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

Jiaqi Li[1], Shuntian Zheng[1], Yixian Shen[2], Jia-Hong Huang[2,3], Xiaoman Lu[1], Minzhe Ni[1], and Yu Guan[1](✉)

[1] University of Warwick, Coventry CV4 7AL, United Kingdom
{jiaqi.li.16, yu.guan}@warwick.ac.uk
[2] University of Amsterdam, Amsterdam 1012 WX, Netherlands
[3] Amazon AGI, The United States of America

Abstract. Video Temporal Grounding (VTG) localizes the temporal boundaries of query-relevant moments in long, untrimmed videos, making video-language-model prohibitively expensive. While recent training-free token pruning has shown success in video question answering, naively applying these objectives to VTG 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, which keeps query-critical patches especially around event boundaries, and connectivity strength, which preserves 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 budgets by balancing query relevance and inter-frame variation to avoid over-pruned 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 context tokens for scene continuity. Extensive experiments 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. Our code is available here.

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 [8, 25, 34, 49]. 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 [22, 24]. Recently, VTG methods have begun to leverage Video-Language Models (VLMs), benefiting

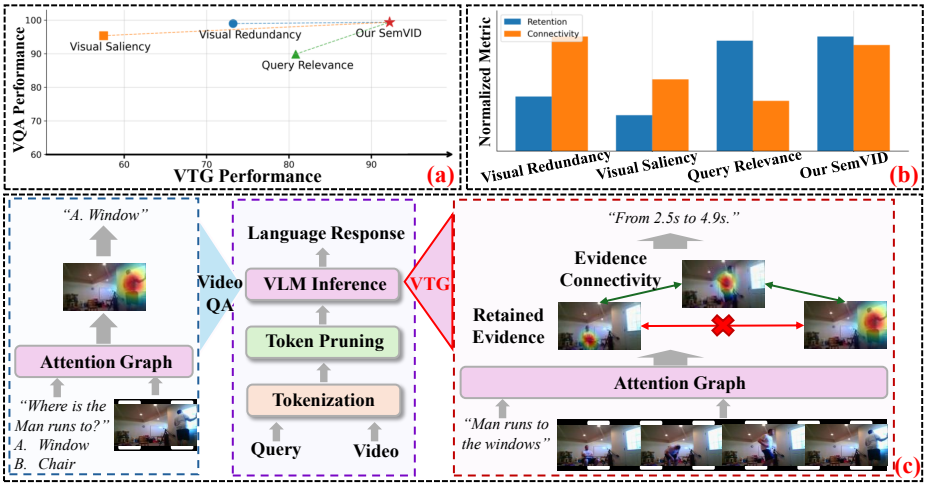

Fig. 1: Comparison between existing pruning objectives and SemVID for VTG. **(a)** Performance comparison between VTG and VideoQA tasks. **(b)** Diagnostics of pruning objectives on evidence retention and cross-frame connectivity. **(c)** VTG requires long-range evidence aggregation rather than a single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

from their strong cross-modal understanding and reasoning ability, and have achieved promising performance on diverse VTG benchmarks [6, 22].

Despite recent progress, deploying VLM-based VTG remains expensive. A video is typically tokenized into thousands of patch tokens, and the attention cost scales quadratically with sequence length [9, 30, 33]. This challenge is amplified for long videos: precise boundary localization often requires dense sampling [35], yet increasing the sampling rate quickly makes prefill consumption prohibitive.

As described in [15, 16], 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 [11, 18, 38], VTG fundamentally differs: it requires temporally coherent evidence to localize event boundaries and to reason about how events evolve over time [6, 39]. 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 [3]. VS prioritizes salient regions, but saliency often concentrates tokens on a few standout frames, leaving large portions of the timeline underrepresented [29]. 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 [7,17].

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 [5,48]. 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 [9, 14], resulting in long visual sequences and quadratic attention cost [33]. This motivates training-free token pruning to accelerate [42]. However, pruning for VTG is particularly challenging. Unlike coarse VideoQA, VTG requires long-range evidence that is both retained and temporally traceable [6,39]. Aggressive pruning can appear safe on coarse QA yet fail on localizations [10].

Visual Redundancy. VR reduces duplicate content by merging or discarding redundant tokens. PruneVid [13] clusters frames into scenes and compresses static tokens within each scene. FastVID [29] further preserves spatiotemporal structure during compression. ToMe [3] merges similar tokens around fixed anchors. TokenSculpt [47] 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 [7, 23, 29, 46]. FastVID computes saliency by vision-encoder attention [29]. However, attention-based scores can be unstable and biased by attention sinks [21, 40], leading to the selected tokens not corresponding to semantically informative regions [12, 37, 42]. 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 [13]. IVTP first estimates intra-vision importance and then filters tokens with instruction-related semantics [12]. LGTTP increases token density for temporally relevant segments based on query cues [17]. Despite their appeal, attention-based relevance can be biased by attention sinks and may even underperform simple baselines [36, 37, 46]. 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 [19, 26, 27]. 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

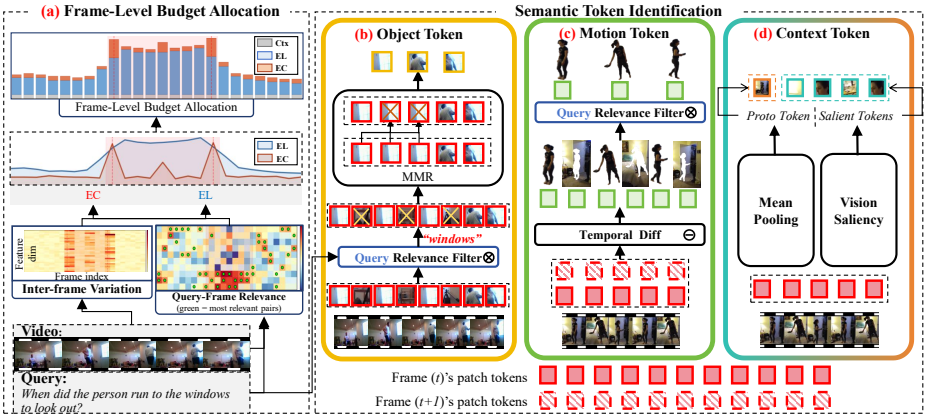

Fig. 2: Overview of SemVID semantic-oriented pruning. (a) **Frame-level budget allocation**: assigns per-frame token budgets by jointly considering query-frame relevance and inter-frame variation. Given per-frame budgets, SemVID then outputs three roles of tokens. (b) **Object token**: uses Maximal Marginal 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.

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.

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. Relying only on query relevance tends to concentrate the budget on a few locally discriminative frames, producing isolated evidence snapshots. By also allocating tokens to transition-rich frames,

SemVID preserves the intermediate state changes that make these evidence snapshots temporally traceable. 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:

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

where $\hat{V}_{\mathrm{glb}}^{(t)} \in \mathbb{R}^{D}$ and $\hat{Q} \in \mathbb{R}^{N \times D}$ denote normalized features of each frame and query tokens. This score provides a lightweight estimate of where query evidence is injected. Since both frame and query features are normalized, high similarity indicates frames whose global semantics are more aligned with the query.

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:

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

Large $s_{\mathrm{EC}}^{(t)}$ 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

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

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

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

This allocation mainly follows the weight $w^{(t)}$, while k_{ctx} serves as an auxiliary safeguard for, avoiding token-empty gaps that would break temporal continuity.

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. It approximates query-to-vision evidence injection using simple feature similarity, avoiding the overhead of materializing full cross-attention maps.

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. MMR selects query-relevant patches while penalizing candidates that are too similar to already selected ones. This encourages diverse object evidence and avoids wasting limited budget on near-duplicate patches around the same region. 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 requires capturing the state transitions and maintaining a traceable path that links evidence across frames. To facilitate cross-frame routing, SemVID introduces motion tokens as explicit relays to capture state changes.

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 preserve relay tokens that

bridge evidence before and after the transition. 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:

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

Given that irrelevant background motions usually have low query relevance, we keep the top-$k_{\text{mot}}^{(t)}$ tokens according to $s_{\text{mot}}^{(t,p)}$, which makes motion selection query-aware by combining temporal change with query relevance. As a result, SemVID can retain meaningful transitions related to the query while suppressing irrelevant background or camera-induced 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 $p_{\text{proto}}^{\star}$ that best represents the frame background by matching the frame-global mean feature:

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

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

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

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 preservation and propagation. We therefore quantify the quality of pruned graph with Evidence Retention (ER) and Connectivity Strength (CS).

Evidence Retention. An effective pruning method should not remove the patches that carry query-critical evidence, so that query could still reach the same visual evidence as in the full video. To measure this, we compare the evidence landing map before pruning, $\boldsymbol{\pi}_{\text{full}}^{(1)}$, with the one after pruning, $\boldsymbol{\pi}^{(1)}$, where $\boldsymbol{\pi}^{(1)}$ denotes the distribution of visual patches that the query finally routes to in the attention graph (Appx. A.3). A direct cross-entropy comparison is

problematic because pruned tokens have zero probability after pruning, which can make the loss infinite. We therefore first compute the amount of original evidence mass that is still kept by the retained token set: $\rho = \sum_{v\in\mathcal{V}'} \boldsymbol{\pi}^{(1)}_{\text{full}}(v)$. Here, ρ measures how much query-induced evidence from the full video remains after pruning. We then use ρ to build a smoothed post-pruning distribution $\bar{\boldsymbol{\pi}}^{(1)}$ and define ER as

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

where $\mathcal{V}' \subseteq \{V^{(1,1)}_{\text{patch}}, \ldots, V^{(T,P)}_{\text{patch}}\}$ is the retained token set. We exponentiate the negative cross-entropy to bound ER in $(0, 1]$. If pruning removes query-critical tokens, $\pi^{(1)}$ will deviate from the full one, leading to lower ER.

Connectivity Strength. To quantify whether pruning preserves a traceable evidence chain, we measure how much attention mass can be routed across adjacent frames, so that information could move smoothly from one frame to the next. At layer ℓ, we define the cross-frame routing mass from frame t to $t{+}1$ as $\mathit{\Gamma}^{(\ell)}_{\mathcal{V}'}(t) = \sum_{u_i\in\mathcal{V}'_t}\sum_{u_j\in\mathcal{V}'_{t+1}} \mathbb{P}^{(\ell)}(u_i\to u_j)$, 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 $\mathit{\Gamma}^{(\ell)}_{\mathcal{V}'}(t)$ over time and layers to obtain CS:

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

where L is the layer amount. Intuitively, larger $\mathit{\Gamma}^{(\ell)}_{\mathcal{V}'}(t)$ means that the retained tokens form stronger temporal links across frames. In other words, the pruned video still provides a connected path for multi-hop evidence propagation, which helps the model localize temporal boundaries more reliably.

4 Experiments

4.1 Experimental Setting

Benchmarks. We evaluate SemVID on Video Temporal Grounding (VTG) in two standard long-video grounding benchmarks, Charades-STA [31] and ActivityNet-Grounding [4]. In the appendix, we also evaluate VideoQA on Video-MME [11] and LongVideoBench [38]. The details of benchmarks are recorded in Appx. C.
Metrics. Following [28,44], 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 [41], Qwen2.5-VL-7B-Instruct [2], and LLaVA-OneVision-7B [20]. To

our knowledge, this is the first study of pruning on Qwen3-VL. We also adapt FastVID [29] and VisionZip [42] to Qwen3-VL as baselines. Unless noted, we set $\alpha = 0.6$ (Eq. 3), $\lambda_{\text{mmr}} = 0.8$ (Eq. 6), $\beta = 0.5$ (Eq. 8), and $k_{\text{ctx}} = 3$ (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.

	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 [42]	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 [29]	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 [42]	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 [29]	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 [42]	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 [29]	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 [42]	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 [29]	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 $1e^{-4}$.

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

Method	ActivityNet-Grounding							Charades-STA						
	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 [7]	Out-of-Memory (OOM)					-	-	Out-of-Memory (OOM)					-	-
VScan [45]	Out-of-Memory (OOM)					-	-	Out-of-Memory (OOM)					-	-
DART [37]	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 [3]	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 [47]	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 [29]	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 [42]	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.

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, indicating 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 [32], and irregular token retention can distort the position-ID trajectory (Fig. 4), introducing and amplifying bias in timestamp perception [10,43]. 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%

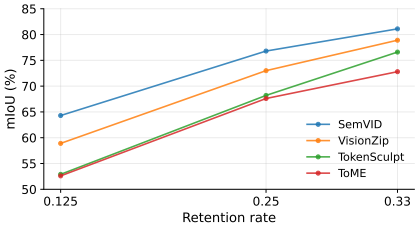

Fig. 3: mIoUs on Charades-STA under different token retention ratios.

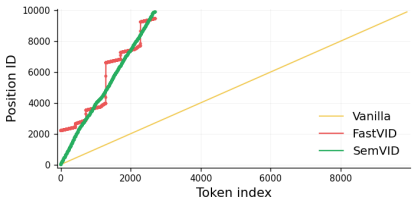

Fig. 4: Position-ID trajectories before and after pruning. Pruning alters the trajectory relative to the original.

	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 [29]			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.

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.

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 Appx. D.5, FastVID relies on sparse anchor frames, which overcompresses 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,

	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
	$\mathbf{k_{\text{ctx}} = 3}$	**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).

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 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

	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 [42]	1307	12.5	4.8	8.7	710.9	184.3	895.2	1.4×	19.89	49.3
FastVID [29]	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.

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*

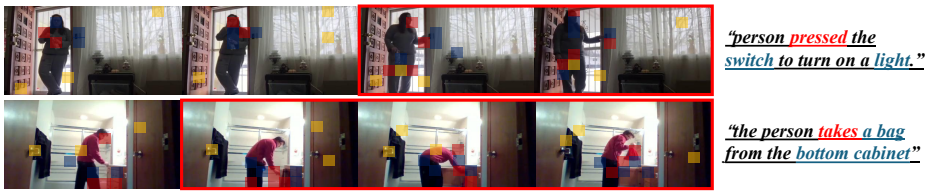

Fig. 5: Visualization results. blue boxes denote object tokens, Red boxes indicate motion tokens, and yellow boxes represent context tokens.

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 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 (Appx. D.5), potentially weakening coverage of subtle actions. A related failure case is that background motion may produce strong local variation and interfere with motion-token selection. In practice, these effects are largely mitigated by our per-frame context floor and query-aware motion filtering: the former prevents weak evidence from being completely pruned, while the latter suppresses query-irrelevant transitions. Nevertheless, more robust motion identification remains an important future direction. 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)

2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. 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)
9. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. 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)
11. 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)
12. 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)
13. 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)
14. Jiang, T., Wu, L., Xia, S., Li, S., Yan, Z., Yang, H., Qiao, Y., Wang, Y.: Imagine before you predict: Interleaved latent visual reasoning for video event prediction. arXiv preprint arXiv:2606.05769 (2026)
15. 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)
16. 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)
17. 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)
18. 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)

19. 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)
20. 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)
21. 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)
22. 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)
23. Liu, Y., Gehrig, M., Messikommer, N., Cannici, M., Scaramuzza, D.: Revisiting token pruning for object detection and instance segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2658–2668 (2024)
24. 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)
25. 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)
26. 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)
27. 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)
28. 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)
29. 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)
30. 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)
31. 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)
32. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
33. 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)

34. 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)
35. 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)
36. 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)
37. 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)
38. 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)
39. 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)
40. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453 (2023)
41. 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)
42. 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 Computer Vision and Pattern Recognition Conference. pp. 19792–19802 (2025)
43. 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)
44. 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)
45. 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)
46. 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)
47. Zhang, S., Li, X., Jiang, T., Zeng, X., Wang, L.: Tokensculpt: Pruning with min-max spatio-temporal duplication for video grounding. OpenReview (2026)
48. 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)
49. 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)