FATE: FRAME-AND-TOKEN EVIDENCE PRESERVATION FOR EFFICIENT MLLM-BASED VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have shown strong capabilities for video temporal grounding (VTG), but processing video inputs incurs substantial computation and memory costs, especially as video duration increases. Existing efficiency methods typically reduce visual computation through frame selection before visual encoding or token compression after encoding, with most approaches focusing on individual reduction stages. However, these stages are closely connected in VTG: frame selection determines which temporal evidence is available for encoding, while token compression determines which fine-grained visual information remains available for localization. We introduce FATE, a training-free framework that coordinates visual reduction across these stages through shared temporal evidence. FATE determines *where to look* through coverage-preserving frame selection and *how much to preserve* through relevance-adaptive token allocation. After visual encoding, query-motion-guided token preservation further determines *what to preserve* within each temporal patch through informative seed selection and coverage-aware completion. Together, these components provide a consistent evidence-guided strategy for preserving grounding-relevant information across the visual processing pipeline without additional training of the grounding MLLM. Experiments on three VTG benchmarks under varying frame and visual-token budgets demonstrate consistent improvements on Ego4D-NLQ and ActivityNet Captions while maintaining competitive performance on Charades-STA.

1 INTRODUCTION

Video temporal grounding (VTG) aims to localize the temporal segment in a video that corresponds to a natural-language query. Recent multimodal large language models (MLLMs) have demonstrated strong semantic understanding and reasoning capabilities for videos (Zhang et al., 2023; Maaz et al., 2024), making them increasingly promising for temporal grounding (Li et al., 2025; Zhang et al., 2026; Zhu et al., 2026a). However, applying MLLMs to VTG incurs substantial computation and memory costs because video frames are encoded into long sequences of visual tokens (Wang et al., 2024; Bai et al., 2025). These costs grow rapidly with video duration as more frames and visual tokens must be processed (Shen et al., 2025; Fan et al., 2026b). This motivates reducing visual processing costs while retaining the information needed for temporal localization.

Existing approaches reduce the computational cost of video MLLMs mainly through two directions: selecting informative frames before visual encoding and compressing visual tokens after encoding. Query-aware frame selection and visual-token compression have demonstrated promising efficiency improvements for video understanding (Zhang et al., 2025; Sun et al., 2025; Zhu et al., 2026b; Shen et al., 2025; Cho et al., 2026; Fan et al., 2026b). Recent VTG-specific methods further introduce task-aware reduction before or after visual encoding, including temporal window selection (Hou et al., 2023) and query-guided token compression (Guo et al., 2025; Fan et al., 2026a; Li et al., 2026b). While effective, these approaches largely focus on individual stages of visual reduction, leaving the interaction between frame selection and token compression less explored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

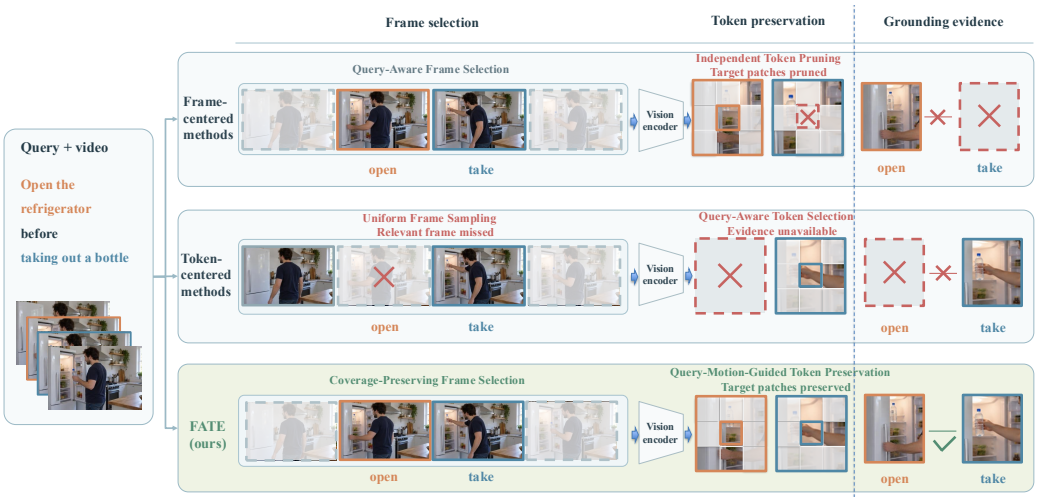

Figure 1: Teaser of FATE. Frame-centered methods can retain query-relevant frames, yet independently applied token pruning may still discard target visual patches. Conversely, token-centered methods cannot recover relevant evidence when it has already been missed by frame sampling. FATE coordinates coverage-preserving frame selection with query-motion-guided token preservation to retain grounding-relevant evidence across both stages.

For MLLM-based VTG, however, frame selection and token compression are connected through the flow of grounding-relevant information. As illustrated in Fig. 1, frame selection determines which temporal content is provided to the visual encoder, while token compression determines which fine-grained information from the selected frames remains available to the grounding model. Information discarded during frame selection cannot be recovered by later token-level processing, and the information retained at earlier stages therefore constrains what can be preserved at later stages. This motivates a cross-stage perspective for efficient MLLM-based VTG, where frame- and token-level reduction decisions are coordinated rather than treated as independent operations.

To this end, we introduce FATE, a training-free framework for coordinating visual reduction across different stages of MLLM-based temporal grounding. FATE first constructs temporal evidence from decomposed query events and their temporal relations. This shared evidence guides *where to look* through coverage-preserving frame selection and determines *how much to preserve* through relevance-adaptive token allocation. After visual encoding, query-motion-guided token preservation further determines *what to preserve* within each temporal patch by selecting informative seed tokens and preserving complementary visual information through coverage-aware completion. Together, these components provide a consistent evidence-guided strategy across frame selection, token-budget allocation, and fine-grained token preservation, without additional training of the grounding MLLM. Experiments across multiple VTG benchmarks and visual budgets, together with component-wise ablations, demonstrate the effectiveness of the proposed framework.

Our main contributions are summarized as follows:

- We introduce a cross-stage perspective for efficient MLLM-based VTG and propose FATE, a training-free framework that coordinates frame-level and token-level visual reduction across the grounding pipeline.
- We develop an evidence-guided reduction strategy in which shared temporal evidence guides coverage-preserving frame selection and relevance-adaptive token allocation, while query-motion-guided token preservation with coverage-aware completion preserves informative and complementary visual information within each temporal patch.
- We evaluate FATE on three VTG benchmarks under varying frame and visual-token budgets, together with component-wise ablations. FATE consistently improves over SemVID on Ego4D-NLQ and ActivityNet Captions across the evaluated token-retention ratios while

remaining competitive on Charades-STA. Under a 512-frame budget on Ego4D-NLQ, FATE improves mIoU from 7.7 to 10.5 compared with SemVID.

2 Related Work

2.1 Video Temporal Grounding with MLLMs

Early VTG methods primarily rely on task-specific architectures to align video and language representations and predict temporal boundaries (Lei et al., 2021; Moon et al., 2023). With the development of MLLMs, recent studies increasingly perform temporal grounding within pretrained video-language models by improving temporal awareness, query representations, or long-video search (Ren et al., 2024; Wu et al., 2025; Pramanick et al., 2025; Hannan et al., 2025; Li et al., 2025; Zhang et al., 2026; Zhu et al., 2026a). These approaches demonstrate the growing potential of MLLMs for temporal grounding, but processing dense video inputs remains computationally expensive, especially as video duration increases (Shen et al., 2025; Fan et al., 2026b). This motivates efficient visual processing strategies that reduce redundant frames and tokens.

2.2 Query-Aware Frame Selection for Video MLLMs

Uniform frame sampling is widely used in video MLLMs (Bai et al., 2025; Zhu et al., 2026a), but it ignores the input query and may allocate limited frame budgets to irrelevant content. Recent methods explore query-aware or adaptive frame selection through text–image matching, list-wise optimization, or learned policies (Zhang et al., 2025; Sun et al., 2025; Yu et al., 2025; Qin et al., 2026; Hu et al., 2025; Zhu et al., 2026b). Other works incorporate query structure or temporal reasoning into frame selection (Li et al., 2026a; Wang et al., 2026; Ben-Ami et al., 2026). Most of these methods target general video understanding rather than temporal grounding, and frame selection alone does not address the large number of visual tokens produced from the retained frames.

2.3 Visual Token Compression for Video MLLMs

A complementary line of work reduces the visual-token sequence through pruning, merging, or representative subset selection. LongVU (Shen et al., 2025) performs adaptive spatio-temporal compression, FLoC (Cho et al., 2026) uses facility-location-based representative selection, while PruneVid (Huang et al., 2025) and FlashVID (Fan et al., 2026b) exploit query relevance or spatio-temporal redundancy for training-free compression. These methods primarily target general video understanding. More recent approaches study visual compression specifically for temporal grounding. VTG-LLM (Guo et al., 2025) introduces slot-based token compression, GroundVTS (Fan et al., 2026a) performs fine-grained query-guided token sampling, and SemVID (Li et al., 2026b) allocates token budgets according to semantic relevance and temporal variation. While these methods demonstrate the value of task-aware visual reduction, the interaction between pre-encoder frame selection and post-encoder token reduction remains comparatively underexplored for VTG.

3 Method

3.1 Problem Formulation and Overview

Given an untrimmed video $V = \{x_t\}_{t=1}^{T}$ and a natural-language query q, video temporal grounding aims to predict a temporal interval $\hat{\tau} = (\hat{t}_s, \hat{t}_e)$ that localizes the moment described by q. We focus on efficient MLLM-based VTG under constraints on both pre-encoder frames and post-encoder visual tokens. Let B_f denote the maximum number of frames selected for the visual encoder, and B_t the maximum number of encoded visual tokens retained for the grounding MLLM.

These two budgets constrain visual information at different stages of the grounding pipeline. Frame selection determines which temporal content is available for encoding, while token reduction determines which fine-grained visual information remains available for localization. This motivates coordinating visual reduction across stages rather than treating the two decisions independently.

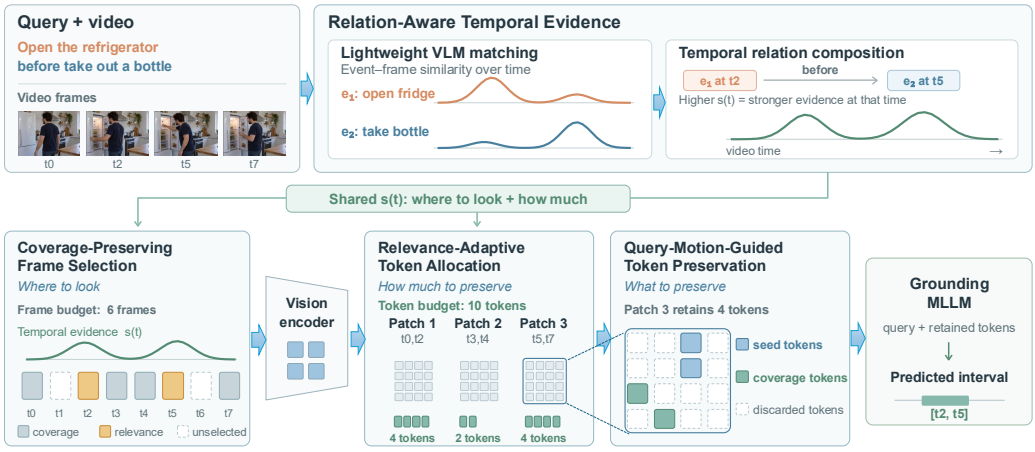

Figure 2: Overview of FATE. A shared relation-aware temporal evidence signal guides pre-encoder frame selection (*where to look*) and post-encoder token allocation (*how much to preserve*). Within each encoded temporal patch, query-motion-guided token preservation selects informative and complementary visual tokens under the allocated budget (*what to preserve*). Together, these components coordinate visual reduction across stages for efficient video temporal grounding.

As illustrated in Fig. 2, FATE first constructs *relation-aware temporal evidence* along the video timeline. The shared temporal evidence guides *coverage-preserving frame selection* to determine *where to look* and *relevance-adaptive token allocation* to determine *how much to preserve*. Within each encoded temporal patch, *query-motion-guided token preservation* further determines *what to preserve* by selecting informative tokens and complementing them with coverage-aware selection.

Sec. 3.2 describes temporal evidence construction, Sec. 3.3 presents frame selection and token allocation, and Sec. 3.4 introduces fine-grained token preservation.

3.2 Relation-Aware Temporal Evidence

The subsequent compression stages require a temporal signal that indicates where grounding-relevant visual evidence is likely to occur. For compositional queries with multiple visual events and temporal relations, we decompose the query into simpler events and compose their temporal evidence according to the expressed relations.

Specifically, we represent the query as $\mathcal{Q} = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E} = \{e_i\}_{i=1}^{M}$ denotes a set of visually grounded events and $\mathcal{R}$ contains their temporal relations. For each event e_i, we compute frame-wise semantic similarity using a lightweight vision-language model, producing an event-wise temporal curve $a_i(t)$. The raw similarity curve is normalized along the temporal dimension using z-score normalization followed by temperature-scaled sigmoid mapping, such that $a_i(t)$ reflects the relative temporal saliency of the event within the current video. Since the video-side frame embeddings are query-independent, they can be precomputed and reused across different queries. We average the normalized event curves to obtain a relation-agnostic baseline $s_{\text{base}}(t)$.

Event-wise matching identifies where individual query components are visually supported, but it does not capture how these events should be organized in time. We therefore explicitly compose the event curves according to the temporal relations in $\mathcal{R}$. For example, given a relation e_i `before` e_j, evidence around e_i is supported by the presence of e_j at a later time, while evidence around e_j is supported by e_i at an earlier time. We define

$$r_{i \prec j}(t) = \max \left\{ a_i(t) \max_{t' > t} a_j(t') e^{-(t'-t)/\tau_r},\ a_j(t) \max_{t' < t} a_i(t') e^{-(t-t')/\tau_r} \right\}, \tag{1}$$

where τ_r controls the temporal decay of cross-event support. The `after` relation is handled by reversing the event order, while `overlap` uses symmetric distance-decayed support between the

two event curves. For queries that explicitly specify the earliest or latest occurrence, the `first` and `last` operators suppress candidates according to competing evidence earlier or later on the timeline.

Each valid temporal relation produces a relation-aware evidence curve. We combine these relation curves with the event curves that are not involved in any explicit temporal relation to obtain $s_{\mathrm{rel}}(t)$. We then interpolate between the relation-agnostic and relation-aware signals:

$$s(t) = (1 - \lambda_r) s_{\mathrm{base}}(t) + \lambda_r s_{\mathrm{rel}}(t), \tag{2}$$

where λ_r controls the contribution of temporal-relation evidence. The resulting temporal evidence $s(t)$ is shared by the subsequent frame-selection and temporal token-allocation stages.

3.3 Evidence-Guided Temporal Allocation

The temporal evidence $s(t)$ constructed in Sec. 3.2 provides a shared signal for coordinating frame selection before visual encoding and temporal token allocation after encoding.

Coverage-Preserving Frame Selection. Uniform sampling provides broad temporal coverage but ignores the grounding relevance indicated by $s(t)$, whereas global top-K selection can concentrate the frame budget around a few relevance peaks. We therefore combine global temporal coverage with relevance-guided sampling under the frame budget B_f. Specifically, we partition the video timeline into fixed-duration temporal blocks. A portion of the frame budget is distributed approximately uniformly across the blocks to maintain temporal coverage. For the remaining budget, we identify the globally highest-scoring frame candidates according to $s(t)$ and allocate additional capacity to blocks containing more such candidates, while using diminishing returns to prevent excessive concentration in a small number of blocks.

Within each block, we place approximately uniform temporal anchors and allow them to shift within local neighborhoods toward more relevant moments. Let K_m denote the number of frames assigned to block $\mathcal{B}_m$, and let $\mathcal{W}_{m,k}$ denote the local search window around its k-th anchor. We jointly determine the selected positions by

$$\{t^*_{m,k}\}_{k=1}^{K_m} = \underset{t_{m,k} \in \mathcal{W}_{m,k},\, t_{m,1} < \cdots < t_{m,K_m}}{\arg\max} \sum_{k=1}^{K_m} s(t_{m,k}), \tag{3}$$

which favors grounding-relevant frames while preserving their temporal ordering and the coarse coverage established by the block-wise allocation.

Relevance-Adaptive Token Allocation. After frame selection, the retained frames are encoded into temporal patches. For each temporal patch i, we compute a patch-level relevance score $\bar{s}_i$ by averaging the temporal evidence over its constituent selected frames. Given the total token budget B_t, each patch first receives a minimum budget $b_{\min}$ so that every encoded temporal patch retains a basic visual representation. The remaining budget is then distributed according to the patch-level relevance:

$$b_i = b_{\min} + \frac{\bar{s}_i}{\sum_{j=1}^{N} \bar{s}_j} \left(B_t - N b_{\min} \right), \tag{4}$$

where N denotes the number of encoded temporal patches. Consequently, more relevant temporal patches retain greater visual capacity while all encoded patches preserve a minimum amount of information. The resulting budgets $\{b_i\}_{i=1}^{N}$ determine how many visual tokens are retained from each temporal patch, while the next section determines which tokens to preserve under these budgets.

3.4 Query-Motion-Guided Token Preservation

Given the per-patch budget b_i, we next determine which visual tokens to retain. Selecting tokens solely according to token-wise relevance may concentrate the limited budget on visually redundant content and overlook complementary information. We therefore first select informative seed tokens using query relevance and local temporal variation, and then use the remaining budget to improve the representational coverage of the selected set.

Query-Motion Seed Selection. For each visual token $z_{i,j}$ in temporal patch i, we consider its relevance to the grounding query and its local temporal variation. Query relevance is measured by the maximum positive cosine similarity between $z_{i,j}$ and the original query-token embeddings obtained from the input embedding space of the grounding MLLM. To capture local temporal variation, we compare the same spatial-token position across adjacent temporal patches:

$$m_{i,j} = \frac{1}{|\mathcal{N}(i)|} \sum_{i' \in \mathcal{N}(i)} \left[1 - \cos(z_{i,j}, z_{i',j})\right], \tag{5}$$

where $\mathcal{N}(i)$ denotes the adjacent temporal patches of patch i. The query-relevance and motion scores are independently normalized within each temporal patch and linearly combined, with λ_q controlling the contribution of query relevance. We select the top $\lfloor \rho b_i \rfloor$ tokens according to the combined score as the initial seed set $\mathcal{S}_i^{(0)}$, where ρ controls the fraction of the per-patch budget reserved for query-motion seed selection.

Coverage-Aware Completion. Query-motion seeds may still focus on similar visual content, leaving complementary information within the same temporal patch insufficiently represented. We therefore use the remaining token budget to improve the representational coverage of the selected set. For a selected set $\mathcal{S}$ in temporal patch i, we define its coverage as

$$C_i(\mathcal{S}) = \sum_u \max_{v \in \mathcal{S}} \left[\cos(z_{i,u}, z_{i,v})\right]_+ . \tag{6}$$

Starting from $\mathcal{S}_i^{(0)}$, we greedily add tokens that yield the largest marginal increase in $C_i(\mathcal{S})$ until the allocated budget b_i is reached. This completion step preserves visual information that is not already well represented by the seed tokens.

4 Experiments

Benchmarks and Metrics. We evaluate FATE on three widely used VTG benchmarks: Ego4D-NLQ (Grauman et al., 2022), ActivityNet Captions (Krishna et al., 2017), and Charades-STA (Gao et al., 2017), covering videos with diverse durations and scenarios. For ActivityNet Captions and Charades-STA, we follow the evaluation subsets used by TimeLens2 (Zhu et al., 2026a), while Ego4D-NLQ is evaluated on its standard validation set. We report mean Intersection over Union (mIoU) and Recall at Rank 1 under IoU thresholds of 0.3, 0.5, and 0.7 (R1@0.3, R1@0.5, and R1@0.7).

Baselines. We use TimeLens2-4B (Zhu et al., 2026a) as the grounding backbone in our main experiments and report its performance without additional visual-token compression as a backbone reference. We compare with SemVID (Li et al., 2026b), a training-free visual-token compression method designed for video temporal grounding, and FlashVID (Fan et al., 2026b), a training-free token-compression method for efficient video MLLMs. For controlled comparisons, FlashVID, SemVID, and FATE are evaluated with the same grounding backbone and maximum input-frame budget, and under the same visual-token retention ratio. Results taken directly from the original papers are explicitly marked in the corresponding tables.

Implementation Details. We follow the dataset-specific preprocessing and generation settings of TimeLens2-4B for the main experiments. Query decomposition is performed using DeepSeek-v4-pro, and FATE uses CLIP ViT-L/14 as a lightweight temporal scout for constructing query-dependent temporal evidence. No additional training or fine-tuning of the grounding MLLM is required. The visual-token retention ratio denotes the fraction of encoded visual tokens retained for input to the grounding MLLM. FATE-specific hyperparameters, query-decomposition prompts, and additional implementation details are provided in the Appendix.

4.1 Main Results under Varying Token Retention Ratios

FATE consistently improves over SemVID on Ego4D-NLQ and ActivityNet Captions across all evaluated token-retention ratios, while remaining competitive on Charades-STA.

Table 1: Temporal grounding performance under different visual-token retention ratios.

	Ego4D-NLQ Long-form				ActivityNet Captions* Medium-length				Charades-STA* Short-form			
Method	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
Uncompressed reference												
TimeLens2†	26.0	18.1	10.0	18.6	74.6	65.5	47.2	59.0	80.3	66.6	37.5	57.7
Retain 25% visual tokens												
FlashVID	11.1	6.7	3.5	7.9	68.5	58.4	41.4	53.7	72.3	57.6	32.3	51.7
SemVID	13.2	8.7	4.7	9.7	72.8	63.6	45.7	57.4	**75.0**	60.9	**34.5**	53.7
FATE	**15.6**	**10.2**	**5.5**	**11.0**	**73.8**	**64.7**	**46.6**	**58.3**	74.3	**61.1**	34.3	**53.8**
Retain 12.5% visual tokens												
FlashVID	10.3	6.3	3.1	7.4	67.6	57.2	39.1	52.3	70.6	55.8	31.5	50.5
SemVID	12.4	8.3	4.5	9.1	72.0	63.2	44.5	56.7	72.3	59.0	32.9	52.0
FATE	**14.4**	**9.4**	**5.1**	**10.2**	**72.8**	**63.9**	**45.5**	**57.4**	**72.9**	**59.0**	**33.0**	**52.2**
Retain 6.25% visual tokens												
FlashVID	8.8	4.9	2.5	6.3	65.0	54.3	37.4	50.2	66.1	51.9	28.7	47.4
SemVID	11.1	7.1	4.0	8.1	69.5	60.6	42.5	54.6	**69.3**	**54.9**	**30.5**	**49.3**
FATE	**12.9**	**8.4**	**4.6**	**9.2**	**70.4**	**61.4**	**43.0**	**55.2**	68.7	54.5	30.0	48.9

* Evaluation is conducted on the subsets sampled by TimeLens2 (Zhu et al., 2026a). † Results are reported from the original TimeLens2 paper.

Table 1 compares FATE with SemVID and FlashVID under different visual-token retention ratios across the three VTG benchmarks. We fix the maximum frame budget to 2048 and vary the final token-retention ratio from 25% to 6.25%. We additionally report the uncompressed performance of TimeLens2 as a reference.

FATE consistently outperforms SemVID on Ego4D-NLQ and ActivityNet Captions across all evaluated token budgets. On Ego4D-NLQ, FATE improves mIoU over SemVID by 1.3, 1.1, and 1.1 points at retention ratios of 25%, 12.5%, and 6.25%, respectively. The gains remain stable as the token budget decreases, showing that FATE maintains its advantage under increasingly constrained visual-token budgets. On ActivityNet Captions, FATE also achieves consistent improvements, with mIoU gains of 0.9, 0.7, and 0.6 points across the three retention ratios.

The difference is smaller on Charades-STA. FATE achieves comparable performance to SemVID at 25% and 12.5% retention, improving mIoU from 53.7 to 53.8 and from 52.0 to 52.2, respectively, while showing a slight decrease from 49.3 to 48.9 under the most aggressive 6.25% setting. Overall, the results show that FATE provides consistent improvements on Ego4D-NLQ and ActivityNet Captions while remaining competitive on Charades-STA across most compression settings.

4.2 Effect of Frame Budget

FATE maintains relatively stable performance as the frame budget decreases, with its largest advantage over SemVID appearing under the 512-frame setting.

We vary the maximum number of frames processed by the vision encoder from 2048 to 512 while fixing the visual-token retention ratio at 12.5%.

As shown in Figure 3, FATE maintains relatively stable performance as the frame budget decreases. At 2048 frames, FATE improves mIoU over SemVID from 9.1 to 10.2. At 1024 frames, FATE achieves 10.7 mIoU compared with 9.0 for SemVID, while under the 512-frame setting it achieves 10.5 compared with 7.7. The recall metrics show the same pattern, with the largest gains over SemVID consistently occurring at the 512-frame budget.

Whereas SemVID degrades substantially at 512 frames, FATE maintains similar mIoU across the three evaluated frame budgets. These results suggest that coverage-preserving, relevance-guided frame selection is particularly useful when only a limited portion of the video can be processed.

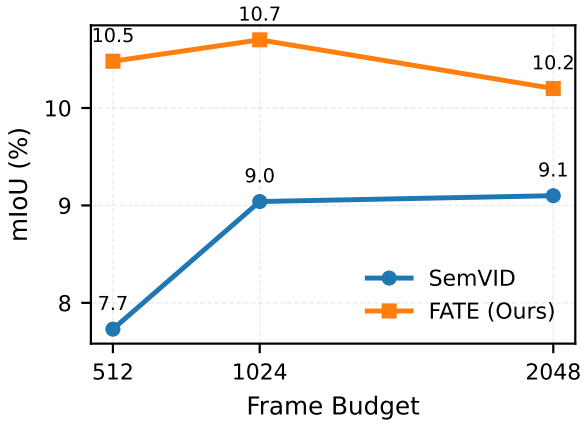

Figure 3: Temporal grounding performance on Ego4D-NLQ under different maximum frame budgets with a fixed 12.5% visual-token retention ratio.

Table 2: Efficiency comparison on Ego4D-NLQ. Runtime is reported per query.

Method	Frames	Encoded Tokens	Retained Tokens	Encoder Time	MLLM Time	mIoU
SemVID	2048	113,351	14,169	4.30 s	3.13 s	9.1
FATE	512	85,917	10,740	3.26 s	2.38 s	**10.5**

4.3 Efficiency Analysis

FATE achieves higher grounding accuracy with substantially fewer input frames and visual tokens, while reducing both visual-encoding and grounding inference time.

We further compare the downstream inference cost of FATE and SemVID using the frame-budget settings in Sec. 4.2. With only 512 input frames, FATE achieves 10.5 mIoU on Ego4D-NLQ, outperforming SemVID with 2048 frames, which obtains 9.1 mIoU. Table 2 reports the number of visual tokens processed by the vision encoder and retained for the grounding MLLM, together with the corresponding runtime. Compared with SemVID, FATE reduces the number of encoded visual tokens from 113,351 to 85,917 and the number of retained visual tokens from 14,169 to 10,740. The measured encoder time decreases from 4.30 s to 3.26 s, while the grounding MLLM inference time decreases from 3.13 s to 2.38 s. These results demonstrate a better accuracy–efficiency trade-off under constrained visual budgets. Query decomposition is performed using an external language-model service and is therefore excluded from runtime measurement.

4.4 Ablation and Design Analysis

Frame Selection and Token Allocation. We first examine how temporal evidence is used for frame selection and token-budget allocation. For each variant, we modify only the indicated component while keeping all other settings identical to the full FATE. Table 3 reports the results on Ego4D-NLQ at a 12.5% visual-token retention ratio.

Uniform frame sampling achieves an mIoU of 9.5, while selecting the globally highest-scoring frames improves it to 10.0. The full coverage-preserving frame-selection strategy further reaches 10.2 mIoU. These results indicate that temporal relevance improves frame selection over uniform sampling, while preserving broader temporal coverage provides an additional benefit over purely selecting the highest-scoring frames.

We further replace relevance-adaptive token allocation with uniform allocation while keeping the remaining components unchanged. Uniform allocation reduces mIoU from 10.2 to 9.9, indicating that distributing token capacity according to temporal evidence provides complementary gains after

Table 3: Ablation of frame selection and token allocation on Ego4D-NLQ at a 12.5% visual-token retention ratio.

Variant	R1@0.3	R1@0.5	R1@0.7	mIoU
Uniform Frame Selection	13.6	8.6	4.3	9.5
Top-K Frame Selection	13.9	9.4	5.0	10.0
Uniform Token Allocation	13.8	9.0	4.8	9.9
Full FATE	**14.4**	**9.4**	**5.1**	**10.2**

Table 4: Ablation of fine-grained token preservation on Ego4D-NLQ at a 12.5% visual-token retention ratio.

Variant	R1@0.3	R1@0.5	R1@0.7	mIoU
Coverage Only	14.1	9.3	4.9	10.1
Query + Coverage	13.8	8.8	5.1	9.8
Motion + Coverage	14.1	9.3	4.6	9.9
Query + Motion Seeds Only	13.9	8.7	4.7	9.8
Full FATE	**14.4**	**9.4**	**5.1**	**10.2**

frame selection. Together, these results support using the same temporal evidence to determine *where to look* through frame selection and *how much to preserve* through token-budget allocation.

Fine-Grained Token Preservation. We next analyze how visual tokens are preserved within each temporal patch. The full strategy combines query relevance and temporal motion for seed selection, followed by coverage-aware completion. We compare this design with variants that remove either the seed-selection stage, the coverage-completion stage, or one of the two seed signals.

Coverage-aware completion alone already provides a strong baseline, reaching 10.1 mIoU, whereas query-motion seed selection without coverage achieves 9.8 mIoU. This highlights the importance of preserving complementary visual information beyond the highest-scoring tokens. Combining query relevance, temporal motion, and coverage-aware completion achieves the best overall performance, reaching 10.2 mIoU and 14.4 R1@0.3. Although the gain over coverage-only selection is modest, query and motion cues can provide additional guidance when combined with coverage-aware preservation.

5 Conclusion

In this work, we presented FATE, a training-free framework that takes a cross-stage perspective on efficient MLLM-based video temporal grounding. Rather than treating frame selection and token reduction as independent operations, FATE coordinates visual reduction across pre- and post-encoder stages to better preserve grounding-relevant information. Specifically, relation-aware temporal evidence guides coverage-preserving frame selection and relevance-adaptive token allocation to determine *where to look* and *how much to preserve*, while query-motion-guided token preservation determines *what to preserve* within each encoded temporal patch. Experiments across three VTG benchmarks and multiple visual budgets show consistent improvements over SemVID on Ego4D-NLQ and ActivityNet Captions while maintaining competitive performance on Charades-STA. Notably, on Ego4D-NLQ, FATE achieves 10.5 mIoU with a 512-frame budget, compared with 9.1 mIoU for SemVID using 2048 frames. Ablation studies further demonstrate the contributions of frame selection, token allocation, and fine-grained token preservation. Overall, our results suggest that coordinating visual reduction across stages provides an effective approach to efficient MLLM-based temporal grounding.

AI Use Statement

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, assist with the interpretation of experimental results, and assist with code generation

and implementation. We did not use generative AI tools to generate synthetic datasets, formulate or prove mathematical claims, translate the manuscript, clean or reformat datasets, or conduct qualitative or thematic data analysis; other required-disclosure tasks were not applicable to this work. Additionally, we used generative AI tools for literature discovery and summarization, organizing the structure of the paper, drafting portions of the manuscript, and editing the text to improve clarity and readability. We reviewed all AI-assisted work, verified AI-assisted code through execution and experimental testing, and independently checked experimental results and literature-related claims. We take responsibility for the final content of this work, including text, claims, code, results, and artifacts produced with the aid of generative AI.

Reproducibility Statement

We provide detailed descriptions of the proposed method and evaluation protocol in the main paper. The Appendix includes the query-decomposition procedure, dataset-specific preprocessing settings, implementation details, and hyperparameters used in our experiments. All comparisons are conducted under controlled frame and visual-token budgets with consistent grounding backbones and evaluation settings. Additional configurations and experiment specifications are included in the supplementary materials, and we plan to release the source code and configuration files to facilitate reproduction of the reported results.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. Himu: Hierarchical multimodal frame selection for long video question answering, 2026. URL https://arxiv.org/abs/2603.18558.

Janghoon Cho, Jungsoo Lee, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. Floc: Facility location-based efficient visual token compression for long video understanding. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 102973–102996, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/a87d3a6b4884108c0dff392c1c7e2d82-Paper-Conference.pdf.

Rong Fan, Kaiyan Xiao, Minghao Zhu, Liuyi Wang, Kai Dai, and Zhao Yang. Groundvts: Visual token sampling in multimodal large language models for video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10408–10418, June 2026a.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. Flashvid: Efficient video large language models via training-free tree-based spatiotemporal token merging. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 98730–98759, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/a0698477f358a6adeab2c2b44f109e5e-Paper-Conference.pdf.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990, 2022. doi: 10.1109/CVPR52688.2022.01842.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. VTG-LLM: Integrating Timestamp Knowledge into Video LLMs for Enhanced Video Temporal Grounding. In *AAAI Conference on Artificial Intelligence*, pp. 3302–3310, 2025. doi: 10.1609/AAAI.V39I3.32341. URL https://mlanthology.org/aaai/2025/guo2025aaai-vtg/.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19012–19022, June 2025.

Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, W.K. Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. CONE: An efficient COarse-to-fiNE alignment framework for long video temporal grounding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8013–8028, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.445. URL https://aclanthology.org/2023.acl-long.445/.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-llm based video frame selection for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712, June 2025.

Xiaohu Huang, Hao Zhou, and Kai Han. PruneVid: Visual token pruning for efficient video large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1024. URL https://aclanthology.org/2025.findings-acl.1024/.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 706–715, 2017. doi: 10.1109/ICCV.2017.83.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11846–11858. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/62e0973455fd26eb03e91d5741a4a3bb-Paper.pdf.

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then ground: Adapting frame selection to query types for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11369–11380, June 2026a.

Jiaqi Li, Shuntian Zheng, Yixian Shen, Jia-Hong Huang, Xiaoman Lu, Minzhe Ni, and Yu Guan. Keeping the evidence chain: Semantic evidence allocation for training-free token pruning in video temporal grounding. *arXiv preprint arXiv:2603.05663*, 2026b.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 64426–64455. Curran Associates, Inc., 2025. doi: 10.52202/085713-2159. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5d2e24df9cfaad3189833b819c40b392-Paper-Conference.pdf.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23023–23033, 2023.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24297–24308, October 2025.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16944–16953, June 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 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 (CVPR)*, pp. 14313–14323, June 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/shen25j.html.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24090–24101, October 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Shihao Wang, Guo Chen, De-An Huang, Zhiqi Li, Minghan Li, Guilin Liu, Jan Kautz, Jose M. Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed

temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24640–24650, June 2026.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13754–13765, June 2025.

Sicheng Yu, CHENGKAI JIN, Huanyu Wang, Zhenghao Chen, Sheng JIn, ZHONGRONG ZUO, Xiaolei XU, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning to query frames for video large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 84154–84179, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d18d208fa9c333483e5724ade7beff0f-Paper-Conference.pdf.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10419–10429, June 2026.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22056–22065, October 2025.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. Timelens2: Generalist video temporal grounding with multimodal llms, 2026a. URL https://arxiv.org/abs/2607.17423.

Zirui Zhu, Hailun Xu, Yang Luo, LIU YONG, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 98463–98486, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/a010577f93a8301dada61f0b97cda923-Paper-Conference.pdf.

APPENDIX OVERVIEW

The appendix provides implementation details, methodological specifications, and additional experimental analyses for FATE. Appendix A describes the experimental setup and implementation details, Appendix B presents the query-decomposition procedure and temporal relation operators, and Appendix C reports additional results on frame budgets, temporal evidence construction, and backbone generalization.

A EXPERIMENTAL AND IMPLEMENTATION DETAILS

A.1 EXPERIMENTAL SETTINGS

Unless otherwise specified, we follow the dataset-specific preprocessing, grounding prompts, and generation settings of TimeLens2-4B (Zhu et al., 2026a). The settings that differ across benchmarks are summarized in Table 5. For ActivityNet Captions and Charades-STA, we use the evaluation subsets and grounding prompts adopted by TimeLens2.

Table 5: Dataset-specific evaluation settings used in our experiments.

Dataset	Evaluation Set	Sampling FPS	Max. Frames
Ego4D-NLQ	NLQ v2 validation	2	2048
ActivityNet Captions	TimeLens2 subset	4	2048
Charades-STA	TimeLens2 subset	4	2048

All runtime measurements are conducted on an AutoDL server equipped with NVIDIA GeForce RTX 4090 GPUs.

A.2 FATE IMPLEMENTATION DETAILS

FATE uses CLIP ViT-L/14 as the lightweight vision-language model for constructing frame-wise temporal evidence. The video-side CLIP embeddings are query-independent and are precomputed and reused across queries. Table 6 summarizes the default FATE-specific hyperparameters.

Table 6: Default FATE-specific hyperparameters.

Component	Hyperparameter	Value
Temporal evidence	Event temperature	1.0
	Relation weight λ_r	0.5
	Ordering decay	8 s
	Overlap decay	2 s
Frame selection	Temporal block duration	15 s
	Uniform coverage ratio	0.5
	Anchor search radius	1 anchor interval
Token allocation	Minimum tokens per temporal patch	1
	Uniform share of remaining token budget	0
Token preservation	Query-score weight λ_q	0.5
	Query-motion seed ratio ρ	0.5

Query decomposition is performed once per query using a fixed prompt and is described in Appendix B. All FATE components are training-free, and no parameters of the grounding MLLM are updated.

B Query Decomposition and Temporal Relation Details

B.1 Query Decomposition

For each grounding query, we perform query decomposition once using DeepSeek-v4-pro with a fixed prompt. The query is represented as $Q = (E, R)$, where E contains visually recognizable events or states and R contains explicit temporal relations between them. Objects, attributes, spatial relationships, and locations are retained within the corresponding event descriptions rather than represented as separate nodes. We support five temporal relations: `before`, `after`, `overlap`, `first`, and `last`. DeepSeek-v4-pro is used only for inference, and the resulting decomposition is used to construct the temporal evidence described below.

The complete query-decomposition prompt used in our experiments is provided below.

Query Decomposition Prompt

```
Decompose one natural-language video temporal grounding query or
    caption into visual evidence useful for localizing its target
    moment.

Return one JSON object with exactly these two keys, with no Markdown
     or explanation.

Schema:
{
"events": [
{
"id": "e1",
"text": "concise visual event or state"
}
],
"temporal_relations": [
{
"relation": "before|after|overlap|first|last",
"subject": {"type": "event", "id": "e1"},
"reference": {"type": "event", "id": "e2"} or null
}
]
}

Rules:

1. Events:
Decompose the query into concise visual events or states that
    provide temporal localization evidence.

Each event should correspond to a visually recognizable situation in
     the video. Split events only when they provide meaningful
    independent localization evidence; otherwise combine them to
    avoid vague or redundant events.

Prioritize the event or state corresponding to the queried moment.
    Do not create separate events for preparatory actions, resulting
     states, background context, or passive conditions unless they
    are necessary for localization or explicitly requested by the
    query.

2. Event text:
Preserve visually useful information including actions, involved
    objects, attributes, spatial relationships, and locations.
```

```
Keep objects involved in actions or states inside the event
    description. Do not create separate events only for object
    appearance, existence, location, or state.

Remove unnecessary narrative phrasing such as "is shown", "is seen",
     or "the video shows". Do not introduce visual content not
    stated or clearly implied by the query.

Use visually grounded descriptions for states. Avoid abstract
    phrases such as "at a specific location" or "somewhere" unless
    no more specific description is available.

3. Temporal relations:
Add relations only when temporal ordering or overlap between
    distinct events provides useful localization evidence.

Use:
- before: A occurs before B.
- after: A occurs after B.
- overlap: two active events occur during approximately the same
    temporal interval.
- first/last: only when the query explicitly asks for the earliest
    or latest occurrence; reference is null.

Do not create relations for background/contextual descriptions. For
    the same pair of events, output only one relation (do not output
     both before(A,B) and after(B,A)).

Do not convert spatial relationships such as "behind", "in front of
    ", "next to", "inside", or "following behind" into temporal
    relations. Keep them in event text unless an explicit temporal
    relation is stated.

Query: {{query}}
```

B.2 Temporal Relation Operators

Let $a_i(t)$ denote the normalized temporal evidence curve of event e_i, and let τ_r and τ_o denote the decay constants for ordered and overlapping relations, respectively.

Before and After. For a relation e_i `before` e_j, we use distance-decayed cross-event support in both temporal directions:

$$r_{i\prec j}(t) = \max\left\{ a_i(t) \max_{t'>t} a_j(t') e^{-(t'-t)/\tau_r},\ a_j(t) \max_{t'<t} a_i(t') e^{-(t-t')/\tau_r} \right\}.$$

The `after` relation is obtained by reversing the event order:

$$r_{i\succ j}(t) = r_{j\prec i}(t).$$

Overlap. For two events expected to overlap in time, we use symmetric distance-decayed support:

$$r_{i\cap j}(t) = \max\left\{ a_i(t) \max_{t'} a_j(t') e^{-|t'-t|/\tau_o},\ a_j(t) \max_{t'} a_i(t') e^{-|t'-t|/\tau_o} \right\}.$$

This operator favors temporal locations where evidence for the two events occurs within a short temporal distance.

First and Last. For queries explicitly requesting the earliest or latest occurrence of an event, we suppress candidates that have competing evidence earlier or later on the timeline. Given an event curve $a_i(t)$, we first define

$$g_i(t) = \left[\frac{a_i(t) - \eta}{1 - \eta}\right]_+,$$

where $\eta = 0.5$ is a fixed evidence floor. The `first` and `last` relation curves are then

$$r_i^{\text{first}}(t) = a_i(t)\left(1 - \max_{t'<t} g_i(t')\right),$$

and

$$r_i^{\text{last}}(t) = a_i(t)\left(1 - \max_{t'>t} g_i(t')\right),$$

respectively. Thus, `first` suppresses later occurrences when strong evidence already appears earlier, while `last` suppresses earlier occurrences when strong evidence appears later.

Each valid temporal relation produces one relation-aware curve. We average these curves together with event curves not involved in explicit relations to obtain $s_{\text{rel}}(t)$. The final temporal evidence is

$$s(t) = (1 - \lambda_r)s_{\text{base}}(t) + \lambda_r s_{\text{rel}}(t),$$

where λ_r controls the contribution of relation-aware evidence. For queries without valid temporal relations, we simply use $s(t) = s_{\text{base}}(t)$.

B.3 Query Decomposition Examples

We provide several illustrative examples of the structured query representation used by FATE.

Single Event. For the query *"What did I pour in a coffee maker?"*, the decomposition contains a single event:

> `e1`: person pours something into a coffee maker

No temporal relation is added for this single-event query.

Ordered Events. For the query *"A person opens the refrigerator before taking out a bottle."*, the decomposition is:

> `e1`: person opens the refrigerator
> `e2`: person takes a bottle out of the refrigerator
> `before(e1, e2)`

The two event curves are composed using the `before` operator.

Ordinal Relation. For the query *"When did the person first open the refrigerator?"*, the decomposition is:

> `e1`: person opens the refrigerator
> `first(e1)`

The `first` operator suppresses later competing occurrences.

C Additional Experimental Results

C.1 Complete Results under Different Frame Budgets

Table 7 reports the complete grounding results corresponding to the frame-budget analysis in Sec. 4.2. We vary the maximum number of frames processed by the vision encoder from 512 to 2048 while fixing the visual-token retention ratio at 12.5%.

Table 7: Complete results on Ego4D-NLQ under different maximum frame budgets with a fixed 12.5% visual-token retention ratio.

Frame Budget	Method	R1@0.3	R1@0.5	R1@0.7	mIoU
512	SemVID	10.7	6.4	3.4	7.7
	FATE	**14.5**	**9.6**	**5.2**	**10.5**
1024	SemVID	12.8	7.6	4.2	9.0
	FATE	**15.0**	**10.0**	**5.4**	**10.7**
2048	SemVID	12.4	8.3	4.5	9.1
	FATE	**14.4**	**9.4**	**5.1**	**10.2**

FATE outperforms SemVID across all reported metrics at each evaluated frame budget. The largest gains occur under the 512-frame setting, where FATE improves R1@0.3 from 10.7 to 14.5, R1@0.5 from 6.4 to 9.6, and mIoU from 7.7 to 10.5. Because the token-retention ratio is fixed, reducing the frame budget also decreases the absolute number of retained visual tokens, making lower-frame settings more restrictive overall.

C.2 Temporal Evidence Construction

We compare the relation-aware temporal evidence used by FATE with a simpler variant that directly matches the original query against video frames, while keeping frame selection, token allocation, and token preservation unchanged.

The two evidence constructions achieve similar performance on both datasets. Relation-aware evidence improves mIoU from 10.1 to 10.2 on Ego4D-NLQ and from 57.3 to 57.4 on ActivityNet Captions, while individual recall metrics vary slightly in both directions. These results suggest that FATE does not depend strongly on a particular temporal-evidence construction; rather, the shared evidence signal mainly provides a common interface for coordinating frame selection and token allocation across stages.

C.3 Backbone Generalization

We further examine whether the gains of FATE persist with a different grounding backbone. Specifically, we replace TimeLens2-4B with Qwen3-VL-4B-Instruct and evaluate FATE and SemVID on the full Ego4D-NLQ evaluation set under the same 12.5% visual-token retention ratio. All other evaluation settings are kept consistent with the corresponding main experiments.

As shown in Table 9, FATE consistently outperforms SemVID with Qwen3-VL-4B-Instruct. mIoU improves from 2.6 to 3.5, while R1@0.3 increases from 3.6 to 4.9, with similar gains at stricter IoU thresholds. Although the absolute performance of Qwen3-VL-4B-Instruct is lower than that of the VTG-specialized TimeLens2-4B backbone, the advantage of FATE over SemVID is preserved. These results suggest that the gains of FATE are not specific to the TimeLens2-4B backbone.

Table 8: Effect of temporal evidence construction at a 12.5% visual-token retention ratio.

	Ego4D-NLQ				ActivityNet Captions			
Evidence	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
Direct Query	14.0	**9.6**	**5.2**	10.1	**72.9**	63.5	**45.6**	57.3
Relation-Aware	**14.4**	9.4	5.1	**10.2**	72.8	**63.9**	45.5	**57.4**

Table 9: Backbone generalization on Ego4D-NLQ using Qwen3-VL-4B-Instruct at a 12.5% visual-token retention ratio.

Method	R1@0.3	R1@0.5	R1@0.7	mIoU
SemVID	3.6	2.1	1.0	2.6
FATE	**4.9**	**3.1**	**1.5**	**3.5**