TOPOATOM: QUERY-CONDITIONED SEMANTIC ATOMIZATION FOR ZERO-SHOT VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot video temporal grounding (ZS-VTG) increasingly leverages frozen multimodal models to localize natural-language queries without task-specific training. Yet existing approaches largely build localization upon predefined temporal primitives, such as uniformly sampled frames or fixed-length clips, and subsequently measure their relevance to the query. Such fixed granularity overlooks a fundamental property of videos: semantics are structurally organized rather than uniformly distributed along the timeline. We introduce **TOPOATOM**, a training-free framework that replaces predefined temporal primitives with adaptively discovered, query-conditioned semantic atoms. Based on pairwise clip distances, we build an H_0-inspired connectivity filtration, whose component-merging process is compactly represented by a minimum spanning tree and its induced single-linkage hierarchy. Starting from the whole video, TOPOATOM traverses this hierarchy top-down and recursively splits a component only when a finer descendant exhibits stronger query correspondence; otherwise, the component is retained as a semantic atom. These adaptive semantic atoms provide a structure-aware foundation for candidate generation. We further calibrate each candidate using a topology-derived persistence score, moving beyond sole reliance on query relevance by explicitly incorporating inherent connectivity and compactness. Experiments on Charades-STA, ActivityNet-Captions, and QVHighlights consistently demonstrate the effectiveness of TOPOATOM.

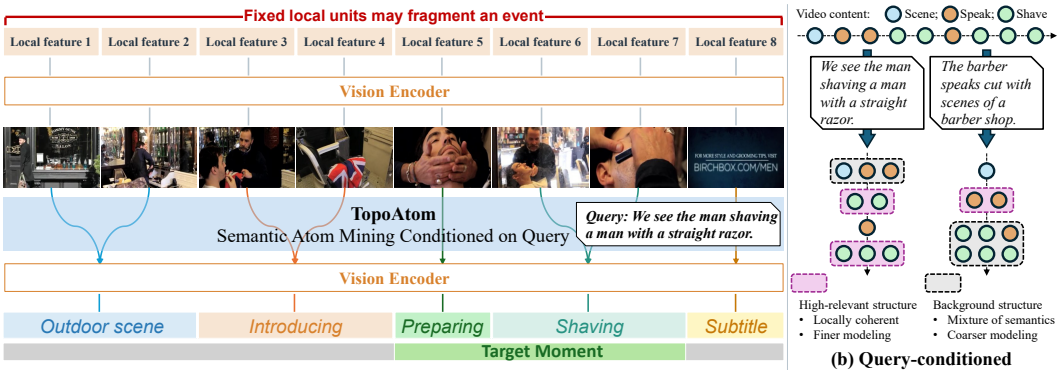
1 INTRODUCTION

Video temporal grounding (VTG) aims to localize the temporal segment in an untrimmed video that best corresponds to a natural-language query. Recent advances in large-scale vision–language models (VLMs) and multimodal large language models (MLLMs) have enabled a particularly appealing setting: zero-shot video grounding, where temporal localization is performed without task-specific training or temporal annotations. Frozen vision–language representations enable localization through similarity estimation and proposal search (Luo et al., 2024; Zheng et al., 2024), while other approaches exploit generated captions or reformulated queries to improve semantic alignment between video content and language descriptions (Xu et al., 2025; Jeon et al., 2026; Lee et al., 2026).

Despite this progress, a fundamental issue in the visual representation underlying zero-shot grounding remains largely underexplored. Most existing methods instantiate video evidence using uniformly sampled frames or fixed-length clips, which are independently encoded before subsequent similarity estimation, proposal construction, or multimodal reasoning. Such a representation implicitly imposes a *uniform temporal granularity* on the video. However, video semantics are inherently non-uniform along the temporal axis: a coherent action may continuously span multiple neighboring clips, a transition may occur within a very short interval, and long portions of a video may contain rich yet query-irrelevant content. Consequently, uniformly partitioning and independently encoding a video can fragment the coherent semantic event or introduce boundaries at locations unrelated to its semantic structure, as illustrated in the upper part of Figure 1(a).

Consider the example in Figure 1(a). When the query asks for "the man shaving a man with a straight razor," the video should be modeled with sufficient granularity around the preparation and shaving-related content, where subtle visual transitions are critical for identifying the target boundaries. In

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) Different granularities to organize video content. **(b) Query-conditioned video structure discovery.**

Figure 1: (a) Top: Existing zero-shot video grounding methods typically divide a video into uniform frames or clips and encode them independently, which can fragment temporally coherent semantics. Bottom: TOPOATOM instead performs query-conditioned video structure discovery. (b) By adaptively grouping locally coherent content conditioned on the query, structures termed *semantic atoms* are constructed to serve as base primitives for composing candidate moments.

contrast, clearly unrelated regions, such as an outdoor scene or a subtitle segment, provide little benefit from further fine-grained decomposition. A different query could induce a substantially different ideal video partition. Therefore, the query provides more than a matching signal used after visual encoding; it can serve as a prior for determining *how the video itself should be organized*.

Motivated by this perspective, we propose **TOPOATOM**, a training-free framework that discovers *query-conditioned structures*, termed *semantic atoms*, from the intrinsic temporal dynamics of a video. TOPOATOM changes the basic localization primitive from a predefined temporal unit to the adaptive semantic atoms discovered jointly from video structure and query relevance, as illustrated in Figure 1(b). The resulting structures preserve coherent temporal semantics, allocate fine-grained modeling capacity according to the query, and provide an explicit structural criterion for evaluating candidate moments. Experiments on QVHighlights, Charades-STA, and ActivityNet-Captions demonstrate consistent improvements, validating the effectiveness of semantic atom discovery and topology-aware candidate evaluation.

Our main contributions are summarized as follows:

- We propose **TOPOATOM**, a training-free zero-shot video grounding framework that, to the best of our knowledge, is the first to introduce topology-inspired analysis into zero-shot VTG and model the intrinsic video structure in a query-conditioned manner.
- We develop a **query-conditioned semantic atom discovery** algorithm that constructs an H_0 merge hierarchy from clip-wise relations and performs query-adaptive top-down pruning to identify semantically coherent video units at adaptive temporal scales.
- We introduce **candidate persistence**, a topology-derived criterion that characterizes the interval between a candidate's internal connectivity and external attachment, providing complementary structural evidence for candidate evaluation.
- We conduct extensive experiments on multiple widely used benchmarks, demonstrating the effectiveness and general applicability of TOPOATOM for zero-shot video grounding.

2 RELATED WORK

Learning temporal localization. Supervised approaches learn from temporal annotations (Gao et al., 2017; Lei et al., 2021b; Lin et al., 2023; Jung et al., 2025; Ma et al., 2025); while delivering precise localization, they suffer from high annotation costs and limited scalability. Weak supervision mitigates this by using only video–sentence pairs, leveraging proposal competition (Zhang et al., 2020), uncertainty-guided self-training (Huang et al., 2023), or global query context (Lv & Su, 2025). Our approach completely avoids model training, requiring no boundary or paired supervision.

Grounding with frozen models. Frozen-model retrieval (Luo et al., 2024) and TFVTG (Zheng et al., 2024) exploit pretrained similarities and temporal proposals. Moment-GPT (Xu et al., 2025)

uses off-the-shelf multimodal language models, while GranAlign (Jeon et al., 2026) aligns rewritten queries and captions at different semantic granularities.

Several recent methods attempt to refine the temporal modeling strategy. For example, TAG (Lee et al., 2025) introduces temporal pooling and temporal coherence clustering to aggregate semantically consistent video content, while T2SGrid (Guo et al., 2026) reorganizes temporally ordered frames into spatial grids so that pretrained multimodal models can better capture local temporal dependencies. Self-SiMS (Lee et al., 2026) derives proposals from video self-similarity and subsequently uses query-aware reasoning. However, their visual modeling is still built upon video content sampled or grouped according to predefined temporal layouts. The receptive field of the basic visual representation is therefore not determined by the intrinsic video organization or query semantics.

Topological Analysis and Persistent Homology. Topological data analysis (TDA), particularly persistent homology, provides a principled way to characterize multi-scale structures by tracking how topological features evolve over a filtration (Chazal et al., 2016). Among different homology dimensions, H_0 persistence captures the evolution and merging of connected components, making it particularly suitable for modeling hierarchical connectivity in representation spaces. Recent studies (Moor et al., 2020; You et al., 2026) have incorporated such topological priors into representation learning as training-time constraints. In contrast, our work employs H_0 connectivity as a test-time structural analysis tool, using the induced merge hierarchy to discover adaptive semantic units and evaluate candidate moments for zero-shot video grounding.

3 METHOD

We introduce TOPOATOM, a training-free framework that replaces uniformly predefined temporal units with query-conditioned semantic atoms for zero-shot video grounding.

3.1 PROBLEM AND LOCAL OBSERVATIONS

Given a video V and query Q, video temporal grounding aims at predicting the segment that semantically matches the query. We sample N overlapping base clips X_i with window L and stride Δ, measured in sampled frames. A frozen VLM with visual encoder $\Phi_v(\cdot)$ and text encoder $\Phi_t(\cdot)$ is used to extract video and text embeddings

$$\mathbf{v}_i = \Phi_v(X_i), \qquad \mathbf{q} = \Phi_t(Q), \tag{1}$$

and measure the cross-modal and segment-level relevance as cosine similarities $s(\cdot, \cdot)$ between the obtained features.

Importantly, these uniformly sampled clips are used only as elementary observations for topology construction and semantic atom discovery. They do not constitute the final localization primitives. After atom mining, TOPOATOM switches the basic unit of video analysis from fixed-length clips to a set of adaptive and temporally continuous semantic atoms.

3.2 CONSTRUCTING VIDEO HIERARCHY

The first stage of TOPOATOM aims to recover the multi-scale semantic organization of the video.

***H_0* connectivity filtration.** For every pair of base clips X_i and X_j in the given video, we compute their visual dissimilarity as

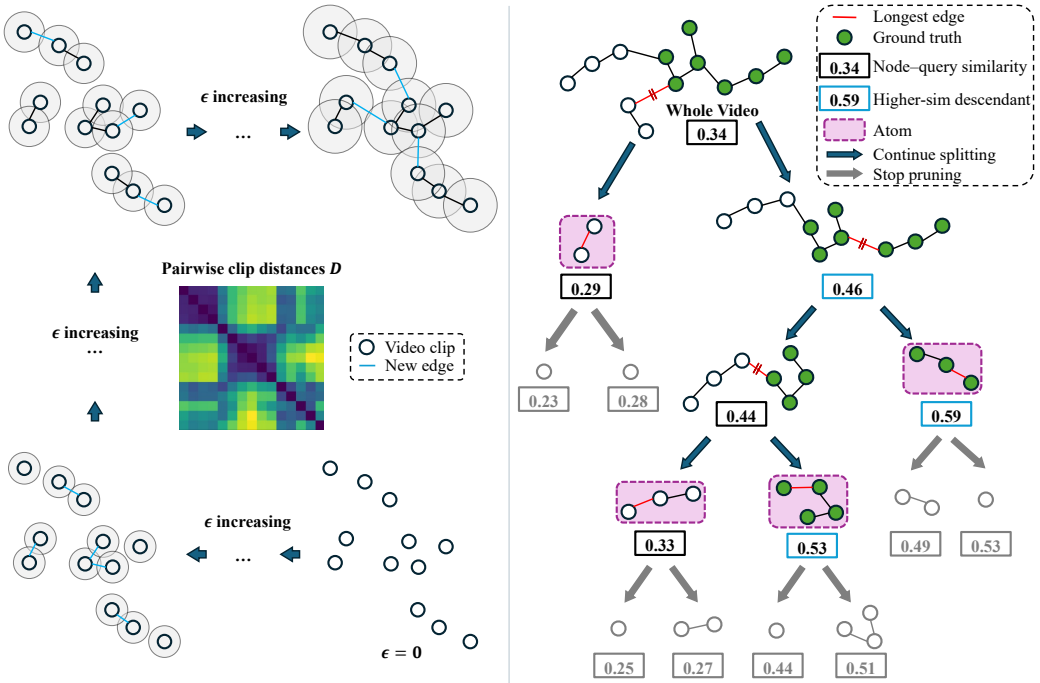
$$D_{ij} = 1 - s(\mathbf{v}_i, \mathbf{v}_j). \tag{2}$$

Based on D, we construct a weighted complete graph $G = (\mathcal{V}, \mathcal{E})$, where each vertex corresponds to a base clip and the weight of edge (i, j) is D_{ij}. For a distance threshold ϵ, we define the thresholded graph as

$$G_\epsilon = (\mathcal{V}, \mathcal{E}_\epsilon), \quad \mathcal{E}_\epsilon = \{(i, j) \mid D_{ij} \leq \epsilon\} \tag{3}$$

At $\epsilon = 0$, each clip forms an isolated connected component. As ϵ gradually increases, new edges appear and previously separated components progressively merge. Tracking this evolution yields the H_0 connectivity filtration of the video.

Since H_0 only depends on connected-component merges, the connected-component evolution of the filtration can be represented compactly by a minimum spanning tree $\mathcal{T} = \mathrm{MST}(G)$ (Carlsson & Mémoli, 2010; Chazal et al., 2016), as illustrated in Figure 2(a). Following the edges of $\mathcal{T}$ in



(a) Bottom-up construction of the H_0 merge hierarchy. (b) Atom mining: query-adaptive top-down pruning

Figure 2: Atom discovery in TOPOATOM. (a) H_0 merge hierarchy construction. As the distance threshold ϵ increases, clips progressively merge into connected components, yielding an H_0 merge hierarchy from pairwise clip distances. (b) Query-adaptive semantic atom mining. We traverse the hierarchy top-down and retain a node as a semantic atom when none of its descendants achieves higher query similarity; otherwise, the node is recursively split. The resulting query-conditioned atoms serve as adaptive primitives for subsequent candidate construction.

ascending order of their weights produces exactly the component-merging events required for H_0 persistence. We organize these events into an H_0 merge hierarchy $\mathcal{H}$.

Each leaf node of $\mathcal{H}$ corresponds to one base clip. Each internal node u represents a connected component formed by merging its child components, and the root represents the entire video. We denote the set of clip indices covered by node u as $\mathcal{I}(u)$. Therefore, $\mathcal{H}$ naturally provides a multi-scale structural decomposition of the video, ranging from individual clips to progressively larger semantic regions.

3.3 QUERY-ADAPTIVE TOP-DOWN SEMANTIC ATOM MINING

The H_0 merge hierarchy captures the intrinsic connectivity structure of the video, but it does not by itself determine the appropriate semantic granularity for grounding. We therefore introduce a query-adaptive top-down pruning procedure that selects a subset of hierarchy nodes as semantic atoms.

Node–query similarity. For each hierarchy node u, the clips in $\mathcal{I}(u)$ may form either one temporally continuous interval or multiple disconnected temporal intervals. We therefore decompose $\mathcal{I}(u)$ into its maximal temporally consecutive runs $\mathcal{R}(u) = \{R_1^u, R_2^u, \ldots, R_{K_u}^u\}$. Let $X(R_k^u)$ denote the video segment covered by run R_k^u. The similarity between query q and node u is then defined as

$$S(u, q) = \frac{1}{K_u} \sum_{k=1}^{K_u} s\big(X(R_k^u), q\big), \tag{4}$$

where $s(\cdot, \cdot)$ denotes the feature similarity.

Although Eq. (4) requires VLM evaluation for hierarchy nodes, it introduces little sequential overhead. The complete merge hierarchy $\mathcal{H}$ is already determined before top-down pruning, and therefore all temporal segments $\{X(R_k^u)\}$ required by every hierarchy node are known in advance. Con-

sequently, all node similarities can be obtained with only a small number of batched forward passes. Inference latency analysis is provided in Appendix A.4.

Top-down pruning criterion. For a hierarchy node u, let $\mathrm{Desc}(u)$ denote the set of all strict descendants of u in $\mathcal{H}$. As presented in Figure 2(b), a node u is retained as a semantic atom when u is a leaf node, or none of its descendants achieves stronger query correspondence than the node itself

$$S(u,q) \geq \max_{v \in \mathrm{Desc}(u)} S(v,q) \tag{5}$$

Intuitively, this indicates that further decomposition does not reveal a more query-relevant substructure. The current component is therefore retained at the current scale under the query-similarity criterion.

Conversely, if $S(u,q) < \max_{v \in \mathrm{Desc}(u)} S(v,q)$, at least one finer-scale descendant provides stronger query correspondence. TOPOATOM therefore continues splitting u and recursively applies the same criterion to its children. This criterion induces a query-dependent semantic resolution. A region is preserved at a coarse scale when its finer structures provide no additional query evidence, while regions containing more discriminative substructures are recursively decomposed to finer scales.

Temporal continuity of final atoms. A retained hierarchy node may contain multiple temporally disconnected runs. We finally decompose these nodes into the continuous components, with each component is treated as an individual semantic atom. The resulting atom set is $\mathcal{A} = \{A_1, \ldots, A_M\}$, ordered chronologically along the video. After this step, all subsequent proposal construction and evaluation operate on $\mathcal{A}$ rather than on the original fixed-length clips.

3.4 CANDIDATE CONNECTIVITY PERSISTENCE

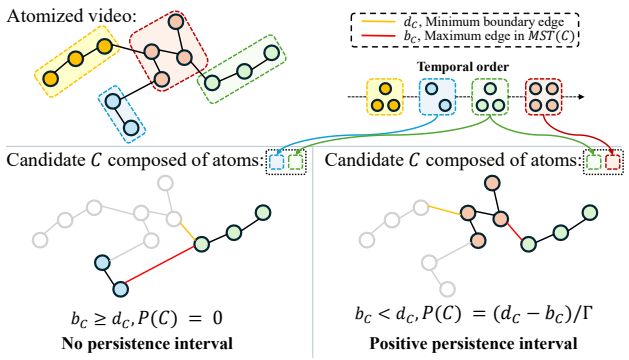


Figure 3: Candidate persistence. For a candidate C, we define its internal connectivity scale b_C as the largest edge weight in $\mathrm{MST}(C)$, and its external attachment scale d_C as the smallest edge connecting C to the rest of the video.

While semantic atoms provide adaptive primitives for query alignment and subsequent candidate construction, query relevance alone is insufficient to guarantee that a candidate forms a structurally coherent region within the video. To address this limitation, we introduce **candidate persistence**, which quantifies the structural integrity of a candidate segment. Specifically, for a candidate C covering a set of underlying base clips $\mathcal{I}_C$, candidate persistence measures how long C exists as an independent connected component during the H_0 filtration, as illustrated in Figure 3.

Internal connectivity scale. By denoting the subgraph of G induced by the clips in $\mathcal{I}_C$ as $G[\mathcal{I}_C]$, we define the internal connectivity scale b_C as the largest edge weight in the minimum spanning tree of this induced subgraph

$$b_C = \max_{e \in \mathrm{MST}(G[\mathcal{I}_C])} w(e). \tag{6}$$

Equivalently, b_C is the smallest threshold at which all clips within candidate C become connected. Thus, b_C characterizes the amount of structural variation that must be tolerated before the candidate forms a complete internal component. A smaller b_C indicates stronger internal compactness.

External attachment scale. We next measure when candidate C first connects to the remaining video. Its external attachment scale is defined as

$$d_C = \min_{i \in \mathcal{I}_C,\ j \notin \mathcal{I}_C} D_{ij}. \tag{7}$$

Equivalently, d_C is the smallest filtration threshold at which any clip inside C becomes connected by an edge to a clip outside C. A large d_C therefore indicates that the candidate is well separated from its surrounding video content.

Table 1: Comparison results on QVHighlights. Best results are indicated in **bold**, second best results are underlined.

Method	Venue	QVHighlights test				QVHighlights val			
		R1,0.5	R1,0.7	mAP,0.5	mAP,avg	R1,0.5	R1,0.7	mAP,0.5	mAP,avg
VideoLLaMa	EMNLP'23	17.1	6.7	18.2	6.2	18.5	6.9	17.8	7.1
VideoChatGPT	ACL'24	21.1	10.2	22.8	9.5	22.4	10.8	21.9	10.3
UniVTG	ICCV'23	25.2	9.0	27.4	10.9	–	–	–	–
Moment-GPT	AAAI'25	58.3	37.7	55.1	35.0	58.9	38.6	55.7	35.9
GranAlign	AAAI'26	59.92	39.3	58.94	38.23	61.94	41.81	59.63	39.12
Self-SiMS	ECCV'26	59.7	42.2	59.2	38.3	61.0	43.2	60.5	39.3
TOPOATOM	-	**62.91**	**42.47**	**59.68**	**39.76**	**63.68**	**43.87**	**60.97**	**40.06**

Together, b_C and d_C define the structural lifetime of candidate C. We define the normalized candidate persistence as

$$P(C) = \frac{\max(0, d_C - b_C)}{\Gamma}, \tag{8}$$

where $\Gamma = \max_{e \in \mathcal{T}} w(e)$ is the largest edge weight of the minimum spanning tree of the complete video. This candidate persistence quantifies the integrity and compactness of internal semantics from a hierarchical perspective, providing a novel and hyperparameter-free candidate connectivity estimation principle complementary to the query alignment. A more detailed definition and property discussion of this candidate persistence $P(C)$ are provided in Appendix B.1. We also analyze the sensitivity of $P(C)$ with respect to small perturbations of edge weights in Appendix B.2.

3.5 ATOM-LEVEL CANDIDATE SCORING AND PREDICTION

Given the adaptively discovered semantic atoms $\mathcal{A} = \{A_1, \ldots, A_M\}$, candidate segments are formed as temporally contiguous subsets of these atoms, yielding $M(M+1)/2$ candidates for M atoms. Note that any subsequent candidate generation and evaluation methods can be seamlessly applied to the atoms discovered by TOPOATOM simply by replacing standard frame- or clip-level units with semantic atoms. However, as the candidate generation strategy is not the primary focus of this paper, we evaluate TOPOATOM using a widely adopted candidate generation and query alignment framework (Zheng et al., 2023a; 2024; Lee et al., 2025).

For a candidate segment C, let $\mathcal{A}_C \subseteq \mathcal{A}$ denote the set of atoms contained in C, and $\overline{\mathcal{A}}_C = \mathcal{A} \setminus \mathcal{A}_C$ denote the background atoms outside C. Because temporal moment localization aims to maximize query relevance within the target moment while minimizing it outside, we define the contrastive atom-level grounding score as:

$$S_{\text{rel}}(C) = \frac{1}{|\mathcal{A}_C|} \sum_{A_m \in \mathcal{A}_C} s(A_m, q) - \frac{1}{|\overline{\mathcal{A}}_C|} \sum_{A_m \in \overline{\mathcal{A}}_C} s(A_m, q). \tag{9}$$

We further integrate the topology-derived candidate persistence term $P(C)$ defined in Eq. (8) to compute the topology-aware candidate score:

$$S(C) = S_{\text{rel}}(C) + \gamma P(C), \tag{10}$$

where γ balances the contribution of structural persistence. Finally, the predicted temporal moment is obtained according to the candidate score (see Appendix A.1 for further details on atomization and candidate strategy).

4 EXPERIMENTS

4.1 BENCHMARKS AND EVALUATION

Datasets and metrics. Charades-STA (Gao et al., 2017), ActivityNet-Captions (Krishna et al., 2017) and QVHighlights (Lei et al., 2021b) are used to evaluate our proposed TOPOATOM. Following previous works, R1,τ is used for evaluation. Charades-STA and ActivityNet-Captions use $\tau \in \{0.3, 0.5, 0.7\}$ and mIoU, the mean rank-1 IoU. QVHighlights uses R1,0.5 and R1,0.7, taking the best overlap with its relevant moments per query. Standard QVHighlights mAP evaluates scored moment lists, and mAP@avg averages thresholds $0.50, 0.55, \ldots, 0.95$ (Lei et al., 2021a).

Table 2: Comparison results on Charades-STA and ActivityNet-Captions. Best results are indicated in **bold**, second best results are underlined.

Method	Venue	Charades-STA				ActivityNet-Captions			
		R1,0.3	R1,0.5	R1,0.7	mIoU	R1,0.3	R1,0.5	R1,0.7	mIoU
UniVTG	ICCV'23	44.09	25.22	10.03	27.12	-	-	-	-
Frozen VLM	WACV'24	56.77	42.93	20.13	37.92	48.28	27.90	11.57	32.37
TFVTG	ECCV'24	67.04	49.97	24.32	44.51	49.34	27.02	13.39	34.10
Moment-GPT	AAAI'25	58.2	38.4	21.6	36.5	48.1	31.1	14.9	30.8
GranAlign	AAAI'26	59.1	39.6	22.7	38.0	50.3	34.0	16.5	33.1
TAG	BMVC'25	67.82	48.58	26.67	45.69	51.88	28.91	15.07	36.55
Self-SiMS	ECCV'26	62.7	39.7	21.0	41.9	49.9	28.2	13.8	34.7
RobustVTG	CVPRW'26	68.63	50.27	**28.92**	46.93	50.25	30.18	13.90	34.49
T2SGrid	CVPR'26	71.4	47.0	20.7	44.9	44.8	26.8	16.1	32.5
TOPOATOM	-	**72.13**	**50.79**	27.58	**47.18**	**53.82**	**35.66**	**16.68**	**37.74**

Implementation. We evaluate TOPOATOM based on frozen Qwen3-VL-Embedding 2B (Qwen Team, 2026). Default clips use a 5-frame window and 2-frame stride, sampled at 1 fps on Charades-STA and 2 fps on QVHighlights and ActivityNet-Captions. The weight of candidate persistence $\gamma = 0.5$ is fixed for all datasets. Experiments run on one NVIDIA RTX 3090 Ti GPU.

4.2 COMPARISON WITH PRIOR WORK

We evaluate TOPOATOM against a broad range of representative zero-shot video grounding methods, including VideoLLaMa (Zhang et al., 2023), VideoChatGPT (Maaz et al., 2024), UniVTG (Lin et al., 2023), Frozen VLM (Luo et al., 2024), TFVTG (Zheng et al., 2024), Moment-GPT (Xu et al., 2025), GranAlign (Jeon et al., 2026), TAG (Lee et al., 2025), Self-SiMS (Lee et al., 2026), RobustVTG (Banditakkarakul et al., 2026), and T2SGrid (Guo et al., 2026). Table 1 presents the comparison on QVHighlights, while Table 2 reports the results on Charades-STA and ActivityNet-Captions.

TOPOATOM achieves the best results on 15 out of the 16 reported metrics. These consistent improvements across all three benchmarks strongly support our central hypothesis: grounding over query-conditioned semantic units provides a more effective representation than relying exclusively on uniformly predefined clips.

Furthermore, recent approaches are typically built upon significantly larger MLLMs/LLMs compared to our 2B TOPOATOM—for instance, Self-SiMS uses an 11B LLaMA-3.2, GranAlign employs at least an 8B LLaMA3 alongside a Qwen2.5-VL-7B, and T2SGrid adopts a Qwen3-VL-8B. Moreover, many of these methods rely on computationally expensive MLLM inference at test time to rewrite queries or generate video captions (e.g., Moment-GPT, GranAlign). In contrast, TOPOATOM solely utilizes a 2B Qwen3-VL-Embedding backbone to extract multimodal features, requiring no test-time MLLM inference such as captioning or query rewriting. Although models like TAG (Lee et al., 2025) and RobustVTG (Banditakkarakul et al., 2026) use a smaller 1.2B BLIP-2 backbone, TOPOATOM substantially outperforms them across Charades-STA, ActivityNet-Captions, and TACoS (detailed in Appendix A.2). A corresponding controlled study in Table 3 further confirms that the superiority of TOPOATOM is genuinely driven by query-conditioned adaptive atomization, rather than solely attributable to stronger backbone features.

4.3 ABLATION STUDY

We conduct component-wise ablations on Charades-STA to analyze the contribution of Atom Mining (AM), Contrastive Grounding Score (CGS), and Candidate Persistence (CP). The results are reported in Table 3. To evaluate the candidate quality, we additionally report oracle performance, where the candidate with the highest IoU with the ground truth is selected.

Effect of AM. The first row removes atom mining and directly enumerates candidates over the original clips using CGS for ranking. This configuration achieves very high oracle performance but substantially lower localization accuracy.

We further restrict the H_0 construction to a temporal-adjacent graph, denoted as "TG only", where edges are included only between temporally neighboring clips. The localization and oracle performances are slightly below the full model. Consequently, constructing the merge hierarchy from the

Table 3: Components contribution study conducted on Charades-STA. AM: Atom Mining (TG only: Temporal-adjacent Graph only; Clustering: TAG-style video clip clustering), CGS: Contrastive Grounding Score (Clip: clip-based CGS), CP: Candidate Persistence.

AM	CGS	CP	Performance			Oracle Performance		
			R1,0.5	R1,0.7	mIoU	R1,0.5	R1,0.7	mIoU
-	✓	-	42.56	17.23	35.37	**99.27**	**98.29**	**96.35**
✓	-	✓	33.63	15.52	23.94	89.17	74.07	77.03
✓	✓	-	48.80	25.02	46.10	89.17	74.07	77.03
✓	Clip	✓	50.22	26.41	46.30	89.17	74.07	77.03
TG only	✓	✓	50.32	26.50	46.28	87.62	73.81	75.60
Clustering	✓	✓	49.24	26.06	46.42	82.49	71.91	74.09
✓	✓	✓	**50.79**	**27.58**	**47.18**	89.17	74.07	77.03

Table 4: Ablation study on hyper-parameter γ, which controls the weight of the candidate persistence term. We report model performance and Top-1 candidate change statistics for $\gamma > 0$. Notably, some Top-1 ranking shifts leave the IoU with respect to ground truth unchanged.

γ	Performance			Top-1 Candidate Analysis ($\gamma > 0$)				
	R1,0.5	R1,0.7	mIoU	Changed	Improved	Worsened	Improvement	Worsening
0	48.80	25.02	47.10	–	–	–	–	–
0.1	49.91	25.90	**47.31**	17.9%	8.2%	7.3%	0.148	0.139
0.3	50.22	26.41	47.30	33.9%	15.6%	14.6%	0.154	0.151
0.5	**50.79**	**27.58**	47.18	42.7%	19.6%	19.0%	0.163	0.164
0.7	50.16	27.01	46.76	48.1%	21.5%	22.1%	0.167	0.177

complete graph provides a more informative structural representation than relying exclusively on temporal adjacency.

To further verify the superiority of TOPOATOM over clustering for video content parsing, we implemented the clustering method following TAG (Lee et al., 2025) and its official repository. The results show degradation in both localization accuracy and proposal quality. Compared with clustering with a fixed number of clusters, TOPOATOM benefits from its adaptive atomization governed by the inherent topological structure of videos and query semantics.

Effect of atom-level CGS. Removing CGS while retaining AM and CP causes a substantial performance drop, since CP measures whether a candidate forms a structurally stable component in the video, but it does not directly measure whether that component matches the query.

We additionally replace atom-level CGS with a clip-based variant, where the similarity of each atom is approximated by averaging the query similarities of its constituent base clips. This reduces all the metrics. The result supports the motivation of TOPOATOM: encoding the coherent semantic region jointly as an atom preserves richer temporal context.

Effect of candidate persistence. Removing CP decreases all the metrics. Candidate persistence complements the cross-modal similarity by favoring candidates that are internally coherent and structurally separated from their surrounding video content.

Table 4 further summarizes the ablation results of the persistence weight γ. Overall performance first improves and then slightly declines as γ increases, while the fraction of re-ranked Top-1 candidates grows monotonically. Although the values for *Improved* and *Improvement* in Table 4 do not substantially exceed those for *Worsened* and *Worsening*, candidate persistence consistently boosts overall localization performance. For instance, even at $\gamma = 0.7$, where *Improved* and *Improvement* fall below *Worsened* and *Worsening*, the overall R1 metrics still outperforms the baseline ($\gamma = 0$). This seemingly counter-intuitive result indicates that candidate persistence achieves a *selective enhancement*: it preferentially boosts high-confidence candidates that already exhibit superior IoU performance. A more detailed analysis of the candidate persistence distribution across various candidates is provided in Appendix A.3.

Effect of the Base Vision–Language Model. Table 5 evaluates the generalization and scalability of TOPOATOM across various frozen VLMs, including InternVideo (Wang et al., 2022) and Qwen3-VL-Embedding variants. Compared to the baseline (CGS only), TOPOATOM consistently yields

Table 5: Performance gains on different VLMs.

Method	InternVideo		Qwen3-VL 2B		Qwen3-VL 8B	
	R1,0.5	mIoU	R1,0.5	mIoU	R1,0.5	mIoU
CGS only	38.03	34.20	42.56	35.37	43.76	36.81
TOPOATOM	**48.89**	**45.29**	**50.79**	**47.18**	**51.39**	**47.62**

Table 6: Effect of window and stride.

L / Δ	R1,0.5	R1,0.7	mIoU
3 / 1	50.13	26.98	46.75
5 / 2	**50.79**	**27.58**	**47.18**
5 / 4	50.25	26.35	46.22
7 / 3	50.73	26.76	46.89

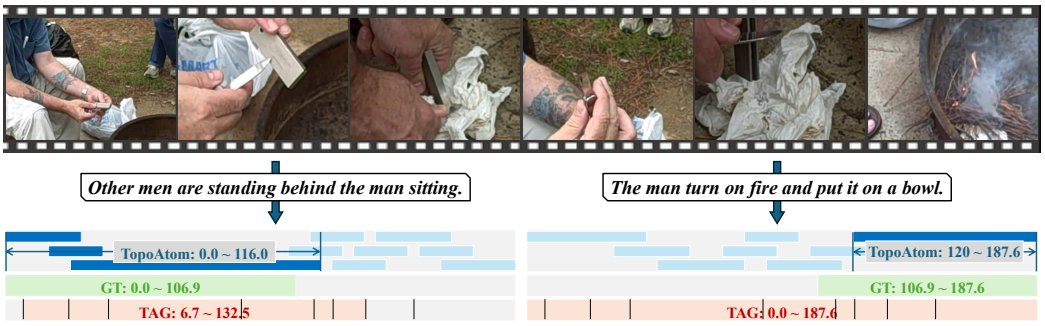


Figure 4: Qualitative visualization of query-conditioned semantic atom discovery. Given different text queries for the same video, TOPOATOM adaptively atomizes video content at flexible granularities (light blue bars: discovered semantic atoms; dark blue bars: atoms used in the final prediction). In contrast, the baseline TAG applies query-agnostic segmentation with indiscriminate boundaries. TOPOATOM maintains a coarser atom for background-focused context (left) while generating finer-grained atoms to capture specific foreground actions and objects (right).

substantial performance improvements. These persistent gains validate that our proposed semantic atomization mechanism is model-agnostic. Furthermore, performance steadily scales up when using stronger backbones like Qwen3-VL 8B (Qwen Team, 2026).

TOPOATOM is specifically designed to exploit the ability of modern VLMs to jointly encode temporally continuous visual content. For this reason, we evaluate backbones that support multi-frame or video inputs. Using image-only VLM to compute the internal-node scores as strictly positive convex combinations of descendant leaf scores will eventually degenerate the query-adaptive pruning toward treating individual frames as atoms, details are provided in Appendix B.3.

Hyperparameter sensitivity. Table 6 studies the sensitivity of TOPOATOM to the initial clip sampling strategy. Overall, the performance remains stable across the tested configurations, indicating that TOPOATOM is not highly sensitive to these hyperparameters.

4.4 QUALITATIVE ANALYSIS

As illustrated in Figure 4, TOPOATOM demonstrates a clear advantage in query-conditioned adaptive granularity for video atomization. For a query focusing on background context ("*Other men are standing behind the man sitting*"), TOPOATOM preserves the relevant segment as a coarse semantic atom, avoiding the indiscriminate over-segmentation observed in TAG. Conversely, when presented with a query involving specific foreground interactions ("*The man turn on fire and put it on a bowl*"), TOPOATOM adaptively increases the temporal granularity to accurately model fine-grained events and key semantic entities (e.g., the bowl). These behaviors highlight the capability of TOPOATOM to dynamically explore video semantics according to varying textual queries. More qualitative analyses are provided in Appendix A.5.

5 CONCLUSION

We presented **TOPOATOM**, a training-free framework for zero-shot video temporal grounding that replaces uniformly predefined frames or clips with query-conditioned semantic atoms. By constructing an H_0 merge hierarchy and performing query-adaptive top-down pruning, TOPOATOM discovers variable-length semantic atoms that better preserve the intrinsic temporal structure of video content. We further introduced candidate persistence to quantify the structural stability of atom-composed moments and combine it with atom-level semantic relevance for final prediction. The experimental results support a broader perspective for zero-shot video grounding: the temporal primitives used for localization should be adaptively discovered from the video structure and the query, rather than fixed in advance.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure (e.g., generating synthetic data sets, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, designing methodology, implementing methods, or interpreting results). Additionally, we used generative AI tools for polishing parts of the manuscript to improve readability. We have reviewed all AI-assisted work. Specifically, all AI-edited text was carefully checked by the authors word-by-word to ensure that the original meaning is strictly preserved. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nutthadech Banditakkarakul, Bo Chen, and Stephen Gould. Towards robust zero-shot video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11143–11152, 2026.

Gunnar Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 11:1425–1470, 2010. URL https://jmlr.org/papers/v11/carlsson10a.html.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. Springer, 2016. doi: 10.1007/978-3-319-42545-0. URL https://arxiv.org/abs/1207.3674.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007. doi: 10.1007/s00454-006-1276-5.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017. URL https://openaccess.thecvf.com/content_ICCV_2017/html/Gao_TALL_Temporal_Activity_ICCV_2017_paper.html.

Chaohong Guo, Yihan He, Yongwei Nie, Fei Ma, Xuemiao Xu, and Chengjiang Long. T2SGrid: Temporal-to-spatial gridification for video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3443–3454, June 2026.

Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18908–18918, 2023.

Mingyu Jeon, Sunjae Yoon, Jonghee Kim, and Junyeong Kim. GranAlign: Granularity-aware alignment framework for zero-shot video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5287–5295, 2026.

Minjoon Jung, Youwon Jang, Seongho Choi, Joochan Kim, Jin-Hwa Kim, and Byoung-Tak Zhang. Background-aware moment detection for video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. URL https://arxiv.org/abs/2306.02728.

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. URL https://arxiv.org/abs/1705.00754.

Jihyun Lee, Cheol-Ho Cho, Woojin Jun, Woojin Jeong, and Jae-Pil Heo. Mitigating modality and language-style gaps for zero-shot video moment retrieval. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.19027.

Jin-Seop Lee, SungJoon Lee, Jaehan Ahn, YunSeok Choi, and Jee-Hyong Lee. TAG: A simple yet effective temporal-aware approach for zero-shot video temporal grounding. In *British Machine Vision Conference*, 2025. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_786/paper.pdf.

Jie Lei, Tamara L. Berg, and Mohit Bansal. QVHighlights evaluation and submission: Official evaluator, 2021a. URL https://github.com/jayleicn/moment_detr/tree/main/standalone_eval. Accessed September 14, 2026.

Jie Lei, Tamara L. Berg, and Mohit Bansal. QVHighlights: Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, 2021b. URL https://arxiv.org/abs/2107.09609.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. UniVTG: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023.

Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Zero-shot video moment retrieval from frozen vision-language models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5464–5473, 2024.

Zezhong Lv and Bing Su. Variational global clue inference for weakly supervised video moment retrieval. *Knowledge-Based Systems*, 311:113071, 2025. doi: 10.1016/j.knosys.2025.113071.

Hongxu Ma, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. MS-DETR: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4514–4523, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7045–7054. PMLR, 2020.

Nuetee et al. Official code repository for tag. https://github.com/Nuetee/TAG, 2023.

Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1847–1856. IEEE, 2024.

Qwen Team. Qwen3-VL-Embedding: Official models and implementation, 2026. URL https://github.com/QwenLM/Qwen3-VL-Embedding. Accessed September 14, 2026.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. InternVideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. URL https://arxiv.org/abs/2212.03191.

Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Ming Li, Wenxin Liang, Yang Li, and Sidan Du. Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8978–8986, 2025.

Junwon You, Dasol Kang, and Jae-Hun Jung. Topological alignment of shared vision-language embedding space. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1207–1215. PMLR, 2026.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023.

Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the ACM International Conference on Multimedia*, 2020. URL https://arxiv.org/abs/2008.08257.

Minghang Zheng, Shaogang Gong, Hailin Jin, Yuxin Peng, and Yang Liu. Generating structured pseudo labels for noise-resistant zero-shot video sentence localization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14197–14209, 2023a.

Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision*, pp. 20–37. Springer, 2024.

Minghang Zheng et al. Official code repository for tfvtg. https://github.com/minghangz/TFVTG, 2023b.

A Additional experimental results

A.1 Implementation details

Atomization. In our experiments, considering that even local video semantics require adequate context to be expressed more accurately, we extract local video representations using overlapping clips as the foundation for video topological structure analysis. Under this strategy, adjacent clips share sampled frames, so temporally neighboring semantic atoms also contain overlapping regions. Overlapping windows provide temporal context for feature extraction. The merge hierarchy is constructed solely from pairwise feature distances, without an additional edge term based on temporal overlap.

Candidate generation and evaluation. Suppose for an atom $A_i \in \mathcal{A} = \{A_1, \ldots, A_M\}$, whose start and end timestamps are denoted as A_i^s and A_i^e, respectively; then the start and end timestamps of a candidate composed of atoms from the l-th to the r-th are A_l^s and A_r^e. For all candidate segments in the candidate pool $\mathcal{C}$, we first compute their base candidate scores $S(C)$ via Eq. (10). Following previous candidate-based temporal grounding practices (Nuetee et al., 2023; Zheng et al., 2023b) that evaluate query alignment by taking the contrastive difference between intra-segment and extra-segment query evidence, we perform min-max normalization on these base scores across all candidates

$$\bar{S}(C) = \frac{S(C) - \min_{C' \in \mathcal{C}} S(C')}{\max_{C' \in \mathcal{C}} S(C') - \min_{C' \in \mathcal{C}} S(C')}. \tag{11}$$

To further enhance candidate scoring through structural overlap consensus, the final prediction score for the i-th candidate C_i is aggregated as follows

$$S_{\text{final}}(C_i) = \bar{S}(C_i) \sum_{C_j \in \mathcal{C} \setminus C_i} \text{IoU}(C_i, C_j)^{\eta}, \tag{12}$$

where $\text{IoU}(C_i, C_j)$ denotes the temporal Intersection-over-Union between candidates C_i and C_j, and η is a scaling exponent controlling overlap sensitivity. The final temporal segment prediction is performed according to S_{final}. To ensure a strictly fair and consistent evaluation, we adopt this exact post-processing protocol with the identical parameter setting ($\eta = 0.6$).

A.2 Performance evaluation on challenging long videos

Table 7: Performance on TACoS dataset.

Methods	R1,0.3	R1,0.5	R1,0.7	mIoU
ChatVTG (Qu et al., 2024)	8.12	3.74	1.25	5.54
Frozen (Luo et al., 2024)	11.20	5.57	-	-
TFVTG (Zheng et al., 2024)	16.40	9.92	4.22	11.88
TAG (Lee et al., 2025)	17.80	9.57	4.12	13.13
RobustVTG (Banditakkarakul et al., 2026)	21.64	12.45	6.75	15.45
TopoAtom	**30.48**	**15.21**	**7.09**	**21.59**

We further conduct evaluations on the TACoS dataset (Regneri et al., 2013), where videos have longer average durations and the target segments exhibit smaller visual differences from the background. The comparison results are presented in Table 7.

A.3 Distribution of candidate persistence

To further demonstrate the candidate persistence property in TopoAtom, we analyze the distribution of candidate persistence for top-1 predictions and oracle candidates, as illustrated in Figure 5. Over half of the candidates in both categories yield candidate persistence > 0. This suggests candidate persistence enables extra hierarchical structure assessment beyond query-candidate semantic matching, and the two are complementary.

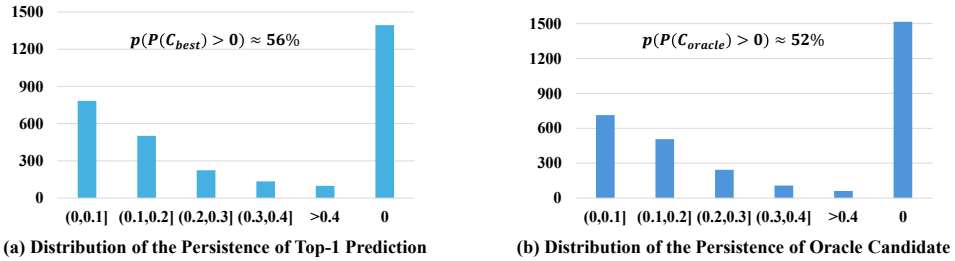


Figure 5: Distributions of the candidate persistence.

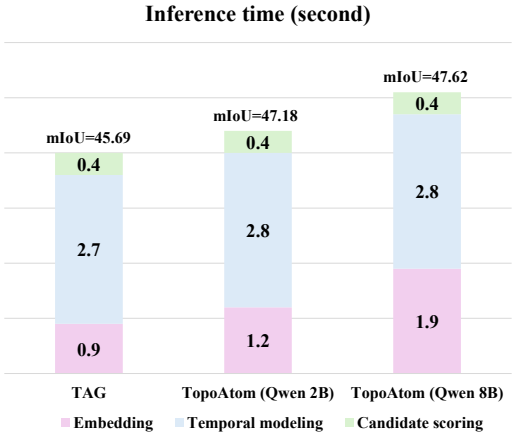


Figure 6: Runtime is decomposed into feature embedding, temporal modeling, and candidate scoring. TAG is faster mainly due to its substantially lower embedding cost, while the differences in temporal modeling and candidate scoring are relatively small. TOPOATOM achieves higher localization accuracy at the cost of additional VLM encoding, with Qwen3-VL-8B reaching the best mIoU.

A.4 INFERENCE EFFICIENCY

Figure 6 compares the inference cost of TAG and TOPOATOM under different VLM backbones on Charades-STA. The main efficiency gap comes from the embedding stage: TAG requires only 0.9s for feature extraction, whereas TOPOATOM takes 1.2s with Qwen3-VL-2B and 1.9s with Qwen3-VL-8B due to the additional encoding of hierarchy-induced video segments. In contrast, the cost of temporal modeling is comparable, ranging from 2.7s for TAG to 2.8s for TOPOATOM, while candidate scoring remains negligible at about 0.4s for all settings. Thus, the extra runtime of TOPOATOM is dominated by VLM embedding rather than the proposed topological modeling or candidate evaluation, and the additional computation is accompanied by consistent gains in localization accuracy.

A.5 ADDITIONAL QUALITATIVE ANALYSIS

Figure 7 provides qualitative visualization of the candidate generation and clip feature distribution. This example illustrates how query-conditioned atom mining can suppress large irrelevant regions and construct candidate moments from locally coherent semantic content, leading to more precise temporal localization.

Figure 8 illustrates both a successful example and a representative failure case. In Figure 8(a), TOPOATOM effectively groups locally coherent content into semantic atoms and predicts 0.0–10.0s, closely matching the ground-truth interval of 0.0–9.5s, while TAG covers the entire video. Figure 8(b) shows a more challenging case. TOPOATOM successfully extracts several relatively independent semantic atoms and substantially narrows the prediction compared with TAG. However, the "man" specified in the query appears only intermittently throughout the relevant portion of the video and is not consistently visible even inside the ground-truth interval. As a result, the query-

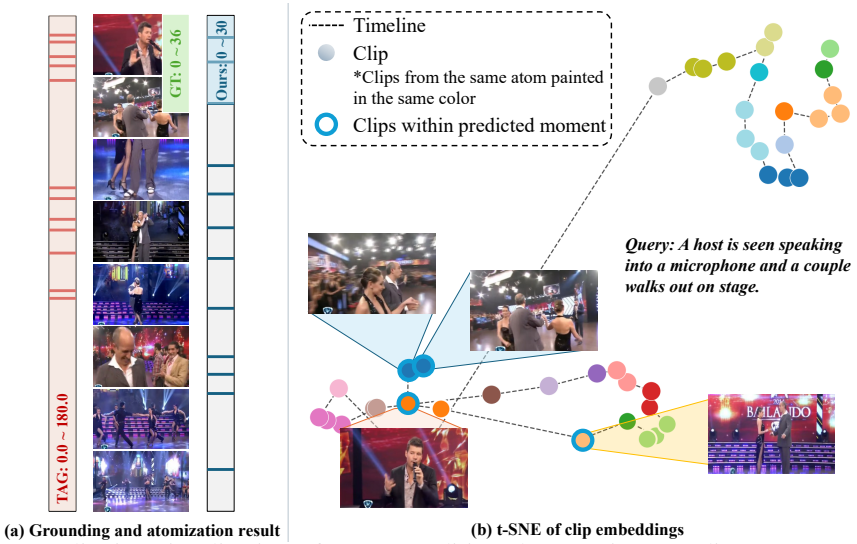


Figure 7: Qualitative visualization of query-conditioned semantic atom discovery. (a) Compared with TAG, TOPOATOM produces more reasonable video splitting and a substantially more compact prediction aligned with the ground-truth moment. The boundaries of the TAG clusters and TOPOATOM atoms are indicated by the lines in their bars, respectively. (b) A t-SNE visualization of clip features, with clips from the same discovered atom shown in the same color.

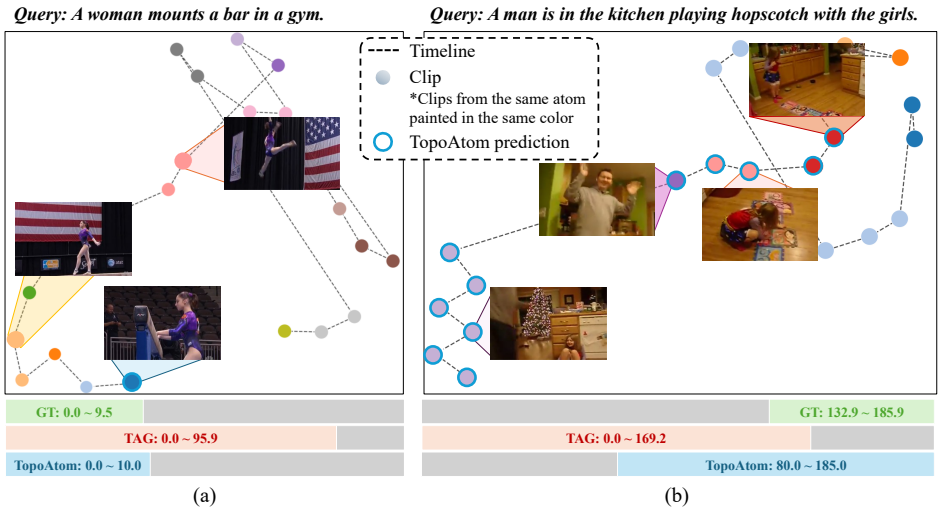


Figure 8: Clips assigned to the same semantic atom are shown in the same color, and blue outlines indicate clips covered by the TOPOATOM prediction. (a) TOPOATOM accurately localizes the target moment by composing query-relevant atoms into a compact prediction. (b) Failure case: although TOPOATOM still produces a better prediction than TAG, the queried "man" appears sparsely and intermittently, even within the ground-truth interval, making it difficult to further isolate a more compact query-relevant segment.

conditioned evidence does not provide a sufficiently stable signal for further atom refinement, and TOPOATOM retains a relatively broad prediction, leading to limited IoU with the ground truth. This case highlights a limitation of the current atom-mining criterion when query-critical entities occur sparsely or discontinuously over time.

B THEORETICAL ANALYSIS

B.1 CANDIDATE PERSISTENCE

Consider a complete finite graph with nonnegative symmetric weights, fixed vertices present at scale zero, and edge inclusion at $w(e) \le \epsilon$. Let $I = \mathcal{I}_C$ be a nonempty proper vertex set. Define b_C and d_C as in the main text, including $b_C = 0$ for a singleton.

Theorem 1 (Lifetime of an exact component set)**.** *For every $\epsilon \ge 0$, I is exactly one connected component of G_ϵ if and only if $b_C \le \epsilon < d_C$. Hence its independent existence interval has length $(d_C - b_C)_+$.*

Proof. For $|I| > 1$, the induced graph is connected at threshold ϵ if and only if it contains a spanning tree with maximum edge weight at most ϵ. An MST minimizes this bottleneck, so the least such threshold is b_C. A singleton is internally connected from zero. The set is externally isolated precisely when no crossing edge has appeared, equivalently $\epsilon < d_C$. Internal connectivity and external isolation together characterize an exact connected component. □

Relation to standard persistence. All ordinary H_0 classes are born at zero in this filtration, and finite deaths occur at component merges. A merged component does not create a new homology class. For example, if two vertices connect at weight 1 and their set connects to a third at weight 3, the set has interval $[1, 3)$, but the finite barcode intervals are $[0, 1)$ and $[0, 3)$. Our b_C is the formation scale of an exact set, not a homology-class birth. The terminology "candidate persistence" refers to this explicitly defined set lifetime (Chazal et al., 2016).

Corollary 1 (Hierarchy laminarity)**.** *Any two vertex sets with positive exact-component lifetimes are either nested or disjoint.*

Proof. Components can only merge as the threshold grows, so intersecting sets must be nested. □

Consequently, persistence is a selective structural bonus.

B.2 SENSITIVITY OF A FIXED CANDIDATE

Stability results for persistence motivate examining sensitivity to tiny perturbations (Cohen-Steiner et al., 2007).

Proposition 1 (Edge perturbation bound)**.** *For two weight matrices W, W' on the same vertices with $\max_{i<j} |W_{ij} - W'_{ij}| \le \delta$, fix a proper candidate index set I. Writing $p_C = (d_C - b_C)_+$, we have*

$$|b_C - b'_C| \le \delta, \qquad |d_C - d'_C| \le \delta, \qquad |p_C - p'_C| \le 2\delta.$$

If $\Gamma, \Gamma' \ge g_0 > 0$, then $|P(C) - P'(C)| \le 3\delta/g_0$.

Proof. The bottleneck representation $b_C = \min_\tau \max_{e\in\tau} w(e)$, over spanning trees of the induced graph, is 1-Lipschitz in the edge sup norm. The cut minimum d_C is also 1-Lipschitz. The positive-part map is nonexpansive, proving the first three bounds. Applying the same bottleneck argument to the full graph gives $|\Gamma - \Gamma'| \le \delta$. Since $p'_C \le \Gamma'$,

$$\left|\frac{p_C}{\Gamma} - \frac{p'_C}{\Gamma'}\right| \le \frac{|p_C - p'_C|}{\Gamma} + \frac{p'_C|\Gamma - \Gamma'|}{\Gamma\Gamma'} \le \frac{3\delta}{g_0}.$$

□

Thus $d_C - b_C > 2\delta$ suffices to preserve positive persistence under these perturbations.

B.3 What scalar averaging implies

Proposition 2 (Degeneracy under convex averaging of leaf scores)**.** *Suppose every internal node is scored as a convex combination with strictly positive weights of its descendant leaf scores, and each leaf is scored by its own value. A non-leaf node is retained by Eq.* (5) *if and only if all its descendant leaf scores are equal. If all leaf scores are pairwise distinct, every retained node is a leaf.*

Proof. A strictly positive weighted average of nonidentical values is strictly below their maximum. A maximizing leaf is a strict descendant, so such an internal node is split. □