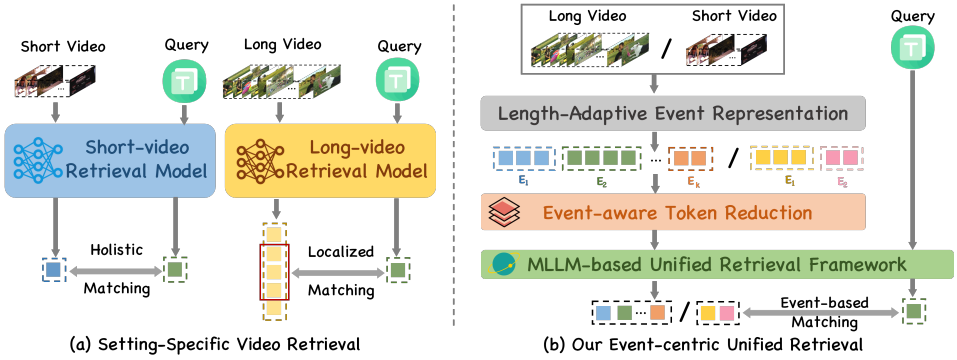

Figure 1: Comparison between setting-specific and event-centric unified paradigms for short- and long-video retrieval. (a) Existing approaches typically employ separate retrieval models for short and long videos, with different representation and matching strategies tailored to their respective settings. (b) UMVR organizes videos into adaptive events, performs Event-Aware Token Reduction before MLLM encoding, and models query-video relevance within a shared retrieval framework, enabling the same model to handle videos with different temporal structures.

fore needs to adapt its representation granularity to each video while supporting both holistic and localized relevance modeling. This motivates us to view events as a natural representation unit for unifying short- and long-video retrieval. Unlike representation units whose granularity is predefined for a particular retrieval setting, temporally coherent events provide a flexible way to organize videos according to their individual temporal structures. Such an event-centric representation can accommodate both holistic and localized video semantics without requiring separate representation schemes for short- and long-video retrieval.

Motivated by this observation, we propose Length-Adaptive Event Representation (LAER), which organizes each video into a variable-length sequence of temporally coherent events. Both the number and temporal extent of these events adapt to the video's temporal structure, allowing simpler videos to be represented by fewer coarse events and more complex videos by finer-grained events. Building on LAER, we develop UMVR, a unified retrieval framework with a shared MLLM backbone, following recent advances in MLLM-based multimodal embeddings (Meng et al., 2025; Lan et al., 2026; Zhang et al., 2024). The encoded event representations are used for localized matching and aggregated for holistic matching, with the two similarities combined into a unified relevance score. This formulation enables the same model to handle short and long videos without separate representation or matching schemes.

Additionally, encoding videos with an MLLM also raises an efficiency concern, particularly for long videos, where processing more frames substantially increases the number of visual tokens. Existing token-reduction methods can alleviate this cost, but typically apply a fixed token budget or compression ratio across the video (Fan et al., 2026; Huang et al., 2025), overlooking variations in visual informativeness across temporal regions. We therefore further introduce Event-Aware Token Reduction (EATR), which adapts the visual-token budget across events according to their informativeness, preserving more tokens for informative events while compressing redundant ones more aggressively.

Overall, our main contributions are summarized as follows:

- We revisit text-based video retrieval from a unified perspective and propose UMVR, a unified MLLM-based framework for both short- and long-video retrieval. Its Length-Adaptive Event Representation (LAER) represent videos into variable-length sequences of temporally coherent events, enabling the same representation and scoring formulation to model both localized and holistic relevance across videos with different temporal structures.
- We develop Event-Aware Token Reduction for efficient MLLM video encoding. It adapts the visual-token budget across events according to their informativeness, retaining more visual tokens for informative events while reducing redundant content more aggressively.
- Extensive experiments on three long-video and four short-video retrieval benchmarks demonstrate the effectiveness and efficiency of UMVR across diverse video lengths and relevance patterns. UMVR maintains strong retrieval performance while retaining only 10% of the visual tokens for MLLM encoding.

2 METHOD

2.1 OVERVIEW

Given a text query q and a candidate video V, video retrieval aims to learn a relevance function $S(q, V)$ that assigns higher scores to matched query-video pairs than to mismatched ones. UMVR applies the same architecture and scoring formulation to both short- and long-video retrieval without introducing task-specific retrieval heads. As illustrated in Fig. 2, UMVR follows an event-centric formulation centered on Length-Adaptive Event Representation (LAER). LAER adaptively organizes each video into K temporally coherent events according to its temporal structure, where both K and the temporal extent of individual events can vary across videos. These events provide the shared representation units for localized and holistic relevance modeling. To efficiently encode the resulting event-structured videos with an MLLM backbone, Event-Aware Token Reduction (EATR) further adjusts the visual-token budget according to the informativeness of different events before MLLM encoding. The compressed events are then jointly encoded by the MLLM to obtain contextualized event representations, which are used directly for localized matching and further aggregated for holistic matching. The two similarities are combined into the final relevance score $S(q, V)$.

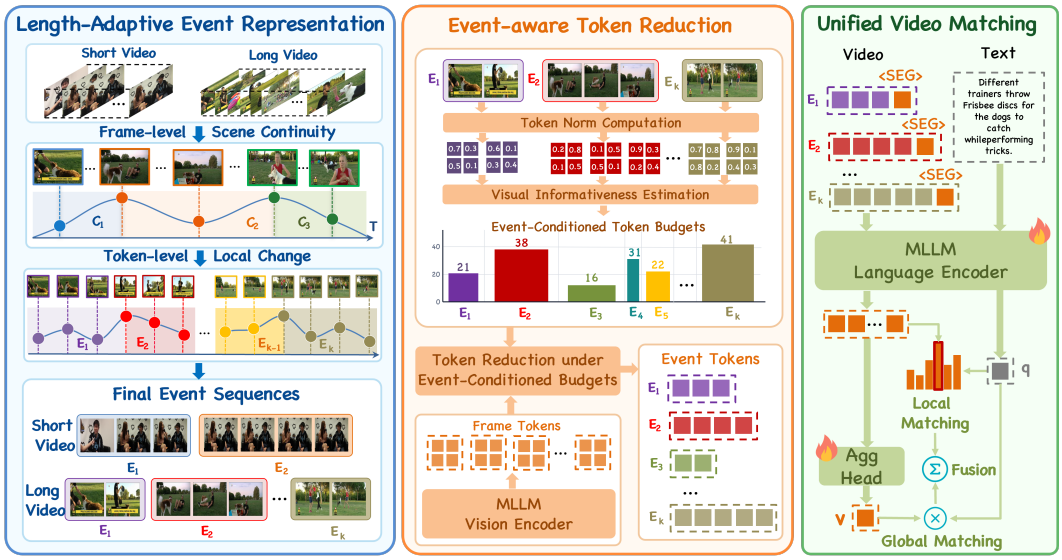

Figure 2: Overview of UMVR. It follows an event-centric formulation in which length-adaptive events provide a shared structure for video representation, token allocation, and relevance modeling. LAER adaptively constructs a sequence of temporally coherent events according to each video's temporal structure, while EATR allocates visual tokens based on event informativeness before MLLM encoding. The resulting event representations support both localized and holistic matching, whose similarities are combined into a unified relevance score for short- and long-video retrieval.

2.2 LENGTH-ADAPTIVE EVENT REPRESENTATION

Length-Adaptive Event Representation (LAER) aims to adapt the representation granularity to the temporal structure of each video. Rather than assigning every video a predefined number of temporal representation units, LAER adaptively constructs a sequence of temporally coherent events whose number and boundaries are determined by the video content. In particular, the resulting representation length K is not fixed across videos, while the temporal extent of individual events can also vary according to the underlying temporal structure. Given a video V, we uniformly sample T frames and encode them with the MLLM vision encoder:

$$\mathbf{X} = \Phi_{\text{vis}}(V) \in \mathbb{R}^{T \times P \times d}, \tag{1}$$

where $\mathbf{x}_{t,p} \in \mathbb{R}^d$ denotes the p-th spatial token of the t-th frame, P is the number of visual tokens per frame, and d is the feature dimension. We construct the event sequence directly from these visual features. Temporal boundaries may arise from major changes in the overall scene as well as finer

changes within a visually consistent scene. We therefore adopt a coarse-to-fine decomposition to obtain the final event partition $\mathcal{E} = \{E_k\}_{k=1}^{K}$, where K varies across videos.

Coarse scene decomposition. We first identify major temporal boundaries from changes in the global visual context. For each sampled frame, we obtain a normalized frame-level representation by averaging its visual tokens:

$$\mathbf{f}_t = \mathrm{Norm}\left(\frac{1}{P}\sum_{p=1}^{P}\mathbf{x}_{t,p}\right), \qquad s_t = \mathbf{f}_t^{\top}\mathbf{f}_{t+1}, \tag{2}$$

where s_t measures the visual continuity between two adjacent frames and $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization. A low similarity s_t may indicate a scene transition. However, the scale and distribution of adjacent-frame similarities vary across videos, making a fixed threshold unsuitable. We therefore measure how prominently each transition forms a similarity valley relative to its surrounding temporal context:

$$h_t = \frac{1}{2}\left(\max_{j<t} s_j + \max_{j>t} s_j - 2s_t\right), \tag{3}$$

A larger h_t indicates that the transition at t is more pronounced relative to the visual continuity on its two sides. We identify a coarse boundary between frames t and $t+1$ when $h_t > \mu_h + \alpha\sigma_h$, where μ_h and σ_h are the mean and standard deviation of the depth scores within the video, and α controls the boundary sensitivity. This adaptive criterion produces M coarse scenes $\{\mathcal{C}_m\}_{m=1}^{M}$, where $\mathcal{C}_m = [b_{m-1}, b_m)$.

Fine-grained event mining. Coarse scene boundaries mainly capture changes in global visual context, but a visually consistent scene may still contain multiple temporally distinct actions. Such changes can be localized to only a small spatial region and may therefore be suppressed by frame-level global averaging. We refine each coarse scene by measuring local visual changes between spatially corresponding tokens in adjacent frames:

$$\delta_{t,p} = 1 - \cos\left(\mathbf{x}_{t,p}, \mathbf{x}_{t+1,p}\right), \qquad b_{m-1} \le t < b_m - 1. \tag{4}$$

A larger $\delta_{t,p}$ indicates a stronger visual change at spatial position p. Since averaging $\delta_{t,p}$ over all spatial locations may dilute localized changes, we retain the top 20% most changed spatial tokens and define the frame-pair change score as:

$$a_t = \frac{1}{R}\sum_{r=1}^{R}\delta_{t,(r)}, \qquad R = \lfloor 0.2P \rfloor, \tag{5}$$

where $\delta_{t,(r)}$ denote the r-th largest value among $\{\delta_{t,p}\}_{p=1}^{P}$. Similar to the coarse stage, we use a scene-adaptive threshold rather than a fixed global threshold. A temporal position t within scene $\mathcal{C}_m$ is selected as a fine-grained boundary when $a_t > \mu_m^a + \beta\sigma_m^a$, where μ_m^a and σ_m^a denote the mean and standard deviation of the frame-pair change scores in $\mathcal{C}_m$, and β controls the boundary sensitivity. We additionally enforce a minimum temporal distance between consecutive boundaries to avoid fragmenting a continuous action into excessively short events. Refining all coarse scenes yields the final event partition $\mathcal{E} = \{E_k\}_{k=1}^{K}$. Both K and the temporal extent of individual events are adaptively determined by the video structure, LAER produces a representation sequence whose granularity varies across videos. The resulting event structure is subsequently used for Event-Aware Token Reduction and event-level relevance modeling.

2.3 Event-aware Token Reduction

Given the length-adaptive event partition $\mathcal{E}$ produced by LAER, we next reduce the visual tokens before MLLM encoding. A token-reduction operator determines which visual tokens are retained under a given token budget, but using the same budget throughout a video ignores substantial variations in visual information and redundancy across events. Visually informative events may require more tokens to preserve useful details, whereas events dominated by static or repetitive content can be represented with fewer tokens. We therefore introduce Event-Aware Token Reduction, which adapts the token budget according to the visual informativeness of each event and performs token reduction under the resulting event-conditioned budgets.

To estimate informativeness efficiently, we use the magnitude of visual-token activations as a lightweight signal. Following prior findings that visual tokens with larger feature norms tend to contain richer semantic information (Luo et al., 2026), we define the dimension-normalized norm of each visual token as $u_{t,p} = \frac{||\mathbf{x}_{t,p}||_2}{\sqrt{d}}$. Averaging over all tokens within an event may dilute informative visual content with background or redundant tokens. We therefore characterize each event using its most responsive visual tokens. Specifically, for each event E_k, we select the top 20% of its visual tokens according to $u_{t,p}$, denoted by Ω_k, and define its visual informativeness as:

$$w_k = \frac{1}{|\Omega_k|} \sum_{(t,p)\in\Omega_k} u_{t,p}, \tag{6}$$

We then convert the event informativeness scores into relative allocation factors. To enlarge the differences between events with different levels of visual informativeness, we apply quadratic scaling and normalize the resulting factors to have a mean of one across events:

$$\hat{w}_k = \frac{K w_k^2}{\sum_{j=1}^{K} w_j^2}. \tag{7}$$

Given a base per-frame token budget B, frames belonging to event E_k are assigned an event-conditioned budget $B_k = B\hat{w}_k$. Consequently, visually informative events retain more tokens, whereas less informative events are assigned smaller budgets and undergo stronger reduction. Given the event-conditioned budgets, we use FlashVID (Fan et al., 2026) as the underlying token reduction operator, which combines attention- and diversity-based token selection with tree-based spatiotemporal token merging. Rather than applying a uniform budget throughout the video, we perform token reduction using the event-specific budget B_k. Let $\widetilde{\mathbf{X}}_k$ denote the compressed tokens of event E_k. The compressed events are concatenated in temporal order while preserving their boundaries, forming an event-structured visual sequence for subsequent MLLM encoding.

2.4 Unified Local-Global Matching

Given the event-structured visual sequence after token reduction, we next construct representations that support both localized and holistic relevance modeling within the same MLLM encoding. Rather than collapsing the entire sequence into a single video representation, we preserve the event structure during encoding so that individual events remain accessible for localized matching. Specifically, we append a shared learnable <SEG> token to each compressed event and concatenate all events in temporal order. The resulting sequence is processed by the MLLM language backbone in a single forward pass. Let h_k^{seg} denote the hidden state at the k-th <SEG> position. We obtain the normalized event representation as $\mathbf{e}_k = \text{Norm}(\mathbf{h}_k^{\text{seg}})$, resulting in $\{\mathbf{e}_k\}_{k=1}^{K}$. Since all events are encoded jointly, each e_k is contextualized by the temporally ordered event sequence while remaining anchored to an individual event. The event representations directly support localized relevance modeling, but some queries may correspond to content distributed across multiple events or to the overall video semantics. We therefore aggregate the event representations into a video-level representation for holistic matching. A lightweight scoring network $\psi(\cdot)$ predicts the contribution of each event, and the weighted event representation is refined by a residual MLP $\phi(\cdot)$:

$$\alpha_k = \frac{\exp(\psi(\mathbf{e}_k))}{\sum_{j=1}^{K}\exp(\psi(\mathbf{e}_j))}, \qquad \bar{\mathbf{e}} = \sum_{k=1}^{K} \alpha_k \mathbf{e}_k, \quad \mathbf{v} = \text{Norm}\left(\bar{\mathbf{e}} + \phi(\bar{\mathbf{e}})\right), \tag{8}$$

During training, gradients from the global aggregation branch are stopped at the event representations. This prevents the global objective from directly driving individual event representations toward the same holistic video semantics, helping them retain the distinctions required for event-level matching while allowing the aggregation head to learn how to compose them into a global representation.

The text query is encoded by the same MLLM, and its normalized end-of-sequence representation is denoted by $\mathbf{q}$. We compute the localized similarity using the most relevant event and the holistic similarity using the aggregated video representation:

$$s_{\text{l}} = \max_{1\le k\le K}\{\mathbf{q}^\top \mathbf{e}_k\}, \qquad s_{\text{g}} = \mathbf{q}^\top \mathbf{v}, \qquad S(q,V) = \lambda s_{\text{l}} + (1-\lambda) s_{\text{g}}. \tag{9}$$

We set $\lambda = 0.8$ for both short- and long-video retrieval. The local similarity captures query-relevant content concentrated within an individual event, while the global similarity captures relevance distributed across the video. Combining the two allows the same representation and scoring formulation to accommodate different relevance patterns without task-specific retrieval heads.

2.5 Training and Inference

Given a mini-batch of N matched query-video pairs $\{(q_i, V_i)\}_{i=1}^{N}$, we optimize a text-to-video contrastive (Oord et al., 2018) objective using the unified relevance score:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(S(q_i, V_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(S(q_i, V_j)/\tau\right)}, \tag{10}$$

where τ is the temperature parameter. For each query, its paired video is treated as the positive target, while the remaining videos in the mini-batch serve as negatives. The same training objective and relevance formulation are applied to both short- and long-video retrieval, without task-specific losses or retrieval heads. At inference time, each text query is encoded into a query representation, while every candidate video is encoded into its event-level and video-level representations as described above. Their local and global similarities are combined using Eq. 9 to obtain the final relevance score $S(q, V)$. Candidate videos are then ranked according to this score. The same inference procedure and scoring function are used for both short- and long-video retrieval.

3 Experiment

3.1 Experimental Setup

Training Data. To train a unified retriever across short and long videos, we consider two complementary sources of video-language supervision. The first is the ActivityNet Captions training split (Krishna et al., 2017), which contains approximately 9K long untrimmed videos with 34K temporally localized descriptions. Compared with other long-video retrieval datasets, ActivityNet exhibits a broader distribution of query-relevant temporal coverage, ranging from brief localized moments to descriptions spanning large portions of a video. This broad coverage makes it particularly suitable for learning our unified local-global retrieval formulation. The second source is the mixed multimodal training collection (Jiang et al., 2024b; Zhang et al., 2025b; Faysse et al.; Yu et al., 2025) used by VLM2Vec-V2 (Meng et al., 2025). Since this collection has no explicit name, we refer to it as MMEB for convenience. To facilitate a clearer comparison with MMEB-trained models while maintaining a video-focused training setting, we use 240K video-QA samples from its LLaVA-Hound subset (Zhang et al., 2025b) and denote this subset by MMEB†. We train UMVR on ActivityNet and MMEB† separately, as well as on their combination.

Evaluation. We evaluate UMVR on three long-video retrieval benchmarks, including ActivityNet Captions (Krishna et al., 2017), QVHighlights (Lei et al., 2021), and TVR (Lei et al., 2020), and four short-video retrieval benchmarks, including MSR-VTT (Xu et al., 2016), VATEX (Wang et al., 2019), MSVD (Chen & Dolan, 2011), and DiDeMo (Anne Hendricks et al., 2017). We use Recall at rank K (R@K) as the evaluation metric, which measures the percentage of queries whose ground-truth video is ranked among the top-K candidates. We report R@1 and R@10 on all datasets.

3.2 Main Results

Table 1 evaluates whether a single UMVR model can support both long- and short-video retrieval within the same architecture and scoring formulation. We compare UMVR with recent MLLM-based retrieval methods, including the universal multimodal embedding models VLM2Vec-V2 (Meng et al., 2025), GME (Zhang et al., 2024), and LamRA (Liu et al., 2025), as well as the reasoning-enhanced embedding models UME-R1 (Lan et al., 2026) and PLUME (He et al., 2026). Since these methods are trained on different data collections, their results are not directly comparable under identical supervision. We therefore report the training data for each method and additionally retrain VLM2Vec-V2 with the same Qwen2-VL-2B backbone and ActivityNet training set as UMVR, providing a controlled comparison under the same backbone and training data.

Table 1: Comparison with MLLM-based retrieval methods across four short-video and three long-video retrieval benchmarks. We report R@1 and R@10 on all datasets together with the training data used by each method. All models use Qwen2-VL-2B as the backbone unless otherwise specified.

Method	Training Data	Short-video Retrieval								Long-video Retrieval					
		MSR-VTT		VATEX		MSVD		DiDeMo		ActivityNet		QVHighlights		TVR	
		R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
VLM2Vec-V2	MMEB	27.8	59.6	26.3	68.4	47.6	84.0	30.3	68.0	16.2	45.1	21.4	54.6	4.4	15.5
GME	Mixed Dataset-8M	34.8	69.5	30.5	68.1	**57.1**	**89.6**	31.6	66.9	**20.2**	53.0	28.3	62.5	6.2	21.7
LamRA	NLI & M-BEIR	30.1	67.0	27.4	66.9	52.2	87.3	28.7	63.2	18.3	51.4	27.6	62.1	6.0	19.2
PLUME	MMEB & CoT Anno	**36.2**	**69.9**	28.2	62.8	56.1	89.4	32.7	65.8	16.9	47.5	22.5	55.5	6.4	19.7
UME-R1	MMEB & CoT Anno	34.3	67.2	29.9	63.7	55.4	89.6	32.4	65.9	19.6	53.5	25.1	58.0	5.7	17.8
UMVR	MMEB†	31.7	66.9	32.0	**74.0**	51.1	88.5	39.2	75.1	19.4	53.3	25.7	64.8	6.8	20.9
UMVR*	MMEB†	33.8	68.7	**34.5**	72.9	54.3	89.1	**41.2**	**75.6**	19.9	**54.1**	**28.8**	**65.2**	**6.9**	**21.4**
VLM2Vec-V2	ActivityNet	30.7	65.7	31.5	71.4	49.3	85.6	32.1	66.3	20.1	56.8	20.4	53.4	5.7	18.4
UMVR	ActivityNet	**35.6**	**69.5**	**38.0**	**77.8**	**53.0**	**91.5**	**40.6**	**75.2**	**23.7**	**60.1**	**30.2**	**69.1**	**8.9**	**25.2**
UMVR	MMEB† & ActivityNet	36.0	70.0	34.8	74.4	56.4	90.6	45.7	76.3	24.2	60.8	31.5	68.5	9.0	26.3

Under the controlled ActivityNet-only training setting, UMVR consistently outperforms VLM2Vec-V2 across all seven benchmarks at both R@1 and R@10. In this setting, VLM2Vec-V2 processes the full visual-token sequence and represents each video with a single global embedding, providing a controlled baseline without the event-centric components introduced in UMVR. The consistent improvements across both long-video and short-video benchmarks therefore demonstrate the effectiveness of UMVR's event-centric formulation and show that the same representation and scoring framework can accommodate different video lengths and relevance patterns.

Beyond the controlled ActivityNet-only setting, we further evaluate UMVR under different sources of training and inference-time supervision. When trained only on the 240K video-QA samples from MMEB†, UMVR remains competitive with methods trained on broader multimodal collections across both long-video and short-video benchmarks. We additionally report UMVR$*$, which incorporates the reasoning data from UME-R1 only at inference time without additional training. Compared with the standard UMVR, UMVR$*$ consistently improves retrieval performance across nearly all metrics, showing that inference-time reasoning can provide complementary information to the proposed retrieval framework. Interestingly, UMVR trained only on ActivityNet outperforms its MMEB†-trained counterpart across all seven benchmarks, suggesting that temporally grounded video-language supervision transfers effectively across different video retrieval settings. Combining MMEB† with ActivityNet further improves performance on several benchmarks, indicating that UMVR can also benefit from broader video-language supervision while maintaining a unified retrieval formulation.

3.3 Video Encoding Efficiency

Table 2 evaluates the efficiency of MLLM video encoding on ActivityNet using a single NVIDIA A100 GPU. We report the average number of video-side tokens, the latency of token reduction and MLLM encoding, average TFLOPs, and retrieval performance. Since all compared methods use the same vision encoder, we exclude the shared vision-encoding latency and focus on the subsequent processing stages. We evaluate UMVR with retention ratios of 50% and 10% to examine its performance-efficiency trade-off under different token budgets. For UME-R1, the reported latency additionally includes its inference-time reasoning stage. Compared with the full-token baselines, UMVR substantially reduces the number of visual tokens processed by the MLLM and correspondingly lowers encoding cost. Although Event-Aware Token Reduction introduces additional reduction overhead, the resulting shorter visual sequence considerably decreases MLLM encoding time and overall computation. At 50% retention, UMVR achieves the strongest retrieval performance in the table while requiring substantially fewer visual tokens than the full-token baselines. Reducing the retention ratio to 10% further lowers the computational cost with only a moderate performance decrease, while still outperforming both full-token baselines in retrieval accuracy. These results show that Event-Aware Token Reduction provides a favorable trade-off between retrieval performance and MLLM encoding efficiency across different token budgets.

Table 2: MLLM video-encoding efficiency and retrieval performance of our method. We evaluate computational overhead and retrieval performance on ActivityNet.

Method	Ret. Ratio	Avg. # Tokens	Post-Vision Encoding Time (ms)			Avg. TFLOPs	Retrieval	
			Compression	LLM	Total		R@1	R@10
VLM2Vec-V2	100%	11784	–	296.9	296.9	44.8	20.1	56.8
UME-R1	100%	12021	–	4204.4	4204.4	51.5	19.6	53.5
UMVR	50%	5008	136.0	142.1	278.1	16.2	**25.2**	**62.3**
UMVR	10%	1085	86.6	33.1	119.7	3.7	23.7	60.1

3.4 Ablation Studies

We next conduct ablation studies to examine the key design choices of UMVR. Unless otherwise specified, all models are trained on ActivityNet. We report results on VATEX and ActivityNet as representative short- and long-video retrieval benchmarks, respectively.

Table 3: Ablation of the main components of UMVR.

Variant	VATEX		ActivityNet	
	R@1	R@10	R@1	R@10
Uniform Token Budget	37.1	76.8	22.1	58.9
Local Matching Only	36.8	77.4	22.9	59.1
Global Matching Only	36.2	76.7	21.5	56.8
UMVR	**38.0**	**77.8**	**23.7**	**60.1**

Main Components. Table 3 evaluates the main design choices of UMVR. To isolate the adaptive budgeting mechanism in Event-Aware Token Reduction, we keep the underlying token-reduction operator unchanged and assign the same per-frame token budget to all events. This modification consistently reduces performance on both VATEX and ActivityNet, showing the benefit of adjusting the reduction budget according to the visual informativeness of each event. We further remove the global and local matching branches separately to evaluate their contributions to the unified relevance formulation. Removing either branch degrades performance on both datasets, confirming that localized and holistic relevance provide complementary information. In particular, removing local matching results in a larger degradation, showing that localized event-level evidence is useful not only for long-video retrieval but also for short-video retrieval.

Length-Adaptive Event Representation. Table 4 examines how different temporal partition strategies affect the event representations used for retrieval. *No Temporal Partition* places the summary tokens only after the entire visual sequence and therefore provides no explicit temporal partition. *Uniform Partition* divides every video into ten equal temporal segments, introducing localized representations but imposing the same predefined granularity across videos. *Coarse Scene Partition* instead determines temporal boundaries from major changes in global visual content, allowing the partition to adapt to the structure of each video. *Coarse-to-Fine Event Partition* further refines these scenes using localized visual changes, allowing both the number and temporal extent of events to adapt to individual videos. As shown in Table 4, content-dependent partitions outperform the fixed alternatives, with Coarse-to-Fine Event Partition achieving the best performance on both VATEX and ActivityNet. These results support adapting the number and temporal extent of representation units to individual videos rather than imposing a predefined temporal partition.

Event-to-Video Representation Aggregation. While individual event representations preserve localized video content, holistic matching requires integrating information across events into a video-level representation. We therefore compare different strategies for aggregating the event representations in Table 5 *Mean Pooling* treats all events equally and directly averages their representations. *SeqTrans* explicitly models interactions among the temporally ordered events with an additional Transformer encoder (Vaswani et al., 2017). Our *learned aggregation* instead assigns different contributions to events and combines them into a video-level representation using the lightweight aggregation module described in Sec 2.4. It achieves the best overall performance on both VATEX and ActivityNet, showing that selectively weighting event representations is more effective than uniform averaging, while additional inter-event sequence modeling is not necessary for constructing a strong holistic representation.

Table 4: Effect of different temporal partition strategies for event representation.

Temporal Partition	VATEX		ActivityNet	
	R@1	R@10	R@1	R@10
No Temporal Parti.	36.4	75.4	20.8	57.7
Uniform Parti.	37.3	77.1	21.7	58.9
Coarse Scene Parti.	37.5	77.3	22.8	58.6
Coarse-to-Fine Event Parti.	**38.0**	**77.8**	**23.7**	**60.1**

Table 5: Effect of different event-to-video aggregation strategies.

Aggregator	VATEX		ActivityNet	
	R@1	R@10	R@1	R@10
Mean Pooling	36.6	77.5	20.4	56.9
SeqTrans	37.5	77.2	22.9	60.0
Learned Agg.	**38.0**	**77.8**	**23.7**	**60.1**

3.5 Further Analysis

Performance across Video Durations. To examine how UMVR behaves across videos with different temporal lengths, we group ActivityNet videos by duration and report retrieval performance within each interval in Fig. 3. UMVR consistently outperforms VLM2Vec-V2 and UME-R1 across all duration ranges, with particularly clear advantages on longer videos. We further report the average number of events produced by LAER within each duration interval. The event count generally increases with video duration, confirming that LAER does not impose a fixed representation length across videos. Instead, the resulting event sequence can vary substantially as the temporal structure of the video changes. Together, these results show that UMVR remains effective across a broad range of video durations while using adaptive event representations.

Transferability of Event-Aware Token Reduction. To examine whether Event-Aware Token Reduction can be applied beyond UMVR, we integrate it into VLM2Vec-V2 and UME-R1. Figure 4 compares full-token encoding with token reduction using either a uniform budget or our event-aware budget under the same 10% retention ratio. Although token reduction inevitably introduces some performance degradation relative to full-token encoding, adapting the reduction budget according to event visual informativeness consistently outperforms using a uniform budget for both models. In both cases, the event-aware strategy recovers more than half of the performance lost under uniform reduction. These results show that the proposed budgeting mechanism is not tied to the UMVR architecture and can be readily incorporated into different MLLM-based retrieval models to improve performance under the same token budget.

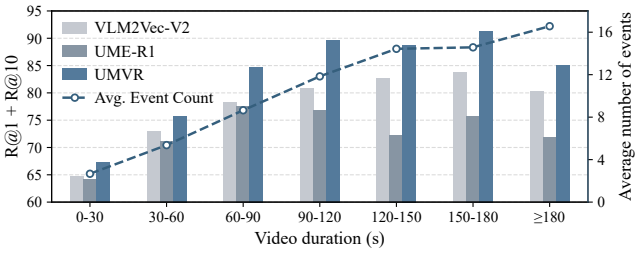

Figure 3: Retrieval performance and average event count across different video durations on ActivityNet.

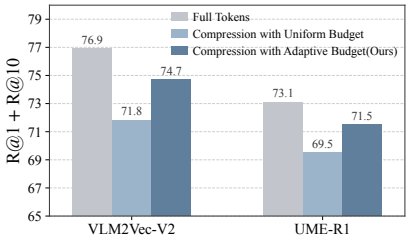

Figure 4: Transferability of Event-Aware Token Reduction across MLLM-based retrieval models.

4 Conclusion

In this work, we presented UMVR, a unified and efficient MLLM-based framework for short- and long-video retrieval. At its core, Length-Adaptive Event Representation (LAER) organizes each video as a sequence of temporally coherent events whose number and temporal extent adapt to its temporal structure, allowing videos with different temporal structures to be handled within the same representation and scoring formulation. Built upon an MLLM backbone, UMVR further incorporates Event-Aware Token Reduction (EATR) to improve video encoding efficiency by adapting the visual-token budget according to event informativeness. The resulting event representations support both localized and holistic relevance modeling within a unified matching formulation. Across seven short- and long-video retrieval benchmarks, UMVR demonstrates strong retrieval performance while retaining only 10% of the visual tokens for MLLM encoding, remains effective across a broad range of video durations, and shows that EATR can be transferred to other MLLM-based retrievers.

AI USE STATEMENT

In the preparation of this paper, we use Large Language Models (LLMs) solely to aid in writing and polishing the text, including improving clarity, grammar, and readability. LLMs are not used for generating scientific content, experimental design, analysis, or conclusions. All technical ideas, experiments, and results reported in this paper are entirely the work of the authors.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics and does not involve the collection of new human subject data or any personally identifiable information. All datasets used in this study are publicly available and widely adopted in the research community. We have carefully ensured that our methodology and experiments comply with relevant ethical standards, including fairness, transparency, and reproducibility.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training procedure, hyperparameters, datasets, and evaluation settings in the main paper and appendix to facilitate reproducibility. Upon acceptance, we will release the implementation and trained models used in this work to support reproduction of the reported results and future research.

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.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1708–1718. IEEE, 2021.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

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.

Jianfeng Dong, Minsong Zhang, Zheng Zhang, Xianke Chen, Daizong Liu, Xiaoye Qu, Xun Wang, and Baolong Liu. Dual learning with dynamic knowledge distillation for partially relevant video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11302–11312, 2023.

Jianfeng Dong, Lei Huang, Daizong Liu, Xianke Chen, Xun Yang, Changting Lin, Xun Wang, and Meng Wang. Dual learning with dynamic knowledge distillation and soft alignment for partially relevant video retrieval. *arXiv preprint arXiv:2510.12283*, 2025.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. FlashVID: Efficient video large language models via training-free tree-based spatiotemporal token merging. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=H6rDX4w6Al.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: efficient document retrieval with vision language models (2024). *URL https://arxiv. org/abs/2407.01449*.

Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5006–5015, 2022.

Chenwei He, Xiangzhao Hao, Tianyu Yang, Yuxiang Ma, Yuheng Jia, Lingxiang Wu, Chaoyang Zhao, Haiyun Guo, and Jinqiao Wang. Plume: Latent reasoning based universal multimodal embedding. *arXiv preprint arXiv:2604.02073*, 2026.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024a.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024b.

Xiaolun Jing, Genke Yang, and Jian Chu. Tc-mgc: Text-conditioned multi-grained contrastive learning for text–video retrieval. *Information Fusion*, 121:103151, 2025.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. In *International Conference on Learning Representations*, volume 2026, pp. 115779–115806, 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.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Jun Li, Jinpeng Wang, Chaolei Tan, Niu Lian, Long Chen, Yaowei Wang, Min Zhang, Shu-Tao Xia, and Bin Chen. Hlformer: Enhancing partially relevant video retrieval with hyperbolic learning. *arXiv preprint arXiv:2507.17402*, 2025.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4015–4025, June 2025.

Bingjun Luo, Tony Wang, Hanqi Chen, and Xinpeng Ding. Enhancing visual token representations for video large language models via training-free spatial-temporal pooling and gridding. *arXiv preprint arXiv:2605.22078*, 2026.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 638–647, 2022.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, et al. Fastvid: Dynamic density pruning for fast video large language models. *Advances in Neural Information Processing Systems*, 38:123553–123581, 2026.

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.

Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16551–16560, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4581–4591, 2019.

Yuting Wang, Jinpeng Wang, Bin Chen, Tao Dai, Ruisheng Luo, and Shu-Tao Xia. Gmmformer v2: An uncertainty-aware framework for partially relevant video retrieval. *arXiv preprint arXiv:2405.13824*, 2024c.

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, 2024d.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2816–2827, 2023.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, volume 2025, pp. 21074–21098, 2025.

Long Zhang, Peipei Song, Jianfeng Dong, Kun Li, and Xun Yang. Enhancing partially relevant video retrieval with robust alignment learning. *arXiv preprint arXiv:2509.01383*, 2025a.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 694–717, 2025b.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

Sa Zhu, Huashan Chen, Wanqian Zhang, Jinchao Zhang, Zexian Yang, Xiaoshuai Hao, and Bo Li. Uneven event modeling for partially relevant video retrieval. *arXiv preprint arXiv:2506.00891*, 2025.

A Appendix

A.1 Related Works

Text-based video retrieval. Text-based video retrieval has traditionally been studied with different representation and matching strategies for short and long videos. For short videos, earlier methods learn joint video-text embedding spaces using dedicated video encoders and temporal aggregation modules (Bain et al., 2021). With large-scale image-text pretraining, subsequent approaches adapt CLIP to videos through frame aggregation, temporal modeling, or query-conditioned pooling (Luo et al., 2022; Gorti et al., 2022; Xue et al., 2022). These methods are typically designed for videos whose associated text can be matched against a compact set of video-level representations. Long-video retrieval introduces a different representation requirement because query-relevant content may occupy only a limited temporal region of an untrimmed video. Partially relevant video retrieval methods therefore preserve finer temporal representations and perform localized matching. MS-SL formulates this problem as multiple-instance learning and measures query-video correspondence at both frame and clip scales (Dong et al., 2022). Subsequent methods improve partial matching through knowledge distillation (Dong et al., 2023), implicit multi-scale clip modeling (Wang et al., 2024d;c), and active moment discovery (Song et al., 2025). Closely related to our event-based representation, UEM replaces uniform temporal segments with uneven events and refines their representations through query-conditioned interaction (Zhu et al., 2025). Despite sharing the same objective of estimating text-video relevance, these two lines of work typically adopt representation granularities tailored to their target retrieval settings, with short-video methods emphasizing more global representations and long-video methods preserving more temporally localized representations. UMVR instead introduces Length-Adaptive Event Representation, which adapts the representation granularity to the temporal structure of individual videos and enables both short- and long-video retrieval within the same event-centric representation and scoring formulation.

MLLM-Based Multimodal Embedding. Recent work has repurposed Multimodal Large Language Models as general-purpose embedding models. E5-V demonstrates that MLLMs can produce universal multimodal embeddings through appropriate prompting and text-only contrastive training (Jiang et al., 2024a). VLM2Vec converts pretrained vision-language models into instruction-aware embedders through contrastive training on diverse multimodal tasks (Jiang et al., 2024b). GME and LamRA improve generalization through large-scale multimodal training and retrieval-oriented optimization (Zhang et al., 2024; Liu et al., 2025), while VLM2Vec-V2 further extends unified embedding learning to videos and visual documents (Meng et al., 2025). More recent approaches incorporate explicit or latent reasoning before embedding extraction to improve complex multimodal matching (Lan et al., 2026; He et al., 2026). These models demonstrate that diverse visual inputs and retrieval tasks can share a pretrained MLLM backbone and embedding space, providing a suitable foundation for a common video retrieval architecture. However, existing MLLM-based embedding methods typically represent a video with a single global embedding. Sharing the backbone therefore does not by itself address the different representation granularities required by videos with different temporal structures and relevance patterns. UMVR retains the shared MLLM embedding framework while introducing Length-Adaptive Event Representation to support both localized and holistic relevance modeling within the same scoring formulation.

Efficient Video Encoding for MLLMs. Encoding videos with MLLMs requires processing large visual-token sequences, motivating extensive work on visual-token reduction through selection, pruning, and merging. VisionZip selects informative visual tokens and merges redundant contextual tokens before language-model processing (Yang et al., 2025). Video-specific approaches further exploit temporal redundancy. PruneVid combines spatiotemporal token merging with query-guided pruning (Huang et al., 2025), while FastVID dynamically partitions videos and applies density-based pruning (Shen et al., 2026). FlashVID combines attention- and diversity-based token selection with tree-based spatiotemporal token merging, providing an effective training-free reduction operator for video MLLMs (Fan et al., 2026). Existing methods primarily focus on determining which visual tokens should be retained or merged under a given reduction budget. However, the amount of retained information is commonly controlled by a predefined token budget or retention ratio rather than adapted to variations in visual information across temporal regions. Applying the same budget throughout a video may therefore allocate excessive capacity to redundant content while insufficiently preserving visually informative regions. UMVR addresses this complementary problem through Event-Aware Token Reduction. It leverages the event structure produced by LAER

Table 6: Ablation of design choices in Event-Aware Token Reduction.

Design Choice	Variant	VATEX		ActivityNet	
		R@1	R@10	R@1	R@10
Reduction Operator	VisionZip	37.1	77.0	22.8	59.6
	FlashVID	**38.0**	**77.8**	**23.7**	**60.1**
Weight Scaling	Linear	**38.0**	77.4	23.2	**60.2**
	Quadratic	38.0	**77.8**	**23.7**	60.1
Event Scoring	Mean	37.3	77.0	22.1	59.3
	Top-20%	**38.0**	77.8	**23.7**	60.1
	Top-40%	37.8	**78.0**	23.5	**60.3**
	Top-60%	37.7	77.1	22.9	59.3
	Top-80%	37.5	77.1	22.2	59.5

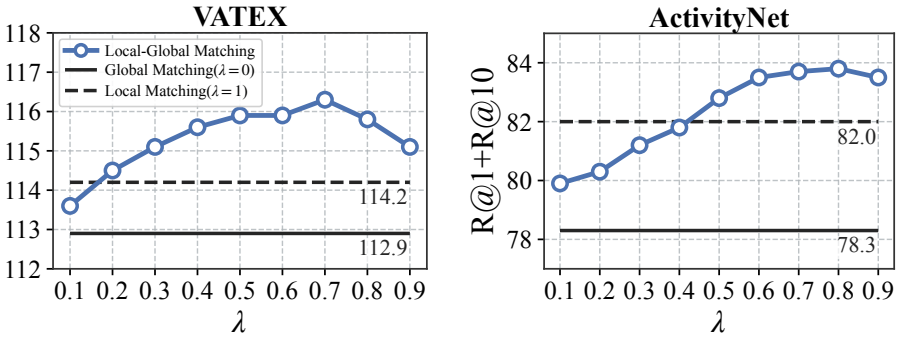

Figure 5: Sensitivity to the fusion weight λ.

to adjust the reduction budget according to event visual informativeness, and then applies an existing reduction operator under the resulting budgets. This separates the decision of how much visual information different events should retain from the token-selection and merging operations used to realize the reduction.

A.2 Implementation Details.

We implement UMVR using Qwen2-VL-2B-Instruct (Wang et al., 2024b) as the MLLM backbone. We uniformly sample 128 frames from each training video. At evaluation, we sample 128 frames for long-video benchmarks and 64 frames for short-video benchmarks to reflect their different temporal extents. A base visual-token retention ratio of 0.1 is used throughout training and evaluation, while the actual per-frame token budget is adjusted according to the visual informativeness of each event. For event decomposition, we set the coarse and fine boundary coefficients α and β to 1.0 and 0.5, respectively. We fine-tune the backbone using LoRA with a rank of 32 and a scaling factor of 64, while jointly optimizing the learnable <SEG> token and the aggregation head. The model is trained on four NVIDIA A100 GPUs with a global batch size of 128, an initial learning rate of 1e-4. The contrastive temperature τ is set to 0.02.

A.3 Additional Analysis.

Sensitivity to Local-Global Fusion. We fix $\lambda = 0.8$ for all experiments, and Figure 5 provides a sensitivity analysis of this choice. On ActivityNet, performance generally improves as more weight is assigned to local matching and peaks at $\lambda = 0.8$, reflecting the importance of localized event-level evidence for long-video retrieval. VATEX shows a similar preference for combining local and global relevance, with the best performance obtained at $\lambda = 0.7$. Importantly, relying exclusively on either local or global matching leads to lower performance on both datasets, confirming that the two relevance signals are complementary. The similar behavior across VATEX and ActivityNet supports using a shared fusion formulation across short- and long-video retrieval without setting-specific scoring strategies.

More Analysis of Event-Aware Token Reduction. Table 6 further examines several design choices in Event-Aware Token Reduction. Specifically, we study the underlying token-reduction operator, the scaling function used to convert event visual-informativeness scores into relative budget factors, and the token subset used to estimate event visual informativeness. We first compare FlashVID with VisionZip (Yang et al., 2025) as the underlying token-reduction operator while keeping the same event partition and reduction budgets. FlashVID consistently performs better on both ActivityNet and VATEX. VisionZip mainly selects dominant visual tokens based on spatial attention, whereas FlashVID additionally reduces redundancy across spatial and temporal dimensions. The results indicate that adapting the reduction budget across events should be coupled with an effective token-reduction operator that preserves useful visual information under the assigned budget.

We next examine how event visual-informativeness scores are converted into relative budget factors. Linear scaling directly preserves the relative scores, whereas quadratic scaling enlarges their differences before normalization, resulting in greater variation in the budgets assigned to different events. The two choices yield similar overall performance. Quadratic scaling improves R@1 on ActivityNet and R@10 on VATEX, matches linear scaling on VATEX R@1, and is only slightly lower on ActivityNet R@10. We therefore adopt quadratic scaling in the final configuration.

Finally, we study how visual-token responses are used to estimate event informativeness. Mean aggregation averages the norms of all visual tokens within an event, whereas Top-$p\%$ scoring uses only the tokens with the largest norms. Top-20% and Top-40% achieve the strongest overall results, while performance generally decreases as a larger fraction of tokens is included. This trend indicates that averaging over many less responsive tokens can weaken the informativeness signal. Top-20% achieves the highest R@1 on both datasets, whereas Top-40% achieves the highest R@10. We adopt Top-20% in the final configuration to prioritize early-rank retrieval. Overall, the results show that Event-Aware Token Reduction is affected more strongly by the choice of reduction operator and event-informativeness estimation than by the precise scaling function used to derive the reduction budgets.

A.4 Qualitative Analysis

Figures 6 and 7 qualitatively illustrate how LAER adapts event representations to different temporal structures and relevance patterns, and how the resulting events further support event-aware token reduction and localized relevance modeling. For each example, we visualize the video-dependent event structure produced by LAER, the visual-token budgets determined according to event visual informativeness, and the resulting query-event similarities. Figure 6 presents two short-video examples with different relevance distributions. In the first example, the query-relevant content occurs only near the beginning of the video despite its short duration. LAER isolates this content into a localized event, which is assigned a relatively large visual-token budget and also produces the highest query-event similarity. This example shows that short videos do not necessarily exhibit purely holistic relevance and can still contain temporally sparse evidence. In the second example, the query corresponds to content distributed across a broader temporal range. LAER separates the interaction into several temporally coherent events, among which the event providing the strongest correspondence to the query obtains the highest similarity, while other semantically related events also retain relatively high scores. Together, these examples show that LAER can adapt its representation to different relevance patterns even within short videos, rather than imposing a uniformly holistic representation based on the retrieval setting. Figure 7 presents a long-video example in which two queries refer to temporally adjacent but semantically distinct activities. At the coarse level, both relevant moments fall within the same scene C_3 because they share a similar environment and overall visual context. LAER further separates this scene into finer events, with E_6 andE_7 closely corresponding to the ground-truth moments of Query A and Query B, respectively. Importantly, the event partition is constructed without access to either query or temporal annotation. Event-Aware Token Reduction also assigns substantial visual-token budgets to both events based only on their visual informativeness, preserving information for multiple potential queries within a single video encoding. When the resulting event representations are matched against the two queries, the similarity peak shifts from E_6 for Query A to E_7 for Query B. Although the two events are temporally adjacent and share similar visual context, their representations remain sufficiently distinct to retrieve the corresponding activities. This example illustrates how LAER preserves localized semantic distinctions within a long video while allowing different queries to retrieve different event-level evidence from the same encoded representation.

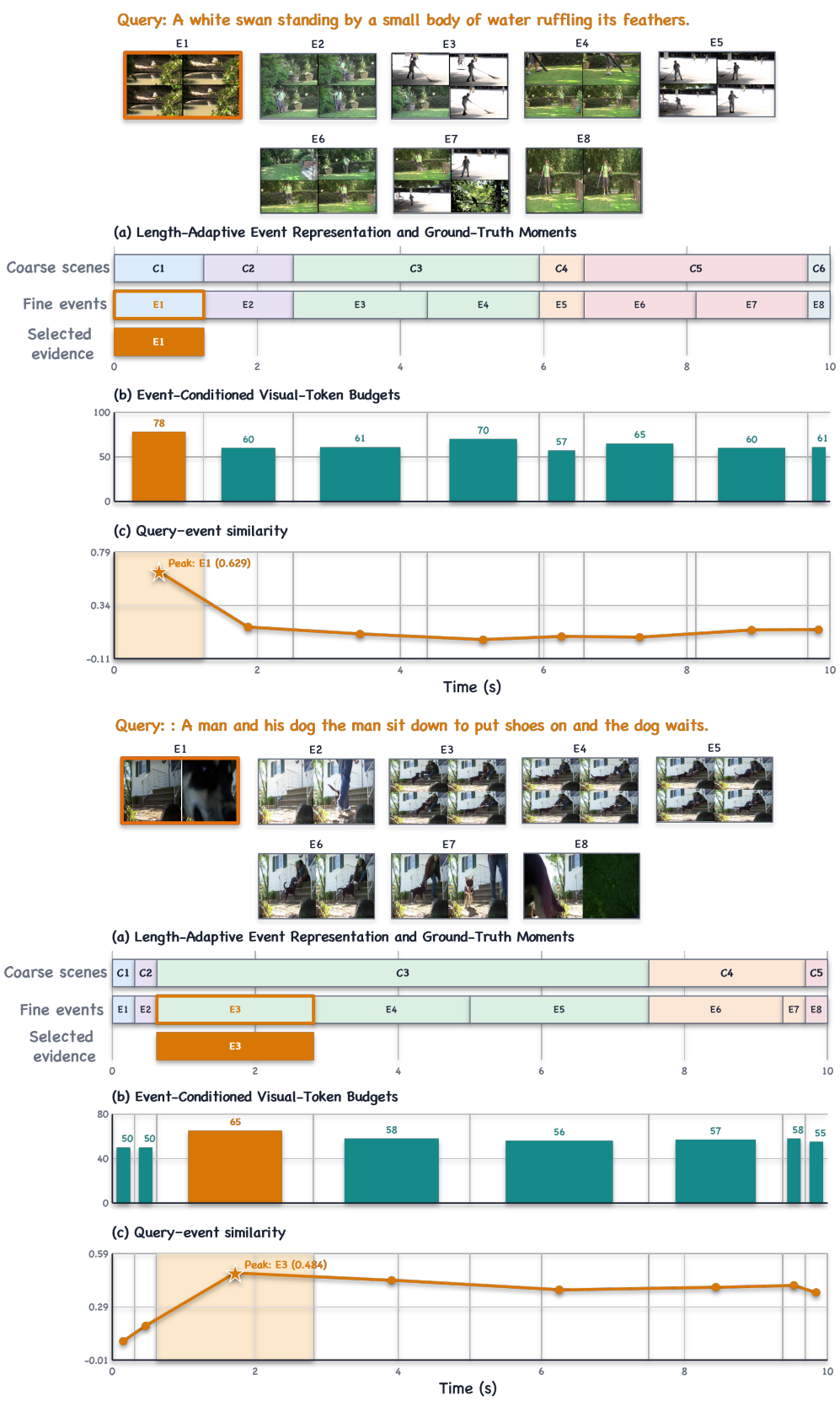

Figure 6: Qualitative examples on VATEX. UMVR captures both sparse relevance at the beginning of a short video and relevance distributed across multiple events.

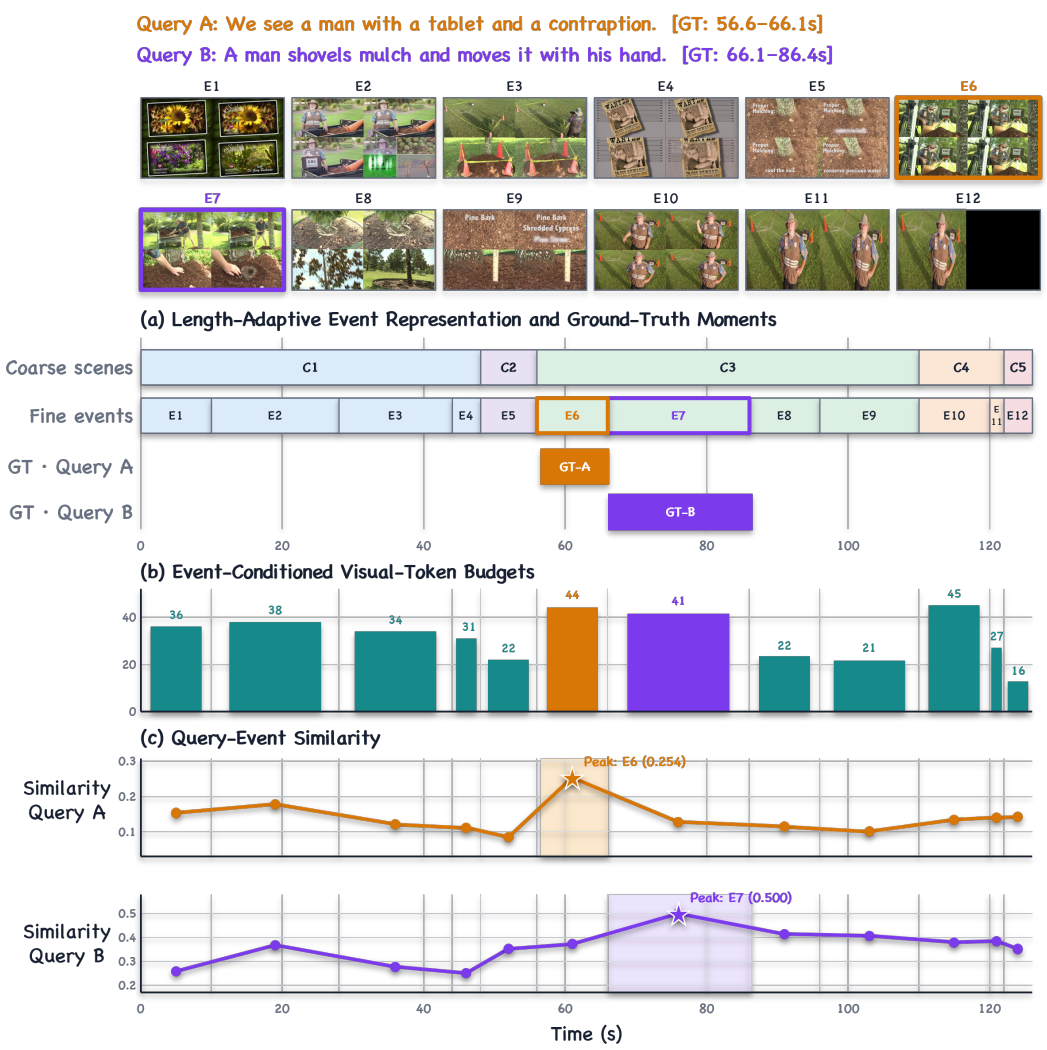

Figure 7: Qualitative example on ActivityNet Captions. Although two adjacent query moments belong to the same coarse scene, the fine-grained decomposition separates them into different events.