Keeping the Evidence Chain: Semantic Evidence Allocation for Training-Free Token Pruning in Video Temporal Grounding

Anonymous ECCV 2026 Submission

Paper ID #5912

Abstract. Video Temporal Grounding (VTG) localizes the temporal boundaries of a query-relevant moment in long, untrimmed videos, making video-language-model pipelines prohibitively expensive. While recent training-free visual token pruning has shown success in video question answering, naively applying these objectives to VTG often causes drastic degradation, as VTG crucially depends on boundary-sensitive evidence and cross-frame reasoning chains. We therefore identify two VTG-specific pruning principles: Evidence Retention (ER), which keeps query-critical patches especially around event boundaries, and Connectivity Strength (CS), which preserves token-level cross-frame connectivity for long-range evidence aggregation. Building on these insights, we propose SemVID, a training-free pruning framework that constructs a compact yet coherent token subset with complementary semantic roles. SemVID first allocates per-frame token budgets by balancing query relevance and inter-frame variation to avoid over-pruned or token-empty segments, and then selects three types of tokens: object tokens for diverse query-critical evidence, motion tokens to capture meaningful transitions and serve as cross-frame relays, and a small set of context tokens for scene continuity. Extensive experiments on VTG benchmarks show that SemVID achieves a strong accuracy-efficiency trade-off, retaining up to 95.4% mIoU with only 12.5% visual tokens and delivering up to a 5.8× prefill speedup, consistently outperforming prior methods under the same budgets.

Keywords: Video Temporal Grounding · Visual Token Pruning

1 Introduction

Video Temporal Grounding (VTG) aims to localize the start and end timestamps of a moment in an untrimmed video that matches a language query [10, 29, 39, 54]. As a core capability for practical video interaction, VTG supports moment retrieval, highlight discovery, and query-driven video summarization, where users need to quickly jump to the exact moment of interest [24, 26]. Recently, VTG methods have begun to leverage Video-Language Models (VLMs), benefiting from their strong cross-modal understanding and reasoning ability, and have achieved promising performance on diverse VTG benchmarks [8, 24].

Despite recent progress, deploying VLM-based VTG remains expensive. A video is typically tokenized into thousands of patch tokens, and the attention cost

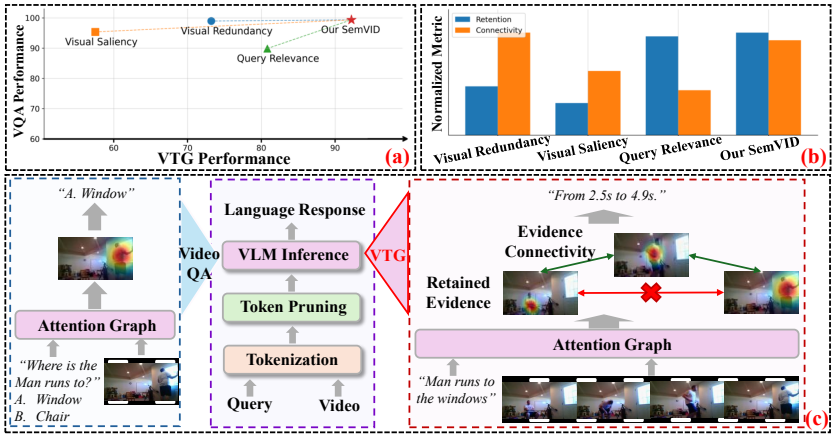

Fig. 1: Comparison between existing pruning objectives and SemVID for VTG. **(a)** Performance comparison between VTG and VideoQA tasks. **(b)** Diagnostics of pruning objectives on evidence retention and cross-frame connectivity. **(c)** VTG requires long-range evidence aggregation rather than a single informative frame. SemVID preserves both query-critical evidence and transition relays to connect evidence across frames.

scales quadratically with sequence length [11,35,38]. This challenge is amplified for long videos: precise boundary localization often requires dense sampling [40], yet increasing the sampling rate quickly makes prefill consumption prohibitive.

As described in [17,18], many visual tokens are redundant and contribute little to performance and thus can be safely removed without compromising accuracy. This naturally motivates training-free visual token pruning to reduce computation. However, pruning strategies designed for VTG remain under-explored. In practice, existing training-free pruning baselines are borrowed from Video Question Answering (VideoQA) and can be categorized by their objectives into Visual Redundancy (VR), Visual Saliency (VS), and Query Relevance (QR).

A straightforward attempt is to directly apply these VideoQA pruning objectives to VTG. While effective for perception-oriented tasks (e.g., object/attribute recognition) that can often be answered from a single informative frame [13,20, 43], VTG fundamentally differs: it requires temporally coherent evidence to localize event boundaries and to reason about how events evolve over time [8,44]. As a result, naively transferring VideoQA pruning tends to discard temporally critical cues and leads to severe performance drops, as shown in Fig. 1(a).

This mismatch can be understood through the lens of VTG-specific requirements. VR removes duplicate content by merging or discarding visually similar tokens. While effective for compression, its query-agnostic criterion can suppress small yet decisive evidence near event boundaries [4]. VS prioritizes salient regions, but saliency often concentrates tokens on a few standout frames, leaving large portions of the timeline underrepresented [34]. This produces insufficient temporal coverage and weakens the evidence continuity, making it difficult to track evolving events. QR keeps tokens most similar to the query. However, it

often repeatedly selects the same locally relevant regions, producing fragmented glimpses that miss the state transitions and disrupt cross-frame reasoning [9,19].

Motivated by these observations, we argue that effective pruning for VTG should satisfy two objectives: **Evidence Retention (ER)** and **Connectivity Strength (CS)**. ER aims to preserve query-critical evidence patches, especially those around temporal boundaries. CS requires that the retained tokens remain well connected across frames so evidence can be aggregated along the video timeline. From an attention-graph perspective, query-conditioned signals are extracted through cross-attention and propagated across frames and layers via stacked self-attention as multi-hop message passing [1]. Pruning alters the graph topology by removing nodes and edges. If boundary-critical evidence tokens or intermediate relay tokens are dropped, the evidence chain becomes fragmented, impeding long-range temporal aggregation and degrading grounding accuracy [7,53]. The more details are demonstrated in Fig. 1(c).

To address the aforementioned challenges, we propose **SemVID**, a training-free pruning framework tailored for VTG. SemVID explicitly optimizes ER and CS by constructing a compact set of tokens with complementary semantic roles. Concretely, SemVID first assigns each frame a token budget that balances query relevance and inter-frame variation, preventing empty or over-pruned segments in long videos. It then identifies three types of tokens: (i) **object tokens** that preserve diverse query-aligned evidence for ER; (ii) **motion tokens** that capture meaningful temporal changes and act as cross-frame relays for CS; and (iii) a small number of **context tokens** as stable anchors to maintain scene continuity. As reflected in Fig. 1(b), SemVID achieves strong performance on both ER and CS, and accordingly, these role-aware tokens form a compact yet coherent evidence chain that keeps evidence both present and traceable while substantially reducing visual tokens (Fig. 1c). Extensive experiments demonstrate that SemVID achieves a strong accuracy-efficiency trade-off on VTG, retaining up to 95.4% mIoU with only 12.5% tokens while delivering a 5.8× prefill speedup, outperforming prior training-free pruning objectives under the same token budget.

Our contributions are as follows:

- We identify that VTG-oriented pruning should follow objectives, Evidence Retention (ER) and Connectivity Strength (CS), which are crucial for preserving boundary-critical evidence and maintaining cross-frame reasoning chains.
- We propose SemVID, a training-free pruning framework tailored for VTG, which explicitly optimizes ER and CS by constructing a role-aware token set that preserves query-critical evidence and maintains cross-frame connectivity.
- We demonstrate a strong accuracy-efficiency trade-off on VTG benchmarks, consistently outperforming prior methods under the same token budgets.

2 Related Work

Training-Free Pruning for VLMs. Modern VLMs tokenize videos into dense patch tokens [11], resulting in long visual sequences and quadratic attention cost [38]. This motivates training-free token pruning to accelerate [48]. However,

pruning for VTG is particularly challenging. Unlike coarse VideoQA, VTG requires long-range evidence that is both retained and temporally traceable [8,44]. Aggressive pruning can appear safe on coarse QA yet fail on localizations [12].

Visual Redundancy. VR reduces duplicate content by merging or discarding redundant tokens. PruneVid [16] clusters frames into scenes and compresses static tokens within each scene. FastVID [34] further preserves spatiotemporal structure during compression. ToMe [4] merges similar tokens around fixed anchors. TokenSculpt [28] introduces structure-aware merging to better respect video geometry. However, VR is typically query-agnostic and tends to remove boundary-sensitive transition cues. We mitigate this via role-aware token allocation, preserving query-critical objects and transitions to safeguard evidence.

Visual Saliency. VS ranks tokens by saliency or attention and keeps the top ones [9,34,52]. FastVID computes saliency by vision-encoder attention [34]. However, attention-based scores can be unstable and biased by attention sinks [23,45], leading to the selected tokens not corresponding to semantically informative regions [15, 42, 48]. Importantly, VS often concentrates tokens on a few dominant frames and causes token-empty gaps that break evidence chains. We address this with a lightweight budget-allocation prior that reserves per-frame tokens to maintain temporal coverage and continuity.

Query Relevance. QR conditions retention on the query, typically by query-token similarity or cross-modal attention. PruneVid combines redundancy reduction with question-guided selection [16]. IVTP first estimates intra-vision importance and then filters tokens with instruction-related semantics [15]. LGTTP increases token density for temporally relevant segments based on query cues [19]. Despite their appeal, attention-based relevance can be biased by attention sinks and may even underperform simple baselines [41, 42, 52]. Furthermore, it tends to extract scattered local patches across frames, leading to fragmented evidence and weakened cross-frame connectivity for multi-hop reasoning [1]. To avoid sink-induced bias, we use simple query-token similarity, whose effectiveness has been widely validated in other fields [21, 30, 31]. Moreover, we promote diversity to avoid repeatedly selecting the same regions, reducing fragmented glimpses.

3 SemVID

3.1 Problem Definition

We consider a Video-Language Model (VLM) that encodes a video into patch tokens and then performs question-answering with a language query. Given a video with T frames, each frame is tokenized into P visual tokens, producing token embeddings $V_{\text{patch}} \in \mathbb{R}^{T\times P\times D}$, where D is the dimension of hidden states. We also compute a frame-level global feature by applying mean pooling to the patch dimension, forming $V_{\text{glb}} \in \mathbb{R}^{T\times D}$. The query is represented by token embeddings $Q \in \mathbb{R}^{N\times D}$, where N is the query token length. For VLM processing, the queries are combined with different instructions depending on the task.

Video Temporal Grounding. Given a video and a query describing an event, VTG aims to localize the start and end timestamps of that event.

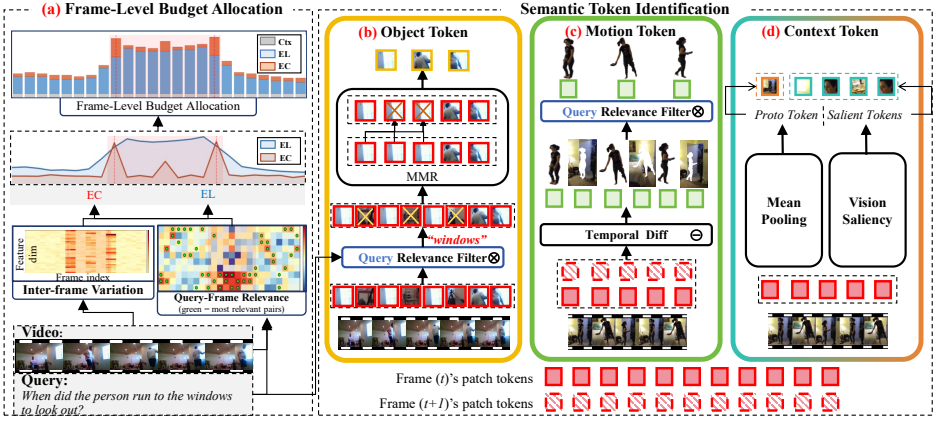

Fig. 2: Overview of SemVID semantic-oriented pruning. (a) **Frame-level budget allocation**: assigns per-frame token budgets by jointly considering query-frame relevance and inter-frame variation. Given per-frame budgets, SemVID then outputs three roles of tokens. (b) **Object token**: uses Maximal Marginal Relevance (MMR) to retain query-relevant yet diverse evidence. (c) **Motion token**: retain query-aligned transitions as relay nodes to bridge long-range evidence and preserve connectivity. (d) **Context token**: selects per-frame anchors by scene-level representativeness and saliency.

SemVID Overview. Given a retention ratio r, we select a subset of visual tokens $\mathcal{V}'$ such that $|\mathcal{V}'| = r{\cdot}(TP)$ without compromising performance. As shown in Fig. 2, SemVID first performs frame-level budget allocation to distribute the budget across T frames, producing per-frame token quotas $\{k^{(t)}\}_{t=1}^{T}$. Given these budgets, we then conduct role-aware semantic token selection within each frame. Since object tokens localize query-critical evidence, we first allocate $\alpha k^{(t)}$ slots to object tokens, where α balances evidence preservation and connectivity. We then use the remaining $(1-\alpha)k^{(t)} - k_{\text{ctx}}$ slots for motion tokens to bridge evidence and finally reserve k_{ctx} context tokens as auxiliary anchors.

3.2 Frame-Level Budget Allocation

SemVID balances two objectives derived from the attention graph: (1) *evidence localization*, which prioritizes frames where the query injects evidence; and (2) *evidence connectivity*, which emphasizes transition-rich frames that serve as temporal relays between evidence-bearing moments. This connectivity-aware allocation is particularly important for VTG, as precise boundary localization depends on linking evidence through intermediate state changes rather than fragment frames. More discussion on these objectives is recorded in Appx. A.1.

Evidence Localization. To quantify which frames are more likely to contain query-critical evidence, we use a lightweight query-relevance over frame features:

$$s_{\text{EL}}^{(t)} = \frac{1}{N} \cdot \hat{V}_{\text{glb}}^{(t)} \hat{Q}^{\top}, \tag{1}$$

where $\hat{V}_{\text{glb}}^{(t)} \in \mathbb{R}^D$ and $\hat{Q} \in \mathbb{R}^{N\times D}$ denote normalized features of each frame and query tokens. This formulation yields a normalized injection prior to frames, indicating the relevance between the query and each frame.

Evidence Connectivity. We also allocate budget to transition-rich frames critical for connecting long-range evidence flow. A cheap proxy is used to localize intermediate frames by measuring the temporal change of frames:

$$s_{\text{EC}}^{(t)} = \begin{cases} \|\hat{V}_{\text{glb}}^{(t)} - \hat{V}_{\text{glb}}^{(t-1)}\|_2, & t \geq 2, \\ s_{\text{EC}}^{(2)}, & t = 1. \end{cases} \tag{2}$$

Large $s_{\text{EC}}^{(t)}$ suggests a likely state transition, which serves as intermediate evidence needed for connectivity. We allocate more tokens to such frames to preserve transition cues and keep the multi-hop evidence path continuous.

Budget Allocation. Given a retention ratio r, we keep $|\mathcal{V}'| = r \cdot (TP)$ tokens in total. We compute a mixed per-frame weight by

$$w^{(t)} = \alpha\, s_{\text{EL}}^{(t)} + (1-\alpha)\, s_{\text{EC}}^{(t)}. \tag{3}$$

We reuse α to balance query relevance and transitions. To avoid token-empty frames, we use a per-frame floor k_{ctx}. The final per-frame budget is

$$k^{(t)} = (K - Tk_{\text{ctx}}) \cdot \frac{w^{(t)}}{\sum_{i=1}^{T} w^{(i)}} + k_{\text{ctx}}. \tag{4}$$

This allocation mainly follows the weight $w^{(t)}$, while k_{ctx} serves as an auxiliary safeguard for per-frame context coverage crucial for evidence preservation.

3.3 Object Token

To localize evidence, a natural strategy is to retain the tokens that the query attends to. However, attention-based query relevance methods that compute cross-attention over all patches require materializing large attention maps, which introduces substantial overhead and becomes prohibitive for long videos.

In our attention-graph formulation, the query-to-vision interaction is characterized by an evidence injection distribution $\boldsymbol{\pi}^{(0)}$, which measures the probability mass assigned by the query to each visual patch via cross-modal attention (see Appx. A.3). We use a lightweight query-to-patch relevance score $\mathbf{s}_{\text{evi}}$ as an efficient estimator of $\boldsymbol{\pi}^{(0)}$, following the spirit of Eq. 1 but on patch features:

$$\mathbf{s}_{\text{evi}} = \frac{1}{N} \cdot \hat{V}_{\text{patch}} \hat{Q}^{\top} \in \mathbb{R}^{T\times P}. \tag{5}$$

$\mathbf{s}_{\text{evi}}$ assigns higher scores to patches that are more likely to be attended by the query, serving as candidate evidence for grounding.

However, directly taking the top-k tokens is prone to redundancy. Since high-scoring patches often cluster around the same object across nearby spatial locations, leading to near-duplicate selections that waste budget and reduce semantic

coverage. We therefore adopt Maximal Marginal Relevance (MMR) to balance *query relevance* and *non-redundancy*: at each step, it selects the candidate patch $p_{\text{obj}}^{\star(t)}$ for frame t that maximizes

$$p_{\text{obj}}^{\star(t)} = \arg\max_{p \in \mathcal{C}} \Big[\lambda_{\text{mmr}} \cdot \mathbf{s}_{\text{evi}}^{(t,p)} - (1 - \lambda_{\text{mmr}}) \cdot \max_{p' \in \mathcal{S}} (\hat{V}_{\text{patch}}^{(t,p)} \hat{V}_{\text{patch}}^{(t,p')\top}) \Big], \tag{6}$$

where $\mathcal{C}$ is the candidate patch index set at t, $\mathcal{S}$ is the already selected patch index set, $\hat{V}_{\text{patch}}^{(t,p)}$ is the normed feature of the patch p at frame t, and $\lambda_{\text{mmr}} \in [0, 1]$ controls the trade-off. By penalizing candidates highly similar to previously selected tokens, MMR produces a compact yet diverse set of query-relevant object evidence. The detailed efficient algorithm is recorded in Appx. B.

3.4 Motion Token

Object evidence is necessary but not always sufficient for VTG. Precise temporal grounding also requires capturing the state transitions around the boundary and maintaining a traceable path that links evidence across frames. To support such temporal traceability, SemVID introduces motion tokens as explicit relay evidence to capture state changes and facilitate cross-frame routing.

In dot-product attention, a token routes less mass to tokens with low feature similarity. Consequently, long-range evidence propagation is most likely to bottleneck at frames where the visual state changes sharply. We therefore identify motion tokens from regions with strong temporal feature variation by scoring each patch with a cheap token-level difference $m_{\text{mot}}^{(t,p)}$:

$$m_{\text{mot}}^{(t,p)} = \begin{cases} \|\hat{V}_{\text{patch}}^{(t+1,p)} - \hat{V}_{\text{patch}}^{(t,p)}\|_2, & t = 1, \\ \|\hat{V}_{\text{patch}}^{(t,p)} - \hat{V}_{\text{patch}}^{(t-1,p)}\|_2, & t = T, \\ \frac{1}{2}\big(\|\hat{V}_{\text{patch}}^{(t,p)} - \hat{V}_{\text{patch}}^{(t-1,p)}\|_2 + \|\hat{V}_{\text{patch}}^{(t+1,p)} - \hat{V}_{\text{patch}}^{(t,p)}\|_2\big), & \text{otherwise.} \end{cases} \tag{7}$$

Large $m_{\text{mot}}^{(t,p)}$ indicates a strong local transition, where temporal connectivity is typically most fragile. Identifying such regions helps bridge evidence across frames. A formal definition of temporal connectivity is provided in Appx. A.3.

Since motion alone may be dominated by camera jitter or background changes, we make motion selection query-aware by fusing motion with query relevance:

$$s_{\text{mot}}^{(t,p)} = (1 - \beta)\, m_{\text{mot}}^{(t,p)} + \beta \max_{\hat{q} \in \hat{Q}} \big(\hat{V}_{\text{patch}} \hat{q}^{\top}\big), \tag{8}$$

We keep the top-$k_{\text{mot}}^{(t)}$ tokens according to $s_{\text{mot}}^{(t,p)}$, which prioritizes query-relevant transitions while suppressing irrelevant background motion.

3.5 Context Token

Without context anchors, pruning can create token-empty gaps or overly query-centric evidence, both of which distort temporal reasoning. To keep each frame interpretable, we retain a small set of query-agnostic context tokens.

First, we select a *proto* token $p^{\star}_{\text{proto}}$ that best represents the frame background by matching the frame-global mean feature:

$$p^{\star(t)}_{\text{proto}} = \arg\max_{p\in\{1,\dots,P\}} \hat{V}^{(t,p)}_{\text{patch}} \hat{V}^{(t)\top}_{\text{glb}}. \tag{9}$$

Then, to complete the context budget $p^{\star(t)}_{\text{ctx}}$, we select the remaining $k_{\text{ctx}}-1$ tokens by a lightweight saliency score $s^{(t,p)}_{\text{sal}} = \left\| V^{(t,p)}_{\text{patch}} \right\|_2$, yielding:

$$\mathcal{P}^{(t)}_{\text{ctx}} = \left\{ p^{\star(t)}_{\text{proto}} \right\} \cup \text{TopK}\left(\left\{ s^{(t,p)}_{\text{sal}} \right\}_{p=1}^{P}, k_{\text{ctx}} - 1 \right), \tag{10}$$

where $\text{TopK}(\cdot, k)$ returns the indices of the k largest scores. Although we keep only a few context tokens per frame, they are vital for perceiving coherent context and scene changes, avoiding fragmented reasoning.

3.6 Evaluation Metrics

Pruning removes nodes and edges in the attention graph, which can degrade both evidence propagation and cross-frame connectivity that are essential for VTG. We therefore characterize the quality of the pruned graph with two measurable quantities: Evidence Retention (ER) and Connectivity Strength (CS).

Evidence Retention. We measure ER by comparing the pre-pruning evidence landing map $\boldsymbol{\pi}^{(1)}_{\text{full}}$ with the post-pruning one $\boldsymbol{\pi}^{(1)}$, where $\boldsymbol{\pi}^{(1)}$ is the distribution of patches that the query ultimately routes to in the attention graph. Direct cross-entropy is ill-defined because pruned tokens receive zero probability mass, which can lead to infinite values. We therefore first compute the retained evidence mass $\rho = \sum_{v\in\mathcal{V}'} \boldsymbol{\pi}^{(1)}_{\text{full}}(v)$, and define ER as the negative cross-entropy:

$$\text{ER}(\mathcal{V}') = \exp\Big(\sum_{v\in\mathcal{V}} \boldsymbol{\pi}^{(1)}_{\text{full}}(v) \log \bar{\boldsymbol{\pi}}^{(1)}(v) \Big), \quad \bar{\boldsymbol{\pi}}^{(1)}(v) = \begin{cases} \rho \cdot \boldsymbol{\pi}^{(1)}(v), & v \in \mathcal{V}', \\ \dfrac{1-\rho}{|\mathcal{V} \setminus \mathcal{V}'|}, & v \notin \mathcal{V}', \end{cases} \tag{11}$$

where $\mathcal{V}' \subseteq \{V^{(1,1)}_{\text{patch}}, \dots, V^{(T,P)}_{\text{patch}}\}$ is the retained token set. We exponentiate the negative cross-entropy so that ER is bounded in $(0, 1]$, where higher ER indicates that pruning preserves a larger portion of query-induced evidence distribution.

Connectivity Strength. To quantify whether pruning preserves a traceable evidence chain, we measure how much attention mass can be routed across adjacent frames. At layer ℓ, we define the cross-frame routing mass from frame t to $t{+}1$ as $\Gamma^{(\ell)}_{\mathcal{V}'}(t) = \sum_{u_i\in\mathcal{V}'_t} \sum_{u_j\in\mathcal{V}'_{t+1}} \mathbb{P}^{(\ell)}(u_i \to u_j)$, where $\mathcal{V}'_t$ denotes the retained tokens in frame t and $\mathbb{P}^{(\ell)}(u_i \to u_j)$ is the layer-ℓ routing probability (i.e., the attention-based transition mass) from token u_i to token u_j in the attention graph. The detailed implementation of $\mathbb{P}^{(\ell)}(u_i \to u_j)$ is provided in Appx. A.2.

We then aggregate $\Gamma^{(\ell)}_{\mathcal{V}'}(t)$ over time and layers to obtain CS:

$$\mathrm{CS}(\mathcal{V}') = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=1}^{T-1}\Gamma^{(\ell)}_{\mathcal{V}'}(t), \tag{12}$$

where L is the layer amount. Intuitively, larger $\Gamma^{(\ell)}_{\mathcal{V}'}(t)$ means stronger temporal links, enabling multi-hop propagation to connect evidence over longer horizons.

4 Experiments

4.1 Experimental Setting

Benchmarks. We evaluate SemVID on Video Temporal Grounding (VTG) in two standard long-video grounding benchmarks, Charades-STA [36] and ActivityNet-Grounding [5]. In the appendix, we also evaluate VideoQA on Video-MME [13] and LongVideoBench [43]. The details of benchmarks are recorded in Appx. C.
Metrics. Following [32,50], we report mean Intersection over Union (mIoU) and $R1$ at tIoU thresholds $m \in \{0.3, 0.5, 0.7\}$. TFLOPs are estimated for efficiency.
Implementation Details. We set up a unified evaluation protocol for fair comparisons detailed in Appx. C. SemVID is evaluated on Qwen3-VL-4B/8B-Thinking [47], Qwen2.5-VL-7B-Instruct [3], and LLaVA-OneVision-7B [22]. To our knowledge, this is the first study of pruning on Qwen3-VL. We also adapt FastVID [34] and VisionZip [48] to Qwen3-VL as baselines. Unless noted, we set $\alpha = 0.6$ (Eq. 3), $\lambda_{\mathrm{mmr}} = 0.8$ (Eq. 6), $\beta = 0.5$ (Eq. 8), and $k_{\mathrm{ctx}} = 3$ (Eq. 10).

4.2 Comparisons with State-of-the-Art Methods

Qwen3-VL. Tab. 1 shows that our Qwen3-VL-based SemVID consistently achieves the best accuracy-efficiency trade-off across model sizes and budgets. Under an extremely aggressive 12.5% budget, SemVID retains up to 95.4% of the original mIoU while largely preserving performance at high IoU ($R1$@0.7), indicating precise boundary localization rather than coarse retrieval.

A key observation is that the gap between methods is well explained by our two attention-graph diagnostics: Evidence Retention (ER) and Connectivity Strength (CS). VisionZip, a redundancy-driven approach, tends to over-merge temporally adjacent visual states, which both suppresses boundary-critical evidence for VTG and removes intermediate relay tokens, leading to pronounced degradations in ER and CS. In contrast, FastVID is saliency-driven and better preserves graph connectivity, but it concentrates the limited budget on a small set of anchor frames, leaving boundary-adjacent evidence sparsely represented and reducing effective ER. SemVID, instead, explicitly allocates tokens to both query-relevant evidence and high-variation transition frames, thereby constructing a comprehensive and temporally continuous evidence chain that jointly sustains ER and CS under aggressive pruning.

At 25% retention, SemVID approaches near-lossless compression on both benchmarks, retaining up to 96.9% of the original mIoU. This strong retention further validates that preserving both ER and CS is crucial for VTG.

Qwen2.5-VL. We further evaluate SemVID on Qwen2.5-VL-7B under a 12.5% visual-token budget in Tab. 2. SemVID again achieves the best overall performance on both Charades-STA and ActivityNet, and the results suggest that VTG accuracy strongly correlates with jointly high ER and CS. Existing baselines largely overlook temporal connectivity, whereas SemVID explicitly optimizes it and improves CS by 23.9% over VisionZip (20.1 to 24.9). Importantly, several query-driven baselines are not deployable on long videos because they require materializing full cross-attention weights, leading to out-of-memory errors. SemVID avoids this overhead by relying on lightweight similarity and temporal-difference proxies, making it practical for long-video inference.

We additionally vary the retention ratio and plot the mIoU trends in Fig. 3. SemVID shows a markedly slower degradation under aggressive pruning, indi-

	ActivityNet-Grounding							Charades-STA						
Method	R1@0.3	R1@0.5	R1@0.7	mIoU	(%)	ER	CS	R1@0.3	R1@0.5	R1@0.7	mIoU	(%)	ER	CS
Qwen3-VL-4B, Retain 12.5% Tokens														
Qwen3-VL-4B	54.60	37.54	23.74	40.33	100	$1e^4$	64.4	79.62	65.99	40.81	56.07	100	$1e^4$	81.5
VisionZip [48]	26.02	15.85	9.58	19.89	49.3	1.6	22.4	56.91	37.20	16.72	36.83	65.7	3.0	26.4
FastVID [34]	45.70	29.32	17.30	33.16	82.2	3.0	**39.7**	55.33	33.32	15.47	35.98	64.2	2.9	29.8
SemVID (ours)	**51.96**	**34.73**	**22.18**	**38.49**	**95.4**	**4.5**	31.6	**74.17**	**56.91**	**30.67**	**49.89**	**89.0**	**5.6**	**33.5**
Qwen3-VL-4B, Retain 25% Tokens														
Qwen3-VL-4B	54.60	37.54	23.74	40.33	100	$1e^4$	64.4	79.62	65.99	40.81	56.07	100	$1e^4$	81.5
VisionZip [48]	30.74	19.62	12.15	23.30	57.8	2.8	30.8	64.11	47.28	23.79	43.09	76.9	3.6	47.3
FastVID [34]	46.54	30.24	18.42	34.33	85.1	3.8	**57.3**	57.93	39.63	19.52	38.13	68.0	3.1	53.7
SemVID (ours)	**52.84**	**35.68**	**22.51**	**39.06**	**96.9**	**6.5**	55.2	**76.91**	**61.08**	**33.17**	**52.31**	**93.3**	**9.8**	**56.1**
Qwen3-VL-8B, Retain 12.5% Tokens														
Qwen3-VL-8B	52.86	35.81	22.86	39.10	100	$1e^4$	94.7	80.03	66.59	42.20	56.67	100	$1e^4$	121.5
VisionZip [48]	10.43	6.18	3.91	8.12	20.8	1.0	14.3	21.43	15.73	7.69	14.52	25.6	1.1	17.7
FastVID [34]	40.53	26.16	15.91	30.61	78.3	3.5	28.4	52.84	35.58	16.36	35.26	62.2	2.8	26.7
SemVID (ours)	**48.32**	**31.93**	**20.47**	**36.21**	**92.6**	**6.8**	**29.0**	**73.45**	**56.24**	**31.11**	**49.93**	**88.1**	**5.7**	**30.5**
Qwen3-VL-8B, Retain 25% Tokens														
Qwen3-VL-8B	52.86	35.81	22.86	39.10	100	$1e^4$	94.7	80.03	66.59	42.20	56.67	100	$1e^4$	121.5
VisionZip [48]	14.24	8.95	5.63	10.92	27.9	1.8	20.1	36.13	28.31	15.86	24.96	44.4	2.9	21.8
FastVID [34]	41.73	27.23	16.93	31.63	80.9	3.9	45.8	55.38	38.58	19.49	37.12	65.5	3.5	24.2
SemVID (ours)	**50.30**	**33.49**	**21.50**	**37.54**	**96.0**	**8.8**	**51.2**	**76.30**	**60.11**	**34.55**	**52.66**	**93.0**	**9.5**	**51.6**

Table 1: VTG results with Qwen3-VL under 12.5% and 25% visual token retention. Percentages are relative to the original. Evidence Retention (ER) and Evidence Retention (CS) are attention graph metrics introduced in Sec. 3.6. The unit for ER is $1e^{-4}$.

	ActivityNet-Grounding							Charades-STA						
Method	R1@0.3	R1@0.5	R1@0.7	mIoU	(%)	ER	CS	R1@0.3	R1@0.5	R1@0.7	mIoU	(%)	ER	CS
Qwen2.5-VL-7B, Retain 12.5% Tokens														
Qwen2.5-VL-7B	29.22	15.77	7.49	22.25	100	$1e^4$	71.8	77.39	59.33	33.82	56.85	100	$1e^4$	90.9
FastV [9]	Out-of-Memory (OOM)					-	-	Out-of-Memory (OOM)					-	-
VScan [51]	Out-of-Memory (OOM)					-	-	Out-of-Memory (OOM)					-	-
DART [42]	16.04	8.24	4.06	12.38	55.6	2.7	14.3	35.65	25.19	12.85	24.30	42.7	2.8	15.7
ToME [4]	20.22	10.24	4.44	15.86	71.3	3.8	18.3	43.46	29.08	14.62	29.93	52.6	3.3	20.7
TokenSculpt [28]	20.67	10.54	4.73	16.20	72.8	4.1	18.8	44.19	29.19	14.81	30.08	52.9	3.6	21.4
FastVID [34]	20.89	10.53	4.66	16.28	73.1	3.9	22.7	44.08	29.96	15.68	30.57	53.8	3.7	21.9
VisionZip [48]	20.87	10.51	4.65	16.33	73.3	4.3	18.9	50.13	33.27	15.88	33.51	58.9	4.1	20.1
SemVID (ours)	**21.74**	**10.69**	**4.63**	**17.21**	**77.4**	**4.5**	**23.1**	**54.01**	**35.67**	**18.23**	**36.54**	**64.3**	**4.2**	**24.9**

Table 2: VTG results with Qwen2.5-VL under 12.5% retention. FastV and VScan materialize full self-attention matrix and lead to OOM in long videos.

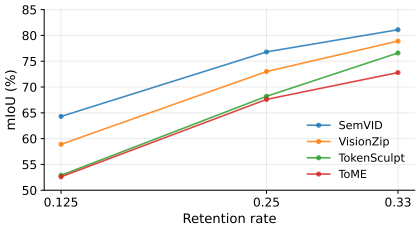

Fig. 3: mIoUs on Charades-STA under different token retention ratios.

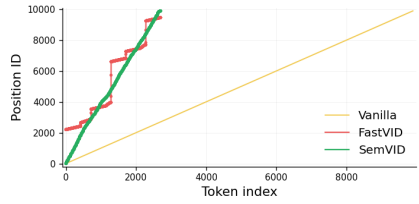

Fig. 4: Position-ID trajectories before and after pruning. Pruning alters the trajectory relative to the original.

	Method	SBA	STI	Charades-STA R1@0.3	R1@0.5	R1@0.7	mIoU	(%)	ER	CS
	Qwen3-VL-4B	-	-	79.62	65.99	40.81	56.07	100	$1e^4$	81.5
	Sampling with Fixed-Interval			61.18	43.60	20.16	39.80	71.0	2.8	32.0
	Random Sampling			63.98	44.67	21.10	41.76	74.4	3.0	27.7
(a)	Sampling with Query-Relevance			67.31	49.56	25.88	44.97	80.2	5.5	20.6
	Semantic Budget, Random Selection	✓		66.25	47.68	23.79	43.26	77.2	3.6	28.1
	Uniform Budget, Semantic Selection		✓	73.01	54.91	29.60	48.50	86.5	5.1	33.2
	Full, Semantic Budget and Selection	✓	✓	**74.17**	**56.91**	**30.67**	**49.89**	**89.0**	**5.6**	**33.5**
	Qwen3-VL-4B	-	-	79.62	65.99	40.81	56.07	100	$1e^4$	81.5
(b)	FastVID [34]			55.33	33.32	15.47	35.98	64.2	2.9	**29.8**
	FastVID + Our Semantic Budget	✓		**73.69**	**55.57**	**29.79**	**48.88**	**87.2**	**5.3**	26.4

Table 3: Ablation results on Semantic Budget Allocation (SBA) and Semantic Token Identification (STI). (a) Comparison with different sampling strategies. (b) Implementing our budget allocation module into FastVID.

cating that semantic evidence allocation is particularly effective when the token budget is the bottleneck under the realistic regime for long videos.

Furthermore, we observe that Qwen2.5-VL suffers a larger pruning-induced performance drop than Qwen3-VL. A plausible reason is the positional encoding. Qwen2.5-VL relies on RoPE [37], and irregular token retention can distort the position-ID trajectory (Fig. 4), introducing and amplifying bias in timestamp perception [12,49]. In contrast, Qwen3-VL uses explicit timestamp tokens before each frame, making temporal cues more robust to pruning.

4.3 Ablation Study

Budget Allocation (BA). Comparing *Random Sampling* with *Semantic Budget, Random Selection* in Table 3(a), BA improves the retained mIoU by 2.8% and boosts ER from 3.0 to 3.6. Since within-frame token selection remains random, this gain isolates the contribution of frame-level budget allocation. Specifically, BA introduces a lightweight prior to estimate per-frame information density, thereby increasing the likelihood of retaining query-relevant evidence and improving ER. Moreover, it prevents the budget from collapsing onto a few salient frames by allocating tokens to boundary-adjacent and transition moments, which helps preserve temporal connectivity and thus maintains CS.

	Method	Charades-STA mIoU	(%)	ER	CS
	Qwen3-VL-4B	56.07	100	$1e^4$	81.5
(a)	Attention Selection	46.57	83.1	5.0	32.8
	Relevance Selection	**49.89**	**89.0**	**5.6**	**33.5**
(b)	$\lambda_{\text{mmr}} = 1$ (w/o MMR)	49.44	88.2	5.4	33.3
	$\lambda_{\text{mmr}} = \mathbf{0.8}$	**49.89**	**89.0**	**5.6**	33.5
	$\lambda_{\text{mmr}} = 0.6$	49.25	87.8	5.3	**33.9**
(c)	$\alpha = 0$ (w/o Obj. Token)	49.48	88.2	5.4	**33.9**
	$\alpha = 0.2$	49.56	88.4	5.5	33.7
	$\alpha = 0.4$	49.79	88.8	5.6	33.6
	$\alpha = \mathbf{0.6}$	**49.89**	**89.0**	5.6	33.5
	$\alpha = 0.8$	49.56	88.4	5.6	30.7
	$\alpha = 1.0$ (w/o Mot. Token)	48.11	85.8	**5.7**	23.3

Table 4: Ablation on object tokens. (a) Different selection strategies. (b) Effectiveness of MMR (Eq. 6). (c) Different object-motion token ratios (Eq. 3).

	Method	Charades-STA mIoU	(%)	ER	CS
	Qwen3-VL-4B	56.07	100	$1e^4$	81.5
(a)	$\beta = 0$ (w/o query-aware)	49.04	87.5	5.5	32.1
	$\beta = \mathbf{0.5}$	**49.89**	**89.0**	**5.6**	**33.5**
(b)	$k_{\text{ctx}} = 0$ (w/o Ctx. Token)	49.20	87.7	**5.6**	31.9
	$k_{\text{ctx}} = 1$ (proto only)	49.35	88.0	**5.6**	32.4
	$\mathbf{k_{ctx}} = \mathbf{3}$	**49.89**	**89.0**	**5.6**	33.5
	$k_{\text{ctx}} = 10$	48.52	86.5	4.8	**34.6**

Table 5: Ablation on motion and context token selection. (a) Different query-aware weights in background change suppression (Eq. 8). (b) Different context token selection strategies (k_{ctx} in Eq. 10 and the proto token in Eq. 9).

To test generality and modularity, we plug our BA into FastVID in Table 3(b), which substantially improves retained mIoU from 64.2% to 87.2%. As shown in Fig. 6, FastVID relies on sparse anchor frames, which over-compresses boundary-adjacent moments, blurs transition cues, and fragments the temporal evidence chain. Our BA reallocates tokens toward query-relevant and high-variation transition frames and thus improving boundary evidence coverage.

Semantic Token Identification (STI). Table 3(a) shows that under a uniform per-frame budget, replacing fixed-interval sampling with STI improves the retained mIoU by 8.7%. This indicates that once temporal coverage is fixed, token selection becomes the dominant factor: STI prioritizes query-aligned evidence while avoiding redundant selections, thereby significantly increasing ER without collapsing CS. Combining BA and STI yields the best results, outperforming either component alone and retaining 88.1% of the original performance. It demonstrates that pruning for VTG requires both BA to maintain frame coverage and evidence-aware STI to preserve diverse evidence and traceable chains.

Object Tokens. Table 4(a) compares two evidence selection strategies. *Attention selection* uses the raw query and key projection matrix from the model's first attention block to compute a lightweight cross-attention between the language and visual patches. However, attention sinks can bias the weights toward non-semantic patches, hindering accurate identification of right evidence. In contrast, using direct query relevance is both more efficient and more precise for evidence localization, resulting in a 5.9% gain in retained performance. Further, Table 4(b) shows that Maximal Marginal Relevance (MMR) expands the evidence set to cover complementary object parts and nearby contextual cues, yielding a denser and more stable evidence extraction, as reflected by higher ER.

Motion Tokens. Ablations on the object-motion ratio in Table 4(c) show that removing motion tokens causes a clear drop in both mIoU and CS. This

	Tokens		TFLOPs		Prefill Time				ActivityNet	
Method	#	(%)	Value	(%)	Pruning	LLM Forward	Total	Speed	mIoU	(%)
Qwen3-VL-4B	10460	100	59.4	100	-	1263.4	1263.4	1×	40.33	100
VisionZip [48]	1307	12.5	4.8	8.7	710.9	184.3	895.2	1.4×	19.89	49.3
FastVID [34]	1370	13.1	5.4	9.1	**23.8**	185.9	**209.7**	**6.0**×	33.16	82.2
SemVID	1307	12.5	4.8	8.7	33.1	184.6	217.7	5.8×	**38.49**	**95.4**

Table 6: Efficiency Comparison on Qwen3-VL. Prefill time refers to the latency to the first generated token, which comprises pruning latency introduced by token selection.

supports our claim that VTG is limited not only by the presence of evidence but also by their connectivity. Our motion tokens serve as relay nodes that strengthen CS, enabling the model to connect evidence via multi-hop attention. Although allocating motion tokens consumes a small fraction of the budget, it is crucial for maintaining a coherent evidence chain, while the majority of tokens should remain dedicated to object tokens to maximize evidence coverage.

Additionally, Table 5(a) shows that query-aware filtering improves CS by suppressing background transitions and focusing on query-relevant state changes.

Context Tokens. Varying k_{ctx} in Table 5(b) shows that a small number of context tokens is necessary. Completely removing context anchors hurts evidence connectivity, while over-allocating context dilutes boundary-critical evidence. This highlights the intended role of context tokens: that they are not competing evidence but lightweight anchors that capture global scene context and stabilize temporal connectivity under aggressive pruning.

4.4 Inference Latency

Table 6 reports an efficiency comparison. SemVID achieves the best accuracy-efficiency balance: under the same 12.5% token budget, it substantially outperforms FastVID and VisionZip in accuracy. In terms of latency, SemVID reduces the prefill time from 1263.4 ms to 217.7 ms, yielding a 5.8× speedup over the original. Although FastVID shows a slightly lower pruning overhead, this difference is negligible compared to the dominant LLM forward cost. Overall, SemVID provides near-FastVID efficiency while preserving markedly stronger performance, making it a practical choice when both throughput and quality matter.

4.5 Visualization

Fig. 5 qualitatively illustrates SemVID's role-aware token selection. SemVID retains noticeably more tokens near event boundaries, where state transitions occur and grounding evidence is most informative. This aligns with the goal of preserving a coherent evidence chain rather than uniformly compressing the video.

We further observe distinct roles for different semantic tokens. *Object tokens* capture query-mentioned entities (e.g., switch, cabinet), treating them as evidence. MMR (Eq. 6) avoids redundant duplicates of a single object. *Motion*

Fig. 5: Visualization results. blue boxes denote object tokens, Red boxes indicate motion tokens, and yellow boxes represent context tokens.

tokens concentrate on action regions and often highlight the decisive transition cues that bridge pre-boundary and post-boundary evidence. *Context tokens* cover stable background anchors that help maintain temporal coherence under aggressive pruning. These qualitative behaviors are consistent with the quantitative ER and CS improvements and the strong VTG performance.

Additional visualizations of the effects of our budget allocation and evidence retention are provided in Appx. D.

4.6 Discussion

While SemVID preserves the evidence chain with role-aware token selection, a limitation of our implementation lies in the coarse Budget Allocation (BA). BA jointly considers the query-frame relevance and inter-frame variation, where the latter is approximated by frame feature differences. This proxy can be biased by camera or irrelevant motions and thus absorb part of the budget (Fig. 6), potentially weakening coverage of subtle actions. In practice, the effect is limited because our allocation enforces per-frame coverage, and subtle actions typically require a small number of evidence tokens. Moreover, motion tokens are eventually filtered by query relevance, which suppresses background transitions and retains meaningful motion evidence. We report a dedicated analysis of sensitivity to motion amplitude and include additional evaluation on VideoQA in Appx. D.

5 Conclusion

We revisit training-free visual token pruning for Video Temporal Grounding (VTG) under an evidence chain formulation and formalize two VTG-tailored metrics, evidence retention and connectivity strength. We propose SemVID, a plug-and-play training-free pruning framework that explicitly optimizes these two objectives. SemVID performs semantic budget allocation and selects a role-aware token set: object tokens preserve diverse query-aligned evidence, motion tokens act as query-filtered transition relays, and lightweight context tokens stabilize scene continuity. Extensive experiments demonstrate that SemVID consistently delivers a strong accuracy-efficiency trade-off, retaining substantially higher performance over strong baselines while significantly accelerating prefill. By explicitly preserving the right evidence and keeping it connected, SemVID provides a simple yet effective recipe for making long-video VTG practical.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 4190–4197 (2020) 3, 4, 19
2. Ainslie, J., Lee-Thorp, J., De Jong, M., Zemlyanskiy, Y., Lebrón, F., Sanghai, S.: Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245 (2023) 21
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 9, 21
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022) 2, 4, 10, 22, 25, 26
5. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015) 9, 20, 21
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 19
7. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 782–791 (2021) 3, 19
8. Chen, H., Wang, X., Chen, H., Feng, W., Song, Z., Jia, J., Zhu, W.: Localizing step-by-step: Multimodal long video temporal grounding with llm. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025) 1, 2, 4
9. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35. Springer (2024) 3, 4, 10, 21, 22, 25
10. Chen, Y.W., Tsai, Y.H., Yang, M.H.: End-to-end multi-modal video temporal grounding. Advances in Neural Information Processing Systems **34**, 28442–28453 (2021) 1
11. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 2, 3
12. Endo, M., Wang, X., Yeung-Levy, S.: Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22826–22835 (2025) 4, 11
13. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR (2025) 2, 9, 21
14. Fu, T., Liu, T., Han, Q., Dai, G., Yan, S., Yang, H., Ning, X., Wang, Y.: Framefusion: Combining similarity and importance for video token reduction on large vision language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22654–22663 (2025) 25
15. Huang, K., Zou, H., Xi, Y., Wang, B., Xie, Z., Yu, L.: Ivtp: Instruction-guided visual token pruning for large vision-language models. In: European Conference on Computer Vision. pp. 214–230. Springer (2024) 4

16. Huang, X., Zhou, H., Han, K.: Prunevid: Visual token pruning for efficient video large language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 19959–19973 (2025) 4, 24, 25
17. Kim, M., Gao, S., Hsu, Y.C., Shen, Y., Jin, H.: Token fusion: Bridging the gap between token pruning and token merging. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1383–1392 (2024) 2
18. Kim, S., Shen, S., Thorsley, D., Gholami, A., Kwon, W., Hassoun, J., Keutzer, K.: Learned token pruning for transformers. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. pp. 784–794 (2022) 2
19. Kumar, Y.: Language-guided temporal token pruning for efficient videollm processing. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 8935–8942 (2025) 3, 4
20. Lei, J., Berg, T., Bansal, M.: Revealing single frame bias for video-and-language learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 487–507 (2023) 2
21. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020) 4
22. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 9
23. Li, J., Wang, G., Zheng, S., Ni, M., Lu, X., Ye, G., Guan, Y.: Towards mitigating modality bias in vision-language models for temporal action localization. arXiv preprint arXiv:2601.21078 (2026) 4
24. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2794–2804 (2023) 1
25. Liu, S., Zhao, C., Zohra, F., Soldan, M., Pardo, A., Xu, M., Alssum, L., Ramazanova, M., Alcázar, J.L., Cioppa, A., Giancola, S., Hinojosa, C., Ghanem, B.: Opentad: A unified framework and comprehensive study of temporal action detection. arXiv preprint arXiv:2502.20361 (2025) 21
26. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10810–10819 (2020) 1
27. OpenAI: Gpt-4o system card. Tech. rep., OpenAI (2024), accessed: 2026-03-03 23
28. QA, L.V.: Tokensculpt: Pruning with min-max spatio-temporal duplication for video grounding (2025) 4, 10, 22
29. Qu, M., Chen, X., Liu, W., Li, A., Zhao, Y.: Chatvtg: Video temporal grounding via chat with video dialogue large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1847–1856 (2024) 1
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 4
31. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 3982–3992 (2019) 4

32. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323 (2024) 9, 21
33. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020) 21
34. Shen, L., Gong, G., He, T., Zhang, Y., Liu, P., Zhao, S., Ding, G.: Fastvid: Dynamic density pruning for fast video large language models. arXiv preprint arXiv:2503.11187 (2025) 2, 4, 9, 10, 11, 13, 21, 22, 23, 24, 25, 26
35. Shinde, G., Ravi, A., Dey, E., Sakib, S., Rampure, M., Roy, N.: A survey on efficient vision-language models. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **15**(3), e70036 (2025) 2
36. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European conference on computer vision. pp. 510–526. Springer (2016) 9, 20, 21
37. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024) 11
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 2, 3
39. Wang, L., Mittal, G., Sajeev, S., Yu, Y., Hall, M., Boddeti, V.N., Chen, M.: Protege: Untrimmed pretraining for video temporal grounding by video temporal grounding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6575–6585 (2023) 1
40. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025) 2, 21
41. Wen, Z., Gao, Y., Li, W., He, C., Zhang, L.: Token pruning in multimodal large language models: Are we solving the right problem? arXiv preprint arXiv:2502.11501 (2025) 4
42. Wen, Z., Gao, Y., Wang, S., Zhang, J., Zhang, Q., Li, W., He, C., Zhang, L.: Stop looking for important tokens in multimodal language models: Duplication matters more. arXiv preprint arXiv:2502.11494 (2025) 4, 10, 21, 22
43. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024) 2, 9, 21
44. Wu, J., Liu, W., Liu, Y., Liu, M., Nie, L., Lin, Z., Chen, C.W.: A survey on video temporal grounding with multimodal large language model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 2, 4
45. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453 (2023) 4
46. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al.: Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. arXiv preprint arXiv:2410.17247 (2024) 21
47. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 9, 21
48. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. In: Proceedings of the Com-

puter Vision and Pattern Recognition Conference. pp. 19792–19802 (2025) 3, 4, 9, 10, 13, 22, 25
49. Ye, X., Gan, Y., Ge, Y., Zhang, X.P., Tang, Y.: Atp-llava: Adaptive token pruning for large vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24972–24982 (2025) 11
50. Zeng, X., Li, K., Wang, C., Li, X., Jiang, T., Yan, Z., Li, S., Shi, Y., Yue, Z., Wang, Y., et al.: Timesuite: Improving mllms for long video understanding via grounded tuning. arXiv preprint arXiv:2410.19702 (2024) 9, 21
51. Zhang, C., Ma, K., Fang, T., Yu, W., Zhang, H., Zhang, Z., Xie, Y., Sycara, K., Mi, H., Yu, D.: Vscan: Rethinking visual token reduction for efficient large vision-language models. arXiv preprint arXiv:2505.22654 (2025) 10, 22, 25, 26
52. Zhang, Q., Cheng, A., Lu, M., Zhang, R., Zhuo, Z., Cao, J., Guo, S., She, Q., Zhang, S.: Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20857–20867 (2025) 4
53. Zheng, L., Li, C., Jia, H., Zhang, X.: Reasoning paths as signals: Augmenting multi-hop fact verification through structural reasoning progression. arXiv preprint arXiv:2506.07075 (2025) 3
54. Zheng, M., Cai, X., Chen, Q., Peng, Y., Liu, Y.: Training-free video temporal grounding using large-scale pre-trained models. In: European Conference on Computer Vision. pp. 20–37. Springer (2024) 1

Supplementary Material

A Preliminary

A.1 Attention Graph.

To simplify the token-to-token interactions in a VLM, we decompose them into language-to-vision cross-attention and vision self-attention. For each Transformer layer ℓ, vision self-attention induces a directed weighted graph $G^{(\ell)} = \big(\mathcal{V}, \mathcal{E}^{(\ell)}, w^{(\ell)}\big)$, where $\mathcal{V}$ is the set of visual tokens, $\mathcal{E}^{(\ell)}$ contains directed token-to-token edges, and $w^{(\ell)}$ assigns a nonnegative weight to each edge. Then we compute the cross-attention from language tokens $\mathcal{Q}$ to visual tokens $\mathcal{V}$. Following prior research [1, 6], we use the weights in the L-th layer, where L is the number of layers within a Transformer model, to induce an **injection distribution** $\boldsymbol{\pi}^{(L)} \in \mathbb{R}^{|\mathcal{V}|}$ over $\mathcal{V}$. Intuitively, $\boldsymbol{\pi}^{(L)}$ tells where the query reads from the visual tokens, which is the starting point of evidence.

A.2 Evidence Flow

Given the attention graph at layer ℓ, $G^{(\ell)} = \big(\mathcal{V}, \mathcal{E}^{(\ell)}, w^{(\ell)}\big)$, we define the Markov transition matrix $\mathbf{P}^{(\ell)} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ whose entries are

$$\mathbf{P}^{(\ell)}_{i,j} = \mathbb{P}^{(\ell)}(u_i \to u_j) = \frac{w^{(\ell, u_i \to u_j)}}{\sum_{j' :\, (u_i \to u_{j'}) \in \mathcal{E}^{(\ell)}} w^{(\ell, u_i \to u_{j'})}}, \tag{13}$$

where $u_i, u_j \in \mathcal{V}$ are patch tokens and $(u_i \to u_j) \in \mathcal{E}^{(\ell)}$ denotes a directed self-attention edge from u_i to u_j. By construction, each row of $\mathbf{P}^{(\ell)}$ sums to one, hence $\mathbf{P}^{(\ell)}$ is row-stochastic and specifies how a token probabilistically routes information to its attended neighbors at layer ℓ.

Let $\boldsymbol{\pi}^{(\ell)}$ denote the evidence distribution over visual tokens at layer ℓ. Starting from the injection distribution $\boldsymbol{\pi}^{(L)}$ induced by cross-attention, evidence is propagated backward through vision self-attention layers by

$$\boldsymbol{\pi}^{(\ell)} = \mathbf{P}^{(\ell+1)\top} \boldsymbol{\pi}^{(\ell+1)} \in \mathbb{R}^{|\mathcal{V}|}, \qquad \ell = L-1, \ldots, 1. \tag{14}$$

The resulting $\boldsymbol{\pi}^{(1)}$ is the **landing distribution** of evidence flow over the input patch tokens. A larger $\pi^{(1)}(v)$ means that token v receives more query-induced evidence after multi-hop attention routing.

A.3 Cross-Frame Connectivity

VTG requires comparing states before and after a boundary, which relies on attention paths that connect evidence across time [1, 7]. We quantify the cross-frame routing strength at layer ℓ by the total transition mass from frame t to $t{+}1$:

$$\Gamma^{(\ell)}(t) = \sum_{u_i \in \mathcal{V}_t} \sum_{u_j \in \mathcal{V}_{t+1}} \mathbb{P}^{(\ell)}(u_i \to u_j), \tag{15}$$

where $\mathcal{V}_t$ is the set of tokens in frame t. Low $\Gamma^{(\ell)}(t)$ indicates weak temporal routing, meaning evidence has difficulty propagating across this boundary.

B Maximal Marginal Relevance (MMR) Algorithm

In this section, we describe the algorithmic procedure of our Maximal Marginal Relevance (MMR) selection, which balances query relevance with token diversity for object evidence. The core idea is to suppress near-duplicate selections and only allow visually similar candidates when they introduce sufficiently different semantic information.

Algorithm 1: MMR-based object token selection

Input: Normalized patch tokens $\{\hat{\mathbf{V}}_{patch}^{(t)} \in \mathbb{R}^{P\times D}\}_{t=1}^{T}$; object budgets $\{k_{\text{obj}}^{(t)}\}_{t=1}^{T}$; trade-off $\lambda_{mmr} \in [0,1]$; temporal weight $\eta \in [0,1]$.

Output: Selected indices $\{\mathcal{S}^{(t)}\}_{t=1}^{T}$ with $|\mathcal{S}^{(t)}| = k_{\text{obj}}^{(t)}$.

Compute patch-query similarities s_{evi} by Eq. (5);
Initialize $\mathcal{S}^{(0)} \leftarrow \emptyset$;
for $t = 1$ **to** T **do**
  Initialize candidates $\mathcal{C} \leftarrow \{1, \ldots, P\}$ and $\mathcal{S}^{(t)} \leftarrow \emptyset$;
  Initialize max-redundancy vector $m[p] \leftarrow 0,\ \forall p \in C$;
  for $i = 1$ **to** $k_{obj}^{(t)}$ **do**
    Select $p^\star \leftarrow \arg\max_{p\in\mathcal{C}} \left(\lambda_{mmr}\, s_{evi}^{(t,p)} - (1-\lambda_{mmr})\, m[p]\right)$;
    Update $\mathcal{S}^{(t)} \leftarrow \mathcal{S}^{(t)} \cup \{p^\star\}$ and $\mathcal{C} \leftarrow \mathcal{C} \setminus \{p^\star\}$;
    Compute similarities between the newly selected token $p^\star$ and all candidates by a matrix–vector product: $\mathbf{s} = \hat{\mathbf{V}}_{patch}^{(t)}\, \hat{\mathbf{V}}_{t,p^\star} \in \mathbb{R}^{P}$;
    Update the running max-sim redundancy: $m[p] \leftarrow \max\big(m[p],\, s[p]\big),\ \forall p \in \mathcal{C}$;
  end
end
return $\{\mathcal{S}^{(t)}\}_{t=1}^{T}$;

A naive MMR implementation recomputes the redundancy term by comparing each candidate with all previously selected tokens, yielding $\mathcal{O}(T^2 \cdot P)$ cost, where T is the frame amount and P is the number of patches within each frame. We reduce it to $\mathcal{O}(T \cdot P)$ by maintaining a running max-sim vector $m[p]$ for each candidate patch p. After selecting a new token $p^\star$, we update m incrementally using a single matrix-vector product and set $m \leftarrow \max(m, \mathbf{s})$.

C Additional Implementation Details

C.1 Benchmarks

Charades-STA [36] contains 3,720 testing queries over 1,334 videos of indoor activities with around 30 seconds per video. The average sequence length after tokenization is 3,909 for Qwen3-VL and 5,322 for Qwen2.5-VL.

ActivityNet-Grounding [5] contains 17,031 queries over 4,885 untrimmed videos, each of which has around 2 minutes describing both indoor and outdoor activities. The average sequence length after tokenization is 10,460 for Qwen3-VL and 11,394 for Qwen2.5-VL.

Video-MME [13] The validation split of Video-MME contains 900 videos (its long-video subset averages about 40 minutes) with 2,700 multiple-choice QA pairs. Following prior settings [34], we use the w/o subtitles pipeline.
LongVideoBench [43] consists of 752 long videos (up to 1 hour) and 1,337 questions across 17 fine-grained categories.

C.2 Unified Protocol

Evaluation pipelines for video-VLM pruning vary substantially (prompts, frame sampling, etc.), which can make comparisons misleading [42]. We introduce **Open V**isual-**P**runing **S**uite **(OpenVPS)**, a unified evaluation protocol to enable fair comparisons for visual pruning. To reduce evaluation cost, we disable the model's thinking-style generation by inserting the `</think>` token immediately after the prompt, following the settings in Time-R1 [40]. The prompts we used are followed by the Qwen3-VL [47], as recorded in Appx. C.4 and C.5.

For preprocessing, we resize videos by setting the shorter side to 480 pixels for Charades-STA [36] and 256 pixels for ActivityNet [5], since ActivityNet clips are substantially longer. This follows the standard practice in OpenTAD [25]. We use the default video loading pipeline of each base model. For Qwen-VL models, we sample raw frames at 2 FPS and apply 2×2 spatiotemporal merging, resulting in an effective visual input rate of 1 FPS. Qwen [3] additionally employs dynamic resolution when the visual token length exceeds a preset threshold, which means the input resolution is automatically reduced. We keep this threshold at 16,384 to match the vanilla setting. This setting with dense frame sampling aligns well with temporal grounding tasks. For LLaVA-OneVision, we uniformly sample 32 frames for VideoQA evaluation, following prior work [34].

For VTG, following [32, 50], we report mean Intersection over Union (mIoU) and $R1$ at tIoU thresholds $m \in \{0.3, 0.5, 0.7\}$, i.e., the percentage of samples whose predicted segment attains IoU larger than m. Auxiliary, we report accuracy for VideoQA in this appendix. For efficiency, we estimate TFLOPs from the final sequence length and model sizes as detailed in Appx. C.3.

All experiments are conducted on three NVIDIA L40 GPUs (48 GB each). Evaluation takes approximately 2 GPU-hours (single L40) on Charades-STA and 18 GPU-hours on ActivityNet. Enabling thinking-style generation substantially increases the runtime while yielding only marginal gains in performance [47].

C.3 TFLOPs

Following prior work [9, 46], we estimate the theoretical Floating-point Operations Per Second (FLOPs) of the Video-VLMs. Qwen [3,47] uses Grouped-Query Attention (GQA) [2] and a SwiGLU-based [33] three-layer feed-forward network. Accordingly, the per-layer FLOPs of the LLM can be expressed as:

$$2nD(h_{kv}d) + 2nD^2 + 2n^2D + 3nDD', \tag{16}$$

where n denotes the number of video tokens, D is the hidden size, D' is the intermediate FFN width, h_{kv} is the number of key/value heads, and d is the head dimension.

C.4 Instruction Prompts for Video Temporal Grounding

```
Given a textual query: {query}.
When does the described content occur in the video?
Please return the final timestamps in seconds directly in one
sentence.
```

C.5 Instruction Prompts for Video Question Answering

```
Select the best answer to the following multiple-choice question
based on the video. Respond with only the letter of the correct
option.
Question: {query}
Possible answer choices: {choices}
The best answer is:
```

D Additional Experiment

D.1 More Retention Ratio of Qwen2.5-VL-based SemVID on Video Temporal Grounding

	ActivityNet-Grounding							Charades-STA						
Method	R1@0.3	R1@0.5	R1@0.7	mIoU	(%)	ER	CS	R1@0.3	R1@0.5	R1@0.7	mIoU	(%)	ER	CS
Qwen2.5-VL-7B, Retain 25% Tokens														
Qwen2.5-VL-7B	29.22	15.77	7.49	22.25	100	$1e^4$	71.8	77.39	59.33	33.82	56.85	100	$1e^4$	90.9
FastV [9]	Out-of-Memory (OOM)					-	-	Out-of-Memory (OOM)					-	-
VScan [51]	Out-of-Memory (OOM)					-	-	Out-of-Memory (OOM)					-	-
DART [42]	17.45	9.11	4.25	13.56	60.9	3.4	20.5	44.41	30.32	15.46	30.11	52.9	3.6	22.0
ToME [4]	23.45	12.16	5.35	18.34	82.4	5.2	28.2	57.02	38.74	18.39	38.44	67.6	4.9	29.1
TokenSculpt [28]	23.42	12.10	5.34	18.41	82.7	5.5	26.6	58.06	38.79	19.01	38.78	68.2	5.1	28.1
FastVID [34]	22.05	11.40	5.12	17.29	77.7	5.0	30.4	64.26	43.79	20.83	41.99	73.8	5.3	33.9
VisionZip [48]	23.92	12.21	5.31	18.64	83.8	5.9	29.3	63.25	42.12	20.10	41.49	73.0	5.7	31.3
SemVID (ours)	**25.06**	**13.07**	**5.85**	**19.50**	**87.7**	**6.1**	**33.1**	**65.73**	**45.19**	**22.00**	**43.66**	**76.8**	**6.0**	**34.8**
Qwen2.5-VL-7B, Retain 33% Tokens														
Qwen2.5-VL-7B	29.22	15.77	7.49	22.25	100	$1e^4$	71.8	77.39	59.33	33.82	56.85	100	$1e^4$	90.9
FastV [9]	Out-of-Memory (OOM)					-	-	Out-of-Memory (OOM)					-	-
VScan [51]	Out-of-Memory (OOM)					-	-	Out-of-Memory (OOM)					-	-
DART [42]	18.62	9.64	4.39	10.88	48.9	1.9	13.6	49.81	33.31	17.45	33.53	59.0	3.9	22.8
ToME [4]	24.13	12.56	5.35	18.84	84.7	6.1	34.3	62.15	42.58	20.13	41.39	72.8	5.5	35.2
TokenSculpt [28]	24.95	12.90	5.69	19.40	87.2	6.7	36.1	65.43	45.32	22.77	43.53	76.6	6.2	35.7
FastVID [34]	24.82	12.66	5.51	19.23	86.4	6.2	38.2	69.41	49.46	24.57	45.98	80.9	6.4	39.6
VisionZip [48]	25.19	12.92	5.75	19.48	87.5	6.8	36.0	67.58	47.69	23.47	44.87	78.9	6.4	36.7
SemVID (ours)	**25.22**	**13.06**	**5.79**	**19.63**	**88.3**	**6.9**	**39.2**	**69.39**	**49.23**	**24.68**	**46.12**	**81.1**	**6.7**	**40.3**

Table 7: VTG results with Qwen2.5-VL under 25% and 33% retention. FastV and VScan materialize full self-attention matrix and lead to OOM in long videos.

Tab. 7 shows that SemVID remains effective under different pruning rates. On both Charades-STA and ActivityNet with Qwen2.5-VL, SemVID consistently

	Motion ↑, Transition ↓		Motion ↑, Transition ↑		Motion ↓, Transition ↓		Motion ↓, Transition ↑	
Method	mIoU	(%)	mIoU	(%)	mIoU	(%)	mIoU	(%)
Qwen3-VL-4B	46.9	100	37.7	100	42.7	100	30.2	100
FastVID [34]	39.6	84.4	33.8	89.7	34.7	81.3	24.7	81.8
SemVID (ours)	**46.5**	**99.3**	**35.1**	**93.1**	**41.3**	**96.7**	**26.9**	**89.1**

Table 8: Sensitivity analysis of motion amplitude and background transition presence under 25% retention ratio.

delivers strong VTG performance at 25% and 33% token retention, achieving the best mIoU scores and performance maintenance under the same budget. These results indicate that our semantic evidence allocation is not tuned to a single compression point but provides a stable accuracy-efficiency trade-off across a wide range of token budgets.

D.2 Sensitivity Analysis of Motion Amplitude

In this section, we examine whether our motion tokens remain effective when the target motion is subtle and the video contains strong background or camera transition artifacts. To identify strong background or camera-induced transition artifacts, we run PySceneDetect, which computes an inter-frame content-difference score based on frame appearance changes in HSV space and flags a scene boundary whenever the score exceeds a threshold. We set the threshold to a relatively low value of 15 so that camera motion is more likely to be detected as a scene change. For motion amplitude, we use GPT-4o-mini [27] to classify each query into small amplitude, large amplitude, or not sure, where the not sure category is used to alleviate potential bias or noise. We then randomly sample instances from the Charades dataset and stratify them into four bins formed by the cross-product of motion amplitude and background transition presence. Sampling continues until each bin contains 100 instances, yielding 400 instances in total. This balanced design mitigates distributional skew and enables a direct comparison of motion-token behavior in the most challenging regime, namely small motions under strong background transitions.

As shown in Tab. 8, VTG accuracy is affected more by background transition artifacts than by action amplitude. For the original setting, introducing strong transitions consistently causes a large mIoU drop (46.9→37.7 and 42.7→30.2), indicating that background or camera changes inject substantial noise into temporal grounding.

FastVID [34] mitigates part of this issue by preserving tokens mainly from sparse anchor frames, which reduces exposure to transition-heavy frames and thus narrows the discrepancy between transition-present and transition-absent cases. However, VTG fundamentally relies on temporally adjacent state changes around boundaries. Aggressively compressing boundary-adjacent moments weakens transition cues and leads to a clear mIoU degradation across all settings.

In contrast, SemVID remains consistently strong under all motion and transition regimes. While our coarse allocation may assign quotas to irrelevant transi-

		Tokens		TFLOPs		VideoMME						
Method	Size	#	(%)	Value	(%)	Short	Medium	Long	Overall	(%)	ER	CS
Qwen2.5-VL-7B	7B	13447	100	124.0	100	76.7	68.2	56.9	67.3	100	$1e^4$	66.3
PruneVID [16]	7B	3295	24.5	23.7	19.1	67.0	59.4	51.6	59.3	88.1	-	-
FastVID [34]	7B	3240	24.1	23.2	18.7	**74.3**	61.3	**52.8**	62.8	93.3	5.2	**33.0**
SemVID	7B	3222	24.0	23.1	18.7	73.8	**62.8**	52.4	**63.1**	**93.7**	**5.4**	30.7

Table 9: VQA results on VideoMME with Qwen2.5-VL-7B under 24% retention.

tions, the subsequent query-filtered motion-token selection suppresses background-dominated changes and prioritizes semantically meaningful motion evidence. This yields uniformly high retention, reaching 99.3% in the easiest setting and maintaining 89.1% even in the most challenging case (Motion ↓, Transition ↑), demonstrating improved robustness to both subtle actions and strong background transition artifacts.

D.3 Performance of Qwen2.5-VL-based SemVID on Video Question-Answering

Although SemVID is primarily developed for VTG, it transfers effectively to general VideoQA. As shown in Tab. 9, under a comparable token budget, SemVID reaches 63.1% overall accuracy on VideoMME with Qwen2.5-VL-7B, outperforming representative pruning baseline PruneVID and yielding a modest but consistent gain over FastVID. Notably, SemVID also achieves higher Evidence Retention (ER) than FastVID, suggesting that preserving structured evidence (diverse object evidence, query-filtered motion cues, and context anchors) benefits not only boundary localization but also general perception tasks. Meanwhile, the relatively smaller role of Connectivity Strength (CS) for VideoQA is expected since many questions can be resolved from one or a few informative frames rather than requiring a temporally continuous evidence chain.

Following the evaluation protocol of FastVID [34], which reports Qwen2.5-VL results only on VideoMME, we restrict Qwen2.5-VL comparisons to VideoMME to ensure a consistent and fair benchmark.

D.4 Performance of LLaVA-OneVision-based SemVID on Video Question-Answering

LLaVA-OneVision is a widely used multimodal model with strong image-level reasoning and a pragmatic video interface, making it a common choice in recent visual token pruning and efficiency-oriented studies. However, LLaVA-OneVision is not intended for fine-grained temporal localization. In its standard inference pipeline, videos are represented by a fixed, sparse set of uniformly sampled 32 frames, and the model does not provide an explicit mechanism to perceive or output timestamps. This sampling strategy is sufficient for general VideoQA tasks whose answers can be supported by object-centric cues from key frames, but it inevitably discards dense motion patterns and long-range temporal dependencies, and therefore no previous study applies LLaVA-OneVision on VTG.

Method	Size	Tokens (%)	LongVideoBench Value	LongVideoBench (%)	VideoMME Short	VideoMME Medium	VideoMME Long	VideoMME Overall	VideoMME (%)
LLaVA-OneVision-7B	7B	100	56.6	100	70.1	56.6	48.8	58.5	100
FastV [9]	7B	25.0	**56.8**	**100.4**	66.0	54.6	47.2	55.9	95.6
VisionZip [48]	7B	25.0	56.0	98.9	68.9	57.4	47.6	58.0	99.1
PruneVid [16]	7B	25.0	55.1	97.3	68.8	54.4	47.7	57.0	97.4
FrameFusion [14]	7B	25.0	54.8	96.8	68.2	55.7	48.6	57.5	98.3
FastVID [34]	7B	25.0	56.3	99.5	**69.9**	56.6	47.7	58.0	99.1
SemVID	7B	25.0	56.4	99.7	68.7	**57.0**	**48.7**	**58.1**	**99.4**
LLaVA-OneVision-7B	7B	100	56.6	100	70.1	56.6	48.8	58.5	100
FastV [9]	7B	15.0	51.5	91.0	58.4	51.7	45.4	51.9	88.7
VisionZip [48]	7B	15.0	54.2	95.8	63.8	54.4	48.3	55.5	94.9
PruneVid [16]	7B	15.0	55.6	98.2	67.9	54.3	48.1	56.8	97.1
FrameFusion [14]	7B	15.0	53.0	93.6	65.8	54.1	46.7	55.5	94.9
FastVID [34]	7B	15.0	56.2	99.3	**69.3**	56.2	47.4	57.7	98.6
SemVID	7B	15.0	**56.4**	**99.7**	68.3	**56.7**	**48.7**	**57.9**	**99.0**

Table 10: VQA results on LongVideoBench and VideoMME with LLaVA-OneVision-7B under 15% and 25% retention.

To align the evaluation with the backbone's strengths and ensure a fair comparison, we focus on VideoQA for LLaVA-OneVision. Moreover, since motion cues are underrepresented under sparse frame sampling (32 frames only for a video of hours length), we set $\alpha = 0.9$ in Eq. (3) and $\lambda_{\text{mmr}} = 0.5$ in Eq. (6) to place greater emphasis on object-centric evidence while improving the diversity of selected semantic cues. This also demonstrates the robustness of our method across different downstream tasks, as it can be readily adapted by simply adjusting the priority of object and motion tokens.

According to Tab. 10, the advantage of our SemVID is most evident on long videos, largely owing to our designs for the diversity (MMR-based object selection) and spatially-temporally structure preserving (semantic budget allocation). In long-form videos, many pruning strategies implicitly concentrate tokens on a few highly salient or query-matched regions, which leads to evidence collapse. By contrast, SemVID explicitly selects diverse semantic evidence and spreads tokens across different objects in different scenes, yielding a compact yet richer evidence pool that better supports long-range aggregation.

D.5 Visualization to Our Semantic Budget Allocation

Fig. 6 reveals that where and how the retained tokens are distributed over time is essential. **ToMe** [4] uniformly pruning spreads tokens over **redundant** frames, under-emphasizing boundary-critical moments. **VScan** [51] relies on **relevance**-driven selection, which repeatedly picks the same query-related regions across frames, collapsing diversity and missing transitions. **FastVID** [34] employs anchor-based aggregation on **salient** patches, sparsifying the timeline and creating token-empty gaps that break cross-frame continuity. In contrast, our **SemVID** explicitly assigns budget according to the **semantic** role of evidence, resulting in a more structured timeline. It increases token density around the ground-truth interval to preserve boundary-critical cues while still reserving a non-trivial budget for intermediate frames.

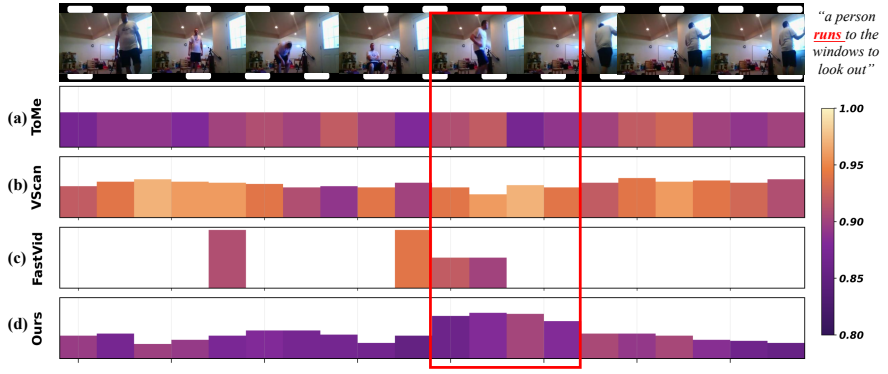

Fig. 6: Per-frame token allocation in VTG. Bar height indicates the **number of retained tokens per frame**; color denotes the average similarity between tokens retained in the current frame and all retained tokens (**darker = more diverse and less redundant**); the red box marks the ground-truth moment. We compare ToMe [4], VScan [51], FastVID [34], and SemVID (ours).

Crucially, the intermediate allocation is not wasted redundancy. It retains motion and context semantics that serve as bridges between temporally separated object evidence, preventing evidence from becoming fragmented and improving cross-frame connectivity for multi-hop reasoning. This budget allocation yields a compact yet diverse evidence set (darker colors) without creating token-empty gaps, which better supports both evidence retention near boundaries and coherent propagation across frames.

D.6 Visualization of the Evidence Retention

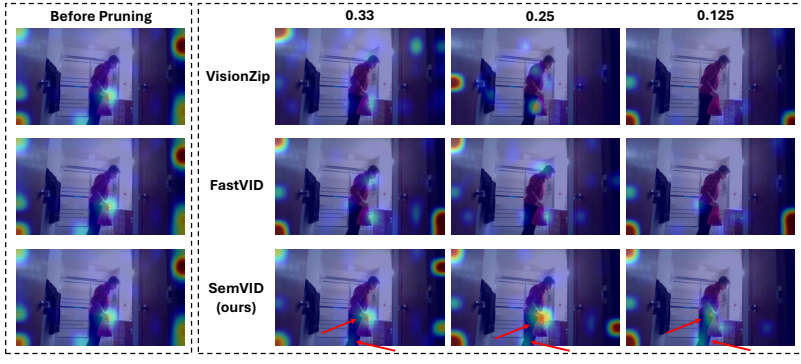

Fig. 7: Comparisons of the attention landing distribution $\boldsymbol{\pi}^{(1)}$ (Eq. 14) over input patches for the query *"the person takes a bag from the bottom cabinet"* on Qwen3-VL-4B, shown under different token retention ratios.

The attention landing distribution $\boldsymbol{\pi}^{(1)}$ is derived by injecting query evidence through the last-layer cross-attention and propagating it backward via cross-layer self-attention over the vision token graph. The resulting distribution at the first layer, referred to as the landing distribution, highlights patches that are most critical to the query. Implementation details are provided in Appx. A.3.

Fig. 7 visualizes the effect of our evidence retention objective. SemVID first selects object-centric tokens to preserve query-critical evidence and further introduces motion tokens to capture foreground semantic transitions, which jointly encourage the attention mass to concentrate on the true evidence regions. Notably, SemVID assigns substantially higher attention to boundary-defining cues such as the red bag and the hands while preserving a large fraction of the pre-pruning attention mass (highlighted by red arrows), whereas other pruning strategies tend to under-attend these regions.