ADAPTTOKEN: ENTROPY-BASED ADAPTIVE TOKEN SELECTION FOR MLLM LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long video understanding remains challenging for Multi-modal Large Language Models (MLLMs) due to high memory costs and context-length limits. Prior approaches mitigate this by scoring and selecting frames/tokens within short clips, but they lack a principled mechanism to (i) compare relevance across distant video clips and (ii) stop processing once sufficient evidence has been gathered. We propose AdaptToken, a training-free framework that turns an MLLM's self-uncertainty into a global control signal for long-video token selection. AdaptToken splits a video into groups, extracts cross-modal attention to rank tokens within each group, and uses the model's response entropy to estimate each group's prompt relevance. This entropy signal enables a global token budget allocation across groups and further supports early stopping (AdaptToken-Lite), skipping the remaining groups when the model becomes sufficiently certain. Across four long-video benchmarks (VideoMME, LongVideoBench, LVBench, and MLVU) and multiple base MLLMs (7B–72B), AdaptToken consistently improves accuracy (e.g., +6.7 on average over Qwen2.5-VL 7B) and continues to benefit from extremely long inputs (up to 10K frames), while AdaptToken-Lite reduces inference time by about half with comparable performance.

1 INTRODUCTION

Understanding long videos is essential for applications such as embodied AI assistants and multi-modal web agents. Recent multi-modal large language models (MLLMs) Chen et al. (2024); Zhang et al. (2024c; 2025a) have achieved strong question-answering and instruction-following performance on short clips. However, they still struggle with long videos (e.g., hour-long inputs), largely because memory demands and context-length limits constrain both the resolution and the number of frames that can be processed Yao et al. (2026).

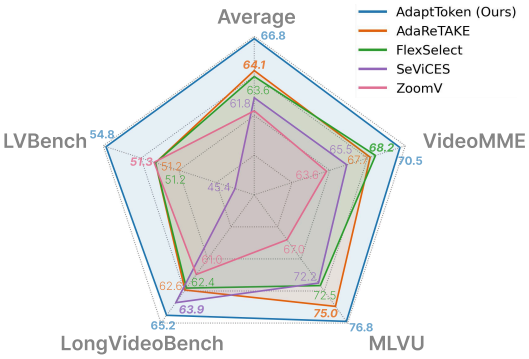

Figure 1: We propose AdaptToken, a flexible and efficient token selection strategy for long video understanding. Compared with state-of-the-art frame/token selection methods on several challenging long-video benchmarks, AdaptToken consistently delivers improved performance.

To mitigate these limitations, recent works Liang et al. (2024); Tang et al. (2025); Hu et al. (2025b); Yao et al. (2025) pre-select relevant or informative frames based on the text prompt and visual content before feeding them into MLLMs. However, frame-level selection often retains substantial irrelevant regions within the chosen frames, which can dilute the useful signal and hurt performance. Alternatively, token-level methods Cheng et al. (2025); Islam et al. (2025); Wang et al. (2025d); Zhang et al. (2025b) directly select or compress the visual tokens provided to the MLLM, typically using cross-modal attention from vision-text encoders Radford et al. (2021) or MLLM layers to score token importance. While this fine-grained approach reduces redundancy compared to frame-level

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.

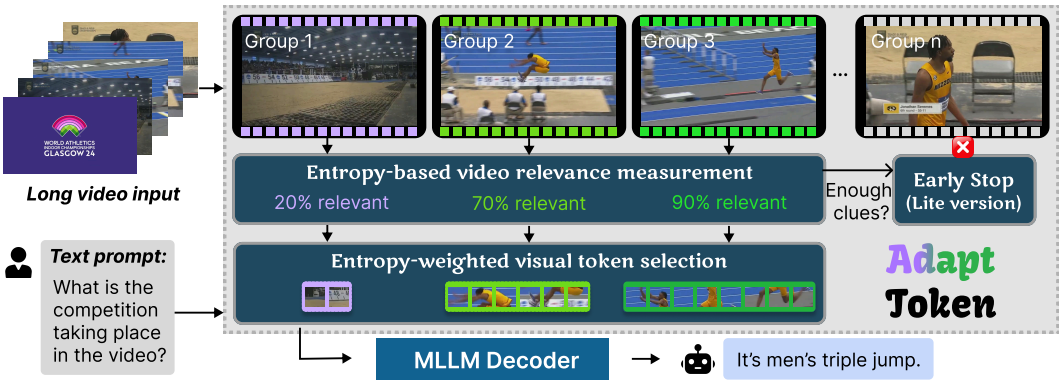

Figure 2: **Overall pipeline of AdaptToken.** AdaptToken processes long videos by dividing them into frame groups and selecting informative tokens within each group based on group relevance estimated from response entropy. It progressively gathers evidence across groups and stops processing once sufficient information has been collected.

selection, two challenges remain. First, token importance is usually computed within individual frames or short clips, lacking a global criterion to allocate tokens across distant clips. Second, the selector typically still processes all sampled frames, incurring unnecessary computation when the MLLM could answer accurately using fewer groups.

To address these limitations, we propose AdaptToken, a training-free long-video token selection framework suitable for various base MLLMs that (i) estimates token importance globally and (ii) decides when to stop examining additional video frames (Fig. 2). Specifically, AdaptToken processes the video group by group to avoid memory and context-length bottlenecks: within each group, it ranks visual tokens via cross-modal attention from a reference MLLM layer, yielding intra-group token relevance scores. To enable global comparisons across groups, we use the model's response entropy as a group-level uncertainty signal: groups that yield more confident (lower-entropy) responses are treated as more prompt-relevant and receive a larger share of the overall token budget. The same entropy signal further supports early stopping (AdaptToken-Lite): once the model becomes sufficiently confident after examining a few groups, we skip the remaining groups to reduce computation. Finally, token importance is not only attributed to prompt relevance. We improve token diversity and temporal coverage with a location-aware global token removal step that suppresses redundant tokens.

We integrate AdaptToken into multiple MLLM architectures Bai et al. (2025b); Chen et al. (2024); Bai et al. (2025a) spanning model sizes from 7B to 72B parameters. Across four challenging long-video benchmarks, VideoMME Fu et al. (2025a), LongVideoBench Wu et al. (2024), LVBench Wang et al. (2025c), and MLVU Zhou et al. (2025), AdaptToken delivers consistent gains over the base MLLMs and prior frame/token selection methods (Fig. 1). Notably, thanks to global awareness during token selection and redundancy removal, AdaptToken scales to extremely long inputs and continues to improve performance even with 10K frames. To the best of our knowledge, prior MLLM long-video understanding methods Cheng et al. (2025); Li et al. (2024c) only report Needle-in-a-Haystack results rather than end-task accuracy at this scale. Our early-stopping variant, AdaptToken-Lite, cuts average inference time by roughly half while maintaining comparable (or even better) performance. In summary, our contributions are threefold.

- We propose AdaptToken, a training-free token selection framework that combines cross-modal attention with response entropy to assess token importance globally for long-video understanding.
- We further use response entropy to trigger early stopping, substantially accelerating inference while maintaining accuracy.
- We demonstrate AdaptToken's effectiveness across diverse MLLMs (multiple architectures and scales) and show consistent improvements on four long-video benchmarks.

2 Related Works

2.1 Multi-modal Large Language Models

Driven by the success of Large Language Models (LLMs) Achiam et al. (2023); Yang et al. (2025); Cai et al. (2024), multi-modal LLMs have emerged to extend text-only models with additional modalities Han et al. (2024); Girdhar et al. (2022); Comanici et al. (2025). Among them, video MLLMs Li et al. (2024a); Wang et al. (2024b); Li et al. (2024b); Xu et al. (2024) aim to provide robust and scalable solutions for understanding and reasoning over video data. A typical architecture comprises a visual encoder and a projection layer that maps visual features into the language embedding space, followed by an LLM backbone that processes the resulting multi-modal token sequence. However, the redundancy in video streams produces a large number of visual tokens, which quickly exhausts the context length and incurs substantial memory and compute overhead. AdaptToken mitigates this challenge by selecting a compact set of informative tokens.

2.2 Long Video Understanding

To process more frames, base MLLMs Zhang et al. (2024b); Bai et al. (2025a) typically extend the context window via long-video training or apply more aggressive temporal/spatial pooling. Nevertheless, efficiency remains a key bottleneck when handling large numbers of frames. Motivated by the redundancy in videos, recent work focuses on extracting prompt-relevant or informative content before feeding it into an MLLM. At the frame level, methods leverage CLIP Radford et al. (2021) vision-text encoders Tang et al. (2025), text-only LLMs Wang et al. (2025f), or learned frame selectors Yu et al. (2024); Yao et al. (2025) to identify relevant frames. However, selected frames may still contain substantial background or irrelevant regions, motivating feature/token-level compression based on cross-modal attention from vision-text encoders Wang et al. (2025b); Mohaiminul Islam et al. (2025) or MLLM layers Wang et al. (2025d); Zhang et al. (2025b). AdaptToken follows this line of work, aiming to improve efficiency while enabling globally informed token selection.

2.3 Certainty Estimation for Model Responses

Humans are more likely to make incorrect statements when uncertain; analogously, recent work studies how to estimate an LLM's confidence to detect or mitigate hallucinations. Self-Evaluation Ren et al. (2023) estimates confidence from the probability of yes/no tokens, while INTUITOR Zhao et al. (2025) learns dedicated confidence tokens via additional finetuning. Confidence can also be inferred directly from the output distribution using metrics such as token-level uncertainty scores Fadeeva et al. (2024), or self-certainty based on KL divergence from a uniform distribution Kang et al. (2025). These estimates have been used to improve test-time reasoning via multi-round voting Fu et al. (2025b). In contrast, AdaptToken leverages model certainty to assess group relevance and derive globally informed token importance for long-video token selection.

3 Method

AdaptToken consists of three main components. Given a long video, AdaptToken first splits it into multiple frame groups. For each group, we compute the group relevance (Section 3.1) and intra-group token importance to enable globally informed token selection (Section 3.2). Next, we apply location-aware global token removal to the selected tokens to improve diversity and coverage beyond relevance (Section 3.3). Finally, AdaptToken supports early stopping: once gathering sufficient evidence, it stops examining additional groups to improve inference efficiency (Section 3.4).

3.1 Group Relevance with Response Entropy

Due to memory and context-length constraints, we cannot feed a large number of video frames (e.g., $>$ 1K frames) into an MLLM at once. A common practice is therefore to process either individual frames Hu et al. (2025a) or small groups of frames Wang et al. (2025d); Zhang et al. (2025b); Pan et al. (2025b) and select relevant frames/tokens within each group. However, inter-group relevance is often poorly handled. Many methods rely on a binary decision: irrelevant frames/groups are either discarded entirely Hu et al. (2025a) or aggressively compressed (e.g., into a single token Cheng et al.

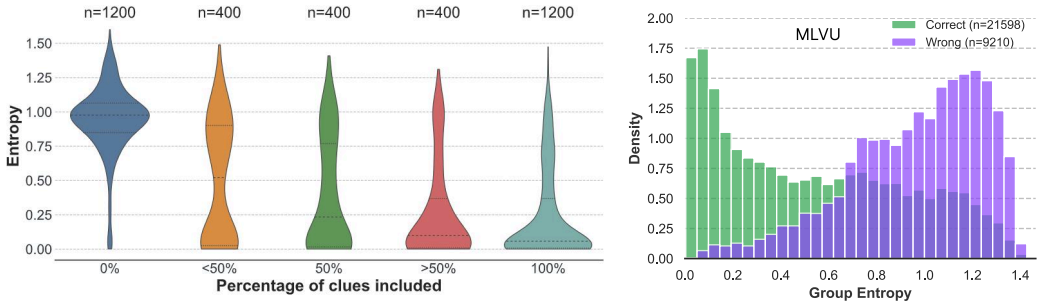

Figure 3: **Response entropy reflects evidence coverage and answer correctness.** We use InternVL2.5 8B with 64 frames. **Left:** Entropy distributions at different clue-coverage levels on a subset of NExT-GQA. **Right:** Entropy distributions for correct and incorrect predictions on MLVU.

(2025)). As the number of groups grows, their contributions to a given prompt can vary substantially, making it beneficial to assess group-wise importance in a finer-grained and continuous manner.

Inspired by recent progress in self-certainty estimation, we use an MLLM's response probabilities to quantify frame-group relevance. Prior work primarily uses self-certainty to improve test-time reasoning in text-only LLMs Kang et al. (2025); Fu et al. (2025b), e.g. by evaluating multiple reasoning traces and voting for the final answer. In contrast, we extend certainty estimation to MLLMs with video inputs and use it to assess the relevance of the input frames to the prompt.

Specifically, given a prefix of input tokens $\{x_1, \ldots, x_i\}$, the LLM backbone autoregressively produces a probability distribution $P_i \in \mathbf{R}^D$ over the next token, where $D = |\mathcal{D}|$ is the vocabulary size. The next token index is $t_{i+1} = \text{argmax}_{j \in \{1,\ldots,D\}} P_i(j)$, and the corresponding token embedding is $x_{i+1} = Embeds[t_{i+1}]$, where $Embeds[\,]$ denotes an embedding lookup table. Beyond selecting the next token, P_i also reflects the model's belief about the output. Denoting the probability of vocabulary item j as $P_i(j)$, the token entropy Kang et al. (2025) is

$$e_i = -\sum_{j=1}^{D} P_i(j) \log P_i(j). \tag{1}$$

Entropy measures the uncertainty of a probability distribution: higher entropy indicates greater uncertainty. We therefore define token certainty as $c_i = -e_i$, and follow Fu et al. Fu et al. (2025b) by averaging certainty over the least-certain portion of the generated response. This focuses the score on uncertain output decisions and limits dilution by highly predictable tokens.

$$C = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} c_i. \tag{2}$$

Here, $\mathcal{G}$ denotes the set of generated tokens with the lowest certainty scores (see Appendix Section C for details). We use C to quantify the relevance of the input frames to the text prompt, based on the hypothesis that higher response certainty indicates that the model has observed more prompt-relevant evidence.

To test whether entropy tracks the amount of prompt-relevant evidence, we use the temporally localized evidence annotations in NExT-GQA Xiao et al. (2024). We construct paired 64-frame inputs from the same question–video pairs while varying the included annotated evidence, and group inputs according to their clue coverage. As coverage increases from 0% to 100%, the response-entropy distribution shifts markedly toward lower values (Fig. 3, left), supporting its use as a group-relevance signal. A complementary Needle-in-a-Haystack experiment in Appendix Fig. 5 compares inputs with and without evidence and distinguishes evidence-supported answers from correct guesses.

We further examine this relationship on long video benchmarks, showing MLVU in Fig. 3 (right) and additional results in Appendix Fig. 6. Given a video with N frames $\{f_1, \ldots, f_N\}$ sampled at a fixed FPS, we partition it into groups of at most K frames, yielding $G = N//K + 1$ groups



Figure 4: **Visualization of AdaptToken token selection.** Two frame groups are presented side by side. For each clip, we first estimate intra-group token relevance via cross-modal attention (heatmaps in the second row), and group-level relevance via response entropy, which measures the model's answer confidence. Based on these signals, we perform global-aware token selection, adaptively allocating a larger token budget to groups that are more relevant to the text prompt (colored masks in the third row). The resulting token set is compact yet information-dense, improving both accuracy and inference efficiency for long-video understanding.

$\{\mathcal{F}_1, \ldots, \mathcal{F}_G\}$. We use interleaved sampling with stride G so that each group spans the full video, while different groups have different temporal sampling offsets. For example, group F_g consists of:

$$\mathcal{F}_g = \{f_g, f_{g+G}, \ldots\}. \tag{3}$$

We then process each group with the MLLM to obtain a response and its associated certainty score. Across benchmarks, higher certainty correlates with a higher probability of answering correctly, indicating that the corresponding frame group contains prompt-relevant evidence.

Compared to prior inter-group relevance measurements Hu et al. (2025a); Cheng et al. (2025), response certainty provides a quantitative and training-free signal for estimating inter-group relevance, which we leverage to derive globally informed token-importance scores.

3.2 Entropy-guided Global Token Selection

With the entropy-based response certainty measure, a video split into G frame groups yields one certainty score C_g per group $\mathcal{F}_g$, which captures group-level relevance. We then estimate token-level relevance within each group, which can be obtained in the same forward pass as C_g.

Token importance can be estimated either using pretrained vision-text encoders Wang et al. (2025e); Mohaiminul Islam et al. (2025) or using attention from MLLM layers Wang et al. (2025d); Zhang et al. (2025b); Wang et al. (2024a). We adopt the latter because MLLMs better capture the semantics of complex text prompts. Concretely, for a given MLLM, we select a late layer with strong retrieval performance (identified via a layer-wise Needle-in-a-Haystack experiment Zhang et al. (2025b)) and use it to compute cross-modal attention. During the same forward pass used to compute group certainty, we extract the visual key embeddings $\{k_1, \ldots, k_V\}$ and text query embeddings $\{q_1, \ldots, q_T\}$ from the selected LLM layer to compute cross-modal attention. We then derive token relevance scores for the visual tokens, $R = \{r_1, \ldots, r_V\}$, by aggregating attention weights over heads and taking the maximum across text queries:

$$r_v = \max_{t \in \{1,\ldots,T\}} \sum_{h=1}^{H} \mathrm{Attn}_h(q_t, k_v), \quad v \in \{1, \ldots, V\}. \tag{4}$$

Here, H denotes the number of attention heads and $\mathrm{Attn}_h(q_t, k_v)$ is the attention weight from query q_t to key k_v in head h. We compute R_g for each frame group to obtain intra-group token-importance scores. To enable global-aware token selection, we adopt a two-stage token-allocation strategy. We

first set an overall token budget B for the entire video, and then allocate a group-level budget B_g for each frame group $\mathcal{F}_g$ from its certainty score C_g using a temperature-scaled softmax:

$$B_g = B \times \Big(\mathrm{Softmax}(\{C_1, \ldots, C_G\}/\tau)\Big)_g. \tag{5}$$

Given B_g, we select the top-B_g visual tokens in group g according to R_g: more relevant groups (higher C_g) retain more informative tokens. Finally, since token location is important for prediction, we also keep the corresponding positional embeddings for the selected tokens.

3.3 Location-aware Global Token Removal

Our two-stage token allocation strategy selects prompt-relevant tokens globally. However, beyond relevance, effective long-video understanding also requires diversity and coverage Tang et al. (2025); Cheng et al. (2025). For instance, neighboring groups $\mathcal{F}_g$ and $\mathcal{F}_{g+1}$ may contain highly similar content. Processing both groups can yield similar certainty scores C and token relevance scores R, causing redundant tokens to be selected.

To mitigate this redundancy, we first select a candidate set slightly larger than the target budget B, leaving room for subsequent removal. We compute the pairwise cosine similarity of candidate token features, S^f. Since redundancy often arises from temporal proximity, we also record the normalized global frame indices of the candidate tokens. The resulting positions d_i express relative location within each video. We then define the temporal similarity term as $S^d_{i,j} = \exp\big(-(d_i - d_j)^2/\sigma\big)$, where σ controls how quickly similarity decays with temporal distance. Appendix Table 11 shows stable MLVU accuracy across the tested decay settings. We combine the two similarities as

$$S = S^f + S^d. \tag{6}$$

Token removal iteratively discards tokens with the highest mutual similarity until the candidate set reaches the target budget B. This reduces redundancy while keeping the context budget unchanged.

3.4 Inference Efficiency with Early-stopping

Splitting the whole video into frame groups largely reduces the memory requirement and also slightly improves the inference speed since the model attentions are computed with shorter lengths. However, all the frame groups still need to be processed and examined, which linearly increases the inference time as the number of frame groups increases. For a super-long video, the MLLM may already gather enough clues using some frame groups without going through the whole video. However, this has been barely addressed by existing frame/token selection methods Wang et al. (2025d); Zhang et al. (2025b); Cheng et al. (2025).

With group certainty C_g, we can decide when to stop examining new frame groups without additional model evaluations. Lower response entropy is associated with more prompt-relevant evidence (Fig. 3, left) and a higher probability of answering correctly on MLVU (Fig. 3, right), but a single confident group may still provide only partial evidence. We therefore stop only after a prescribed number m of examined groups yield low-entropy responses. We use shared stopping settings across the reported benchmarks and backbones (more details are given in Appendix Table 13).

We traverse the interleaved groups in a maximum-margin order to spread the temporal sampling offsets examined before a possible early stop. Appendix Section E.5 specifies the traversal sequence and compares it with contiguous grouping and sequential traversal, supporting the combined design on MLVU. Once stopped, we use the already reviewed frame groups to perform two-stage token allocation and token redundancy removal.

The final selected tokens contain abundant text-prompt-relevant information with high diversity and coverage, which provides a globally consistent compact representation for the original video. We aggregate them following the temporal order and send them again to the MLLM to obtain the final response. Algorithm 1 in Appendix Section B summarizes the procedure.

Table 1: **Comprehensive evaluation across long-video benchmarks.** Results are grouped by model family. Best performance within each block is shown in bold. The first two blocks report representative closed-source and open-source MLLMs as reference points. Subsequent blocks list each base MLLM followed by the corresponding integrated methods.

Model	LLM Size	VideoMME	MLVU	LongVideoBench	LVBench
GPT-4o OpenAI (2024)	-	71.9	64.6	66.7	27.0
Gemini-1.5-Pro Team et al. (2023)	-	73.2	-	64.0	65.7
Gemini-2.5-Pro Comanici et al. (2025)	-	84.3	-	-	78.7
mPLUG-Owl3 Ye et al. (2024)	7B	59.3	63.7	52.1	-
NVILA Liu et al. (2025)	8B	64.2	70.1	57.7	-
ByteVideoLLM Wang et al. (2025a)	14B	64.6	70.1	70.1	-
TPO Li et al. (2025)	7B	65.6	71.1	60.1	-
VideoLLaMA3 Zhang et al. (2025a)	7B	66.2	73.0	59.8	45.3
ViLAMP Cheng et al. (2025)	7B	67.5	72.6	61.2	45.2
InternVL2.5 Chen et al. (2024)	8B	64.2	68.9	59.5	43.4
w/ ZoomV Pan et al. (2025b)	8B	64.4	70.0	63.3	51.5
w/ Triumph Suo et al. (2025)	8B	65.4	70.0	60.7	46.6
w/ SeViCES Sheng et al. (2025)	8B	64.7	72.1	61.7	46.7
w/ FlexSelect Zhang et al. (2025b)	8B	67.0	71.9	60.1	49.7
w/ AdaptToken-Lite (Ours)	8B	68.1	**74.4**	**63.8**	51.3
w/ AdaptToken (Ours)	8B	**68.3**	74.1	63.7	**52.1**
Qwen2.5-VL Bai et al. (2025b)	7B	65.4	70.2	59.5	45.3
w/ TimeSearch-R Pan et al. (2025a)	7B	66.6	71.5	60.1	-
w/ ZoomV Pan et al. (2025b)	7B	63.6	67.0	61.0	51.3
w/ AdaReTAKE Wang et al. (2025d)	7B	67.7	75.0	62.6	51.2
w/ SeViCES Sheng et al. (2025)	7B	65.5	72.2	63.9	45.4
w/ FlexSelect Zhang et al. (2025b)	7B	68.2	72.5	62.4	51.2
w/ AdaptToken-Lite (Ours)	7B	69.8	76.3	65.1	53.3
w/ AdaptToken (Ours)	7B	**70.5**	**76.8**	**65.2**	**54.8**
Qwen2.5-VL Bai et al. (2025b)	72B	73.4	76.3	66.2	47.3
w/ AdaReTAKE Wang et al. (2025d)	72B	73.5	78.1	67.0	53.3
w/ FlexSelect Zhang et al. (2025b)	72B	74.4	76.6	66.4	56.6
w/ AdaptToken (Ours)	72B	**76.1**	**79.8**	**70.5**	**59.7**
Qwen3-VL Bai et al. (2025a)	8B	71.4	78.1	65.9	58.0
w/ AdaptToken (Ours)	8B	**73.8**	**79.3**	**66.9**	**60.6**

4 EXPERIMENTS

4.1 BENCHMARKS AND MODELS

We evaluate AdaptToken on four public long-video benchmarks commonly used by other frame/token selection methods: VideoMME Fu et al. (2025a), MLVU Zhou et al. (2025), LongVideo-Bench Wu et al. (2024), and LVBench Wang et al. (2025c). VideoMME, MLVU, and LongVideoBench include both short and long videos, spanning durations from a few seconds to over an hour and covering diverse tasks, which test robustness across video types and lengths. LVBench instead targets extremely long videos, with an average duration of 4,101 seconds, providing a stringent test of scalability to ultra-long inputs. We evaluate under the LMMS-Eval framework Zhang et al. (2024a) and report the official metrics for each benchmark.

We integrate AdaptToken into multiple MLLMs, including InternVL2.5 8B Chen et al. (2024), Qwen2.5-VL 7B/72B Bai et al. (2025b), and Qwen3-VL 8B Bai et al. (2025a). We primarily benchmark on InternVL2.5 8B and Qwen2.5-VL 7B, two common backbones in prior frame/token selection work, to enable fair comparisons. We additionally evaluate on Qwen2.5-VL 72B to demonstrate scalability to larger models. Finally, Qwen3-VL provides a strong long-context base MLLM via ultra-long-context training; we show that AdaptToken further improves its performance while reducing time and memory consumption. We refer to the full method as AdaptToken, and to the variant with early stopping as AdaptToken-Lite (more implementation details in Appendix B).

4.2 MAIN RESULTS

Comparison with state-of-the-art methods. We integrate AdaptToken into multiple base MLLMs and evaluate on the four long-video benchmarks (Table 1). For fair comparison with prior frame/token selection methods, we focus on two commonly used backbones: InternVL2.5 8B Chen

et al. (2024) and Qwen2.5-VL 7B Bai et al. (2025b). For Qwen2.5-VL 7B, AdaptToken improves performance across all benchmarks, with an average gain of +6.7 over the baseline, and surpasses the previous best results on each benchmark. The largest gains occur on LVBench and MLVU, which contain extremely long videos (often exceeding two hours), highlighting AdaptToken's robustness on ultra-long inputs. We observe a similar trend on InternVL2.5, demonstrating that AdaptToken generalizes well across MLLM backbones.

In summary, across all benchmarks AdaptToken yields larger improvements over the corresponding base MLLMs than competing approaches, demonstrating strong effectiveness for long-video understanding.

Generalization to larger MLLMs. To assess scalability, we further integrate AdaptToken into Qwen2.5-VL 72B. AdaptToken consistently improves performance across all benchmarks (Table 1), achieving an average gain of +5.7 over the 72B baseline and outperforming all existing token-selection methods evaluated at this scale. This indicates that AdaptToken remains effective when applied to larger-scale MLLMs.

Efficiency with early-stopping. Our early-stopping strategy, described in Section 3.4, enables the MLLM to skip uninformative frame groups once sufficient evidence has been collected, thereby reducing inference time. We report the average per-sample inference time for AdaptToken and AdaptToken-Lite (Table 2 and Appendix Table 9). AdaptToken-Lite achieves accuracy comparable to AdaptToken (average difference: -0.7) while reducing inference time by approximately 50% (e.g., from 17.8s to 8.6s on VideoMME). Notably, AdaptToken-Lite still outperforms the previous state of the art. When applied to InternVL2.5 8B, AdaptToken-Lite achieves comparable accuracy overall and slightly outperforms AdaptToken on MLVU and LongVideoBench, further validating the effectiveness of early stopping.

Table 2: **Accuracy and inference-time comparison between AdaptToken and AdaptToken-Lite.** Qwen2.5-VL 7B is used as the base MLLM. AdaptToken-Lite achieves comparable accuracy while reducing average inference time by approximately 50%.

Methods	**AdaptToken-Lite**		**AdaptToken**	
	Acc.	Time (s)	Acc.	Time (s)
VideoMME	69.8	8.6	70.5	17.8
LongVideoBench	65.1	11.0	65.2	18.2
MLVU	76.3	10.1	76.8	21.5
LVBench	53.3	19.3	54.8	32.8

Generalization to advanced MLLMs with long context length. Qwen2.5-VL and InternVL2.5 are trained and evaluated primarily on short video clips, with input limits of 25K tokens for Qwen2.5-VL and 16K tokens for InternVL2.5. By selecting a compact set of informative tokens, AdaptToken effectively increases relevant video content in the input, leading to improved performance. More recent MLLMs, such as Qwen3-VL Bai et al. (2025a), are explicitly designed to handle long-context inputs. Qwen3-VL supports up to 2048 frames and 224K tokens, enabling strong baseline performance on long-video benchmarks (Table 1). We show that AdaptToken remains beneficial even in this setting. When applied to Qwen3-VL 8B and using up to 4096 input frames, AdaptToken yields an average improvement of +1.8 on the four benchmarks while substantially improving efficiency. On LVBench, AdaptToken requires 41 GB of GPU memory and 40.7s per sample on average, compared to 96 GB and 58.5s for the base Qwen3-VL 8B.

4.3 Ablations

Effectiveness of different components. To identify the contributions of different components in AdaptToken, we conduct ablation studies (Table 3). We report results on VideoMME and MLVU, which cover a wide range of video durations. Starting from the InternVL2.5 baseline with 64 input frames (a), simply increasing the maximum number of frames to 256 (b) does not consistently improve performance and even degrades accuracy on MLVU (from 68.9 to 67.2). Adding group-wise token selection while constraining the final context length to a fixed budget B and preserving prompt-relevant tokens (c) leads to a clear performance gain. With

Table 3: **Additive ablation study of AdaptToken components.** TS denotes token selection and TR denotes token removal. InternVL2.5 8B is used as the base MLLM. The final configuration (f) corresponds to the full AdaptToken.

Method	Max frames	VideoMME	MLVU
a. Base InternVL2.5	64	64.2	68.9
b. Base InternVL2.5	256	64.3	67.2
c. b+Group TS	256	66.9	70.1
d. c+Scale up frames	1024	67.1	70.9
e. d+Global TS	1024	68.0	73.6
f. e+Global TR	1024	68.3	74.1

Table 4: **Performance and end-to-end efficiency on MLVU at increasing input-frame caps.** Qwen2.5-VL 7B is the base MLLM. "OOM" denotes out of memory.

Method / Max frames	1,024		2,048		4,096		10,000	
	Acc.	Time (s)	Acc.	Time (s)	Acc.	Time (s)	Acc.	Time (s)
SigLIP2 relevance baseline Tschannen et al. (2025)	66.7	6.5	66.8	8.7	66.7	10.7	65.9	12.6
AdaReTAKE Wang et al. (2025d)	74.6	16.6	75.0	29.7	75.0	35.0	OOM	OOM
FlexSelect-Lite Zhang et al. (2025b)	70.3	6.7	70.1	11.1	OOM	OOM	OOM	OOM
FlexSelect Zhang et al. (2025b)	72.5	10.2	72.7	18.6	OOM	OOM	OOM	OOM
AdaptToken-Lite	74.6	7.2	75.6	9.6	76.3	10.1	76.7	11.4
AdaptToken	74.9	10.8	76.0	18.7	76.8	21.5	77.1	23.4

group-wise token selection in place, further scaling the maximum number of input frames to 1024 (d) yields modest improvements, despite incorporating substantially more video content. This suggests that different frame groups contribute unequally to the task and should not be treated uniformly. Incorporating entropy-guided global token selection (e) significantly outperforms group-wise selection alone by prioritizing more relevant groups. Finally, applying our global token removal procedure (f) further increases relevant token diversity and yields additional performance gains. We reached the same conclusion with Qwen2.5-VL 7B across all four benchmarks (Appendix Table 10).

Runtime breakdown analysis. Profiling with Qwen2.5-VL 7B identifies group-wise inference as the dominant runtime cost. AdaptToken-Lite reduces this bottleneck by evaluating fewer frame groups; the per-stage breakdown is provided in Appendix Section D.3.

Scaling performance and end-to-end efficiency. AdaptToken achieves the highest accuracy across all evaluated frame caps on MLVU (Table 4), improving from 74.9 to 77.1 as the cap rises from 1,024 to 10,000. Both AdaptToken variants remain feasible at 10K frames, where the evaluated AdaReTAKE and FlexSelect configurations run out of memory. Increasing the cap from 4,096 to 10,000 improves accuracy on three of four benchmarks (Appendix Table 8).

AdaptToken-Lite reduces latency by 33–53% relative to full AdaptToken, with accuracy decreases of only 0.3–0.5 points. The substantially lower accuracy of the SigLIP2 baseline further supports using MLLM-derived relevance for token selection (evaluation details in Appendix Section D.2).

Comparison to voting methods. Model self-certainty has primarily been used in prior work for test-time reasoning in LLMs, where multiple reasoning traces are aggregated using voting-based schemes to produce a final answer Kang et al. (2025); Fu et al. (2025b). In contrast, AdaptToken leverages self-certainty to identify prompt-relevant clues within each frame group and performs globally-informed token selection before the final inference pass. We compare AdaptToken with several voting-based aggregation strategies (Table 5). Majority voting selects the most frequent response across groups. Weighted voting assigns weights using group certainty, while Borda voting assigns scores using the groups' certainty ranks. Baseline implementation details are provided in Appendix Section B. While these voting methods improve over the base InternVL2.5 model, their gains are substantially smaller than those achieved by AdaptToken. This comparison underscores the advantage of using self-certainty for globally informed token selection rather than for post-hoc response aggregation.

Table 5: **Comparison to different voting methods.** AdaptToken aggregates tokens from different frame groups in a global-aware manner and obtains better performance over the voting methods (i.e., majority voting, borda voting and weighted voting).

Method	VideoMME	MLVU	LongVideoBench
Base InternVL2.5	64.2	68.9	59.5
Majority voting	64.6	70.9	60.0
Borda voting	64.9	71.1	60.7
Weighted voting	65.3	71.6	61.6
AdaptToken (Ours)	**68.3**	**74.1**	**63.7**

5 Conclusion

We present AdaptToken, a training-free and model-agnostic framework that uses response entropy to guide global token allocation and early stopping for long-video understanding. Across four benchmarks and MLLMs spanning 7B–72B parameters, AdaptToken improves accuracy and scales to 10K frames, while AdaptToken-Lite roughly halves inference time with comparable accuracy. Future work includes improving frame grouping and traversal strategies for better early stopping, and exploring more effective intra-group token relevance scoring methods beyond cross-attention.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing specific functions and individual code blocks. Additionally, we used these tools to aid manuscript writing and polish the language for clarity and readability. All AI-assisted work was subsequently reviewed and verified by the authors. Specifically, we checked the code for functional correctness and reviewed the text for accuracy and consistency with the underlying research. The authors take full responsibility for the final content of this work, including all AI-assisted text, code, and scientific claims.

ETHICS STATEMENT

Our experiments are conducted using existing public benchmarks and publicly available pretrained models. We anticipate minimal ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We will publicly release our code and experimental configurations upon acceptance. Implementation details are provided in Appendix B.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. Scaling video-language models to 10k frames via hierarchical differential distillation. *arXiv preprint arXiv:2504.02438*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin,

et al. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025a.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025b.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16102–16112, 2022.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding. *arXiv preprint arXiv:2502.06428*, 2025a.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13702–13712, 2025b.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. Bimba: Selective-scan compression for long-range video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29096–29107, 2025.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Rui Li, Xiaohan Wang, Yuhui Zhang, Orr Zohar, Zeyu Wang, and Serena Yeung-Levy. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024c.

Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection. *arXiv preprint arXiv:2407.03104*, 2024.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4122–4134, 2025.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. Bimba: Selective-scan compression for long-range video question answering. *arXiv e-prints*, pp. arXiv–2503, 2025.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. *arXiv preprint arXiv:2511.05489*, 2025a.

Junwen Pan, Rui Zhang, Xin Wan, Yuan Zhang, Ming Lu, and Qi She. Timesearch: Hierarchical video search with spotlight and reflection for human-like long video understanding. *arXiv preprint arXiv:2504.01407*, 2025b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In *Proceedings on*, pp. 49–64. PMLR, 2023.

Yuan Sheng, Yanbin Hao, Chenxu Li, Shuo Wang, and Xiangnan He. Sevices: Unifying semantic-visual evidence consensus for long video understanding. *arXiv preprint arXiv:2510.20622*, 2025.

Yucheng Suo, Fan Ma, Linchao Zhu, Tianyi Wang, Fengyun Rao, and Yi Yang. From trial to triumph: Advancing long video understanding via visual context sample scaling and self-reward alignment. *arXiv preprint arXiv:2503.20472*, 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29118–29128, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Han Wang, Yuxiang Nie, Yongjie Ye, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. Dynamic-vlm: Simple dynamic visual token compression for videollm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20812–20823, 2025a.

Lan Wang, Yujia Chen, Du Tran, Vishnu Naresh Boddeti, and Wen-Sheng Chu. Seal: Semantic attention learning for long video representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26192–26201, 2025b.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025c.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*, 2024a.

Xiao Wang, Qingyi Si, Shiyu Zhu, Jianlong Wu, Li Cao, and Liqiang Nie. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5417–5432, 2025d.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024b.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025e.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3272–3283, 2025f.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214. IEEE, 2024.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. *arXiv preprint arXiv:2503.09146*, 2025.

Linli Yao, Long Xing, Yang Shi, Sida Li, Yuanxin Liu, Yuhao Dong, Yi-Fan Zhang, Lei Li, Qingxiu Dong, Xiaoyi Dong, et al. Towards efficient multimodal large language models: A survey on token compression. *Authorea Preprints*, 2026.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024c.

Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. Flexselect: Flexible token selection for efficient long video understanding. *arXiv preprint arXiv:2506.00993*, 2025b.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13691–13701, 2025.

Algorithm 1 AdaptToken for Token Selection

Inputs: Video frames $\{f_1, f_2, ..., f_N\}$, text prompt $\mathcal{T}$, overall token budget B, stopping settings C^* and m
Encode $\mathcal{T}$ into text tokens $\{x_1^t, ..., x_T^t\}$ using vocabulary embeddings $Embeds$
Split the video into frame groups $\{\mathcal{F}_1, ..., \mathcal{F}_G\}$ using Eq. 3
Choose a traversal order $\hat{G}$ that spreads the groups' temporal sampling offsets
Initialize the confidence counter $count = 0$
repeat
  for g **in** $\hat{G}$ **do**
    Encode $\mathcal{F}_g$ into visual tokens $\{x_1^v, ..., x_V^v\}$ using the MLLM vision encoder
    Feed $\{x_1^t, ..., x_T^t\}$ and $\{x_1^v, ..., x_V^v\}$ into the MLLM decoder
    Compute the group-wise certainty C_g using Eq. 1 and Eq. 2
    Compute the token-wise prompt relevance R_g using Eq. 4
    if $C_g > C^*$ **and** early stopping is enabled (AdaptToken-Lite) **then**
      $count \leftarrow count + 1$
    end if
  end for
until $count \geq m$
Update $\hat{G}$ to skip the rest groups. Compute candidate budgets B_g using Eq. 5 with a slightly enlarged overall budget
for g **in** $\hat{G}$ **do**
  Select the top-B_g tokens according to R_g
end for
Compute mutual token similarity matrix S using Eq. 6
Iteratively remove the most similar tokens based on S until B tokens remain
Aggregate the remaining tokens and feed them into the MLLM decoder

A BENCHMARK DETAILS

VideoMME. VideoMME Fu et al. (2025a) is a benchmark for video understanding with diverse video types and durations. It contains 900 videos, with 2,700 manually annotated multiple-choice question-answer pairs across 30 subfields. The dataset is split into three subsets by duration: short (<2 minutes), medium (4-15 minutes), and long (30-60 minutes).

MLVU. MLVU Zhou et al. (2025) spans the widest range of video lengths, from 3 minutes to 2 hours. It includes nine tasks, such as topic reasoning, anomaly recognition, video summarization, and plot question answering.

LongVideoBench. LongVideoBench Wu et al. (2024) targets long-context video understanding with videos up to one hour. It contains 3,763 videos and 6,678 annotated multiple-choice questions across 17 categories, emphasizing referring reasoning that requires retrieving and analyzing multi-modal details from specific temporal segments.

LVBench. LVBench Wang et al. (2025c) focuses on long-video understanding, with an average video duration of 4,101 seconds (4× longer than VideoMME and 5× longer than MLVU). It includes 1,549 annotated multiple-choice question-answer pairs covering tasks such as event understanding, key information retrieval, temporal grounding, and reasoning.

B IMPLEMENTATION DETAILS

Except for the GPU-memory test on Qwen3-VL (conducted on an H200 GPU), benchmark evaluations are distributed over 8 H100 GPUs under the LMMS-Eval framework Zhang et al. (2024a). Per-sample latency is measured with batch size 1 on a single 96GB H100 GPU unless stated otherwise. InternVL2.5 and Qwen2.5-VL are evaluated with a maximum of 16K and 32K input tokens, respectively, following their official settings. We extend both context limits by $16\times$, which increases the maximum number of input frames to 1024 for InternVL2.5 and 4096 for Qwen2.5-VL. These frame counts are upper bounds: frames are sampled at a fixed FPS, and their number is capped

only when the sampled count exceeds the configured maximum. For the token budget B, we follow FlexSelect Zhang et al. (2025b) and use 32 frame-equivalent tokens for fair comparisons, corresponding to 7,010 tokens for Qwen2.5-VL and 8,256 tokens for InternVL2.5. For Qwen3-VL, we increase the budget to 128 frame-equivalent tokens (32,768 tokens). We set the temporal similarity decay parameter to 0.3 in all experiments. The Needle-in-a-Haystack evaluation uses the V-NIAH data introduced in LongVA Zhang et al. (2024b). We combine its needle samples with randomly sampled videos from VideoMME to obtain the correct/incorrect distributions.

For entropy-guided allocation in Eq. 5, we use a softmax temperature $\tau = 2$ across all experiments. For the Borda voting baseline, groups are ranked by certainty and assigned scores $v(r) = (G - r + 1)^p$, where G is the number of evaluated groups, r is the group rank, and $p = 0.9$ following Kang et al. Kang et al. (2025).

C Additional analysis of response entropy

C.1 Needle-in-a-Haystack analysis

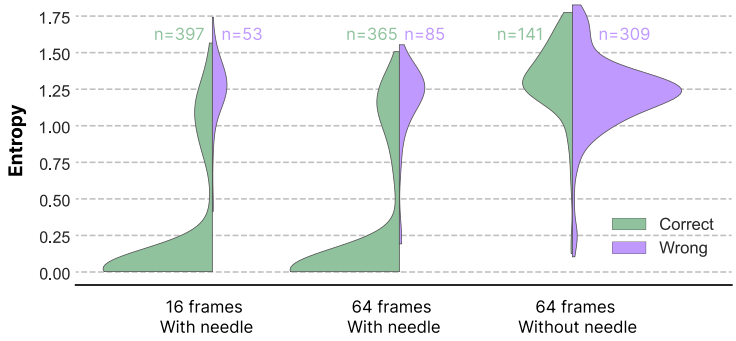

Figure 5: **Needle-in-a-Haystack experiments based on InternVL2.5 8B.** Response entropy distributions for correct vs. incorrect predictions under varying numbers of input frames, with and without needles.

To complement the graded evidence analysis in Fig. 3 (left), we conduct Needle-in-a-Haystack experiments by injecting a needle frame into random videos and asking questions tied to that needle. As shown in Fig. 5, when there is no evidence (i.e., no needle frame) in the input, the MLLM remains highly uncertain even when it guesses correctly, suggesting that the model does not observe prompt-relevant evidence and is instead making a random guess. By contrast, when the input contains the required evidence (i.e., the needle frame), the MLLM typically produces lower-entropy responses. A subset of cases still exhibits high entropy despite the presence of the needle, indicating that the model fails to retrieve or attend to the relevant frame and therefore behaves similarly to the no-evidence condition, making incorrect answers more likely. Moreover, reducing the number of input frames from 64 to 16 decreases the fraction of high-entropy responses, consistent with easier retrieval under shorter contexts.

C.2 Cross-backbone entropy distributions

In the main paper (Fig. 3, right), we analyze group-wise response entropy for InternVL2.5 on MLVU and show that lower entropy is associated with correct predictions, which motivates our entropy-based group ranking and early-stopping strategy. In this appendix experiment, we apply the same analysis protocol to Qwen2.5-VL across four benchmarks to test whether this entropy-correctness relationship generalizes across backbones and datasets. Concretely, we construct 64-frame groups, compute group-wise response entropy under the same prompting/inference setup, and evaluate across all four benchmarks (VideoMME, LongVideoBench, MLVU, and LVBench). As shown in Fig. 6, correct and incorrect cases remain clearly separable in entropy space, and the distributional trend closely matches the InternVL2.5 results. This serves as a direct cross-backbone validation of the main-paper finding: response entropy is a stable uncertainty signal for long-video

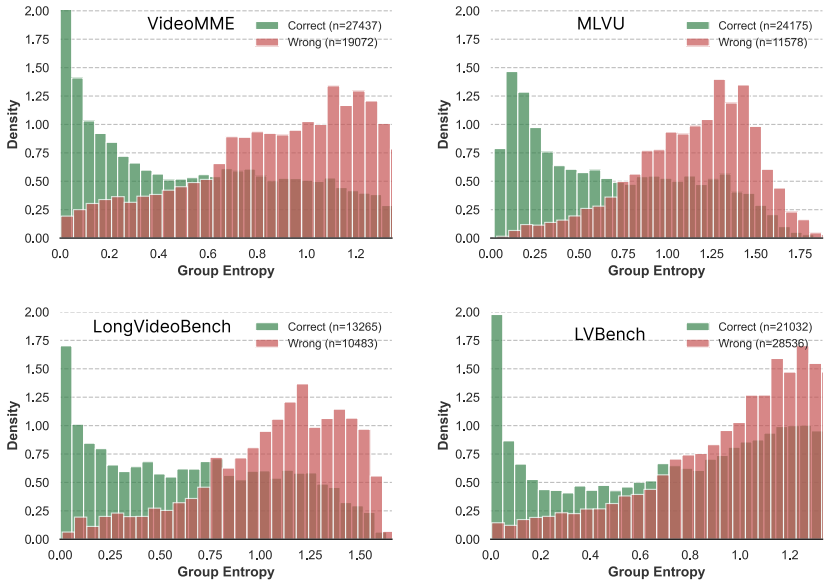

Figure 6: **Entropy distributions across benchmarks.** Qwen2.5-VL 7B is used as the base MLLM.

group selection, enabling the same early-stopping hyperparameters to transfer across models with little or no retuning.

C.3 Additional details for the clue-coverage analysis

For the clue-coverage analysis in Fig. 3 (left), NExT-GQA Xiao et al. (2024) provides temporally localized visual evidence for each question. We construct paired 64-frame groups from the same question–video pairs while varying the included annotated evidence, and evaluate them with InternVL2.5 8B. This paired construction holds the question and source video fixed when comparing evidence levels. We form five bins according to clue coverage: 0%, $(0, 50\%)$, 50%, $(50\%, 100\%)$, and 100%. The numbers displayed above the distributions report the number of evaluated group–question pairs in each bin.

The overlap between the correct and incorrect distributions in the main-paper analysis is expected. A model that misses the relevant evidence may still guess the correct multiple-choice answer, producing a correct but high-entropy prediction similar to the no-needle setting. In contrast, incorrect answers are rare in the genuinely low-entropy region. AdaptToken-Lite therefore stops only in this conservative regime, and its multi-group counter additionally requires the stopping signal to occur in several groups instead of relying on a single confident response.

C.4 Comparison of certainty measures

Table 6: **Comparison of certainty measures.** AdaptToken achieves comparable performance when paired with different self-certainty scores. Qwen2.5-VL 7B is used as the base MLLM.

Method	VideoMME	MLVU
Response KL-divergence	70.0	76.5
Response confidence	70.1	76.7
Response entropy	70.5	76.8

We adopt response entropy to estimate answer certainty, as it demonstrates strong capability in distinguishing relevant from irrelevant frame groups. As discussed in Section 2.3, alternative metrics are also applicable, such as response confidence Fu et al. (2025b) and response KL divergence Kang

et al. (2025). Table 6 compares these measures on VideoMME and MLVU. The results show only marginal differences among the metrics, likely because the generated answers are short (in contrast to long-form mathematical reasoning), where certainty estimates tend to diverge more substantially. Overall, response entropy performs slightly better than the alternatives. We emphasize that our contribution does not lie in proposing a new certainty metric, but in leveraging existing certainty measures to guide input clue estimation for MLLMs.

C.5 Aggregation over low-certainty response tokens

For Eq. 2, we use the lowest-certainty 10% of generated tokens as $\mathcal{G}$ throughout the main experiments. Table 7 compares alternative fractions under the same evaluation setup.

Table 7: **Ablation on the fraction of response tokens used for certainty estimation.** We average the lowest-certainty tokens in each response and evaluate AdaptToken with Qwen2.5-VL 7B on MLVU. The 10% setting is used by default.

Lowest-certainty token fraction	5%	**10%**	20%	50%	100%
MLVU	76.8	76.8	76.8	76.6	76.1

The lowest-certainty tokens have the highest entropy; averaging this lower tail emphasizes uncertain output decisions. Table 7 reports the same MLVU accuracy for fractions of 5–20%, with a modest decrease when a larger fraction is included.

D Additional scaling and efficiency analysis

D.1 Cross-benchmark scaling results

Table 8: **AdaptToken with higher input-frame caps.** Qwen2.5-VL 7B is used as the base MLLM. Frames are sampled at a fixed FPS, subject to the indicated maximum. AdaptToken maintains strong performance as the cap increases to 10K frames.

Max frames	VideoMME	MLVU	LongVideoBench	LVBench
4096	**70.5**	76.8	65.2	54.8
8192	70.2	**77.1**	65.5	54.9
10000	70.1	**77.1**	**65.8**	**55.6**

Table 8 reports accuracy across four benchmarks as the input-frame cap increases from 4,096 to 10,000. Accuracy improves on MLVU, LongVideoBench, and LVBench, with the largest gain on LVBench (54.8 to 55.6). VideoMME decreases slightly from 70.5 to 70.1, indicating that additional frames do not uniformly improve performance across benchmarks.

D.2 Evaluation protocol for the end-to-end efficiency comparison

For Table 4, all methods use Qwen2.5-VL 7B on MLVU, with batch size 1 on a single 96GB H100 GPU. We use the same fixed-FPS candidate-frame sampling and maximum-frame cap for all methods. Videos below the cap retain their FPS-sampled frame count, so increasing the cap need not increase the actual input length or dataset-average latency proportionally.

We report average per-sample end-to-end latency, including visual encoding, token selection or compression, and final answer inference. AdaReTAKE and FlexSelect variants process all candidate frames, whereas AdaptToken-Lite may stop examining new groups once its certainty criterion is met. "OOM" indicates that a method runs out of memory under the evaluated configuration.

We use each method's recommended compression configuration. The methods differ in the representations they compress and the stages at which they incur computation: selecting input visual tokens and compressing layer-wise KV caches produce different storage and processing costs. Equal retained token counts or compression ratios therefore do not define equivalent resource budgets across

methods. This protocol compares practical accuracy–efficiency trade-offs under a shared backbone, candidate-frame sampling, and evaluation hardware; it does not isolate algorithmic differences at a matched retained-token budget.

For the SigLIP2 relevance baseline, we retain the AdaptToken pipeline but replace the MLLM-derived token-relevance scorer with SigLIP2 Tschannen et al. (2025).

D.3 Runtime breakdown analysis

We profile AdaptToken with Qwen2.5-VL 7B and decompose its runtime into four stages. The group-wise inference stage computes response certainty and token relevance. Processing one frame group takes 1.05 s on average, including 0.45 s for visual feature encoding and 0.60 s for LLM-backbone inference. Although the visual encoder is much smaller than the LLM, it incurs comparable latency due to full-attention computation.

The remaining three stages are comparatively inexpensive. Entropy-guided global token selection takes 0.07 s given the group certainties and token-relevance scores from the first stage. Global token removal takes 0.55 s. Finally, MLLM inference over the selected tokens takes 0.31 s due to the reduced input length.

Group-wise inference is therefore the primary computational bottleneck, especially as the number of frame groups increases. By reducing the number of processed groups, AdaptToken-Lite achieves substantial inference-time savings while maintaining comparable accuracy (Table 2 and Table 9).

D.4 Cross-backbone early-stopping efficiency

Table 9: **Accuracy and inference-time comparison between AdaptToken and AdaptToken-Lite.** InternVL2.5 8B is used as the base MLLM. AdaptToken-Lite achieves comparable accuracy while reducing inference time.

Methods	AdaptToken-Lite			AdaptToken		
	Acc.	Groups	Time (s)	Acc.	Groups	Time (s)
VideoMME	68.1	6.9	8.4	68.3	11.0	13.0
LongVideoBench	63.8	7.6	9.2	63.7	11.1	13.0
MLVU	74.4	7.4	9.3	74.1	14.2	16.3
LVBench	51.3	11.3	12.3	52.1	16.2	18.8

In the main paper, we establish early-stopping effectiveness on Qwen2.5-VL 7B (Table 2) using our early stopping version denoted as AdaptToken-Lite. Here, we keep the protocol unchanged and switch only the backbone to InternVL2.5 8B to test whether the gain is model-agnostic rather than architecture-specific. Specifically, we evaluate AdaptToken and AdaptToken-Lite under identical token budgets and benchmark splits, and report three metrics in Table 9: task accuracy, the average number of inferred frame groups per sample, and end-to-end inference time. This appendix experiment therefore serves as a direct cross-backbone validation of the main-paper finding: the entropy-based stopping rule preserves accuracy while reducing computation. Consistent with the Qwen2.5-VL results, AdaptToken-Lite matches or slightly exceeds AdaptToken on LongVideoBench and MLVU, while requiring only about 65% of the inference time. The reduction in processed groups correlates closely with inference time, confirming that group-wise MLLM inference is the dominant runtime cost.

E Additional ablation studies

E.1 Component ablation with Qwen2.5-VL

Table 10 reproduces the additive component analysis with Qwen2.5-VL 7B and supports the same conclusions as the InternVL2.5 results in Table 3. Comparing (b) with (c) shows that group-wise token selection yields consistent gains under the same frame budget. Scaling the candidate frames alone in (d) provides little benefit, whereas entropy-guided global allocation in (e) substantially

Table 10: **Additive ablation of AdaptToken components with Qwen2.5-VL 7B.** TS denotes token selection and TR denotes token removal.

Method	Max frames	VideoMME	MLVU	LongVideoBench	LVBench
a. Base Qwen2.5-VL	768	65.4	70.2	59.5	45.3
b. Base Qwen2.5-VL	1,024	65.1	70.3	59.4	45.6
c. b + Group TS	1,024	68.1	73.0	62.6	52.2
d. c + Scale frames	4,096	68.4	72.8	62.1	52.4
e. d + Global TS	4,096	70.2	75.9	64.7	54.1
f. e + Global TR	4,096	70.5	76.8	65.2	54.8

improves all four benchmarks by prioritizing the most relevant groups. Finally, the global token-removal stage in (f) provides a further consistent gain by reducing redundancy.

E.2 Temporal similarity decay

Table 11: **Ablation on the temporal similarity decay parameter.** Qwen2.5-VL 7B is evaluated on MLVU. The default setting is $\sigma = 0.3$.

σ	0.05	0.1	0.2	**0.3**	0.5	0.8
MLVU	76.3	76.5	**76.9**	76.8	76.7	76.7

Temporal distance is normalized to $[0, 1]$, so σ describes a relative temporal scale within each video. Under uniform temporal sampling, this avoids tying the decay to an absolute frame-index distance. Table 11 evaluates downstream MLVU accuracy across decay settings and shows little variation, with relatively smooth decay ($\sigma > 0.1$) performing well. The default $\sigma = 0.3$ is kept fixed in all main experiments without dataset-specific tuning.

E.3 Token-removal rate

Table 12: **Ablation on the token-removal rate.** AdaptToken is robust to different removal rates.

Removal rate	MLVU
$0.05B$	76.1
$0.1B$	76.8
$0.2B$	76.9
$0.3B$	76.7

AdaptToken initially selects $1.1B$ candidate tokens using entropy-guided allocation, then removes $0.1B$ tokens through global redundancy removal to reach the target budget B. Table 12 varies the number of extra candidates and corresponding removals while keeping the backbone, final token budget, and inference pipeline fixed. The small accuracy differences on MLVU support the default expansion and removal setting.

E.4 Early-stopping hyperparameters

AdaptToken-Lite stops when at least $m = 3$ examined groups have response entropy below $C^* = 0.75$. Lower group thresholds stop more aggressively, whereas larger values require the confidence criterion to be met in more groups; similarly, a higher entropy threshold is more permissive. Table 13 shows similar MLVU accuracy across the tested pairs. We keep the default settings fixed throughout the main experiments.

Table 13: **Ablation on early-stopping hyperparameters.**

Entropy threshold	0.6	0.7	0.7	0.75	0.75	**0.75**	0.8	0.8
Group threshold	1	1	2	1	2	**3**	3	4
MLVU	76.2	75.9	76.2	75.9	76.0	76.3	76.0	76.3

E.5 Frame-group construction and traversal

Interleaved groups follow Eq. 3: each group spans the video, and different group indices correspond to different temporal sampling offsets. Our maximum-margin traversal uses the order $[\mathcal{F}_1, \mathcal{F}_{G/2}, \mathcal{F}_{G/4}, \mathcal{F}_{3G/4}, \ldots]$ to spread the offsets of groups examined early.

Table 14: **Ablation on frame-group construction and traversal order.** InternVL2.5 8B is evaluated on MLVU.

Frame-group input strategy	MLVU
Contiguous video chunks	73.5
Interleaved groups, sequential traversal	74.0
Interleaved groups, maximum-margin traversal	**74.4**

Table 14 compares three grouping and traversal configurations while keeping the remainder of the pipeline fixed. Contiguous groups cover local intervals, whereas interleaved groups span the video. MLVU accuracy improves with interleaving under sequential traversal and improves further with maximum-margin traversal.

F Failure cases and limitations

AdaptToken can struggle when an answer depends on a clue occupying a very small spatial region, such as text on a distant shop sign. Long-video inputs constrain the spatial resolution available per frame, so fine-grained evidence may already be difficult to distinguish during visual encoding or before token selection. In such cases, increasing temporal coverage alone does not recover the missing spatial detail.