Seeing is Believing: Grounding Long-Video Understanding in Spatio-Temporal Visual Evidence

Zhaoyang Wei[1,2*], Guoliang Wang[3*], Guohua Gao[1], Yanchao Hao[2 †], Mingda Li[2], Wenchao Ding[2], Xi Chen[2], Shizhu He[1, 4], Xuehui Yu [2†],

[1] University of Chinese Academy of Sciences
[2] Tencent
[3] University of Sydney
[4]Institute of Automation, Chinese Academy of Sciences
{xuehuiyu, marshao}@tencent.com

Abstract

Although Vision Language Models (VLMs) have excelled at image and video understanding, applying them to hour-long videos is held back by two interrelated challenges: exorbitant computational expense and a qualitative breakdown in long-term temporal reasoning. Thus, models tend to generate answers based on speculation instead of solid visual facts, causing both factually incorrect and plausible hallucinations. This problem is compounded by current benchmarks that, by only emphasizing final answers, lack an effective mechanism to check whether reasoning is substantiated by specific visual evidence. This makes it hard to differentiate between true understanding and pretend comprehension, inhibiting targeted model refinement. To address these interrelated challenges of model fragility and evaluation weakness, we adopt a twofold strategy. First, we present EV^2-Bench, a large-scale benchmark that breaks new ground by an evaluation paradigm built upon spatio-temporal visual evidence, forcing models to justify answers with checkable hints. Second, we put forward DynamicSelect, an adaptive token compression system that efficiently condenses salient information by a dynamic semantic selector and a hierarchical compression strategy. Comprehensive experiments demonstrate that DynamicSelect significantly outperforms the baselines on EV^2-Bench as well as other public benchmarks. Our study offers not only a more effective approach to long-video understanding but also a more stringent evaluation paradigm, indicating the way toward more robust models.

Introduction

The rapid development of Vision Language Models (VLMs) has greatly improved the comprehension of visual content. Although success on image and short video understandings is remarkable, understanding long-form video is still a challenging frontier. Being able to reason over hour-length videos is important for tasks such as analyzing movie plots or comprehending in-depth documentaries. Yet, applying current models to this space is non-trivial. The sheer amount of visual tokens from lengthy videos exceeds the context window of most SOTA transformers and entails prohibitive computation cost, an issue well-noted in recent works (Bai et al. 2025; Zhang et al. 2024c; Chen et al. 2025; Zhang et al. 2025c,b).

This computational bottleneck obscures a more fundamental problem: a qualitative failure at sustained temporal reasoning. Models tend to have a weak perceptual understanding of events far from the beginning or end of the video and resort to response by speculation instead of concrete visual evidence. This issue is compounded by a major shortfall in existing evaluation practices. While existing long-video reasoning benchmarks (Wu et al. 2024; Zhang et al. 2024a; Shen et al. 2024) have made advances, they overwhelmingly emphasize high-level summarization or temporal ordering. Importantly, as recent work has pointed out (Bai et al. 2025), they do not incorporate a stringent mechanism to check if a model's response is rooted in specific, checkable spatio-temporal evidence. This lack of process-level verification means that it is hard to tell apart legitimate reasoning from plausible invention, which makes targeted model improvement challenging. To combat these coupled challenges of model fallibility and evaluation insufficiency, we present a two-fold solution. First, to bridge the essential evaluation gap, we present EV^2-Bench, a novel long-video reasoning benchmark that requires grounding in spatio-temporal visual evidence. Second, to address the information-processing issue underlying model failures, we introduce DynamicSelect, an adaptive token compression framework to enable models to effectively condense salient information from long and redundant video streams.

Our benchmark, EV^2-Bench, is built upon a collection of 1,000 long-form videos, each exceeding 30 minutes with an average duration of 43 minutes. It is specifically designed to assess deep reasoning capabilities by focusing on both intra-event complexities (e.g., dynamic shifts in character relationships, motivations behind actions) and inter-event causal links (e.g., how an action in a prior event precipitates a consequence in a later one). The cornerstone of EV^2-Bench is its novel evaluation paradigm: for every reasoning question, the ground truth contains not only the correct answer but also a set of precise spatio-temporal visual clues—timestamps

*These authors contributed equally.
†Corresponding author.

paired with descriptions of what occurs in that specific region of the frame—that serve as the necessary evidence. Consequently, our evaluation metric assesses a model not just on its final answer, but on its ability to provide the correct supporting evidence, thereby directly measuring its reasoning fidelity and penalizing hallucination. While EV^2-Bench closes the critical evaluation gap by requiring evidential support, it is also essential to address the fundamental information-processing challenge underlying model failures in long-form video comprehension. To this end, we present DynamicSelect, an adaptive token compression framework to enable models to effectively distill salient information from redundant and long video streams. Going beyond static token pruning methods (Chen et al. 2024; Xing et al. 2024), DynamicSelect leverages a dynamic, multi-stage compression procedure. It starts with a dynamic semantic selector that adaptively filters an initial set of visual tokens according to query relevance. This is followed by a hierarchical compression approach: the most relevant tokens are processed by a lightweight bipartite merge to remove local redundancy (e.g., background noise), while a subset of high-scoring but initially unselected tokens are subjected to high-ratio compression. This new "remove redundancy, supplement key information" strategy effectively generates a compact yet complete representation of the video for the model. Our key contributions are threefold:

- We introduce EV^2-Bench, a new benchmark for long-video comprehension that grounds model reasoning in spatio-temporal evidence to distinguish true understanding from plausible generation.
- We propose DynamicSelect, an adaptive token compression method that dynamically extracts salient information from long videos to overcome computational bottlenecks and improve long-term temporal reasoning.
- Through large-scale experiments, we show that DynamicSelect significantly outperforms baselines on multiple benchmarks, while our analysis also reveals the persistent struggles of SOTA models with evidence-based reasoning, highlighting the importance of our work.

Related Work

Long Video Benchmark. Current evaluation systems have moved away from judging short video clips (Yu et al. 2019; Lei et al. 2018) to more focused video reasoning tasks (Song et al. 2024; Wang et al. 2024b; Wu et al. 2024); however, the majority of current benchmarks rarely ask models to provide justification for their answers with explicit visual evidence. By not considering this aspect, they prevent the evaluation of a model's true reasoning capacity (Cobbe et al. 2021; Zeng et al. 2023). Given this, our introduce hour-long videos for **event reasoning**, uniquely requiring models to substantiate answers with precise **spatio-temporal visual evidence**.

Video Token Compression. The issue of efficiently handling long videos has accelerated significant research aimed at the reduction of tokens. Existing methods include attention-priority pruning (Chen et al. 2024; Fu et al. 2024) and the creation of dynamic visual tokens through sampling or fusion methods (Jin et al. 2024). However, these methods are suboptimal, often relying on defective heuristics like attention scores (Zhang et al. 2024b) or neglecting the factor of cumulative redundancy. Our work presents a spatiotemporal adaptive approach that considers both cross-modal queries and inter-frame connections to maximize the effectiveness of redundancy removal.

Task and Benchmark Construction

Task Taxonomy

We propose a hierarchical task taxonomy to assess multi-step reasoning in long videos, shown in Figure 3. It is organized along two central viewpoints: 1) **Intra-Event Understanding:** This tests a model's capacity for fine-grained analysis within one continuous event. Tasks prompt models to go beyond core perception to more in-depth reasoning regarding targets, time, space, and the scene's effect on character actions. 2) **Inter-Event Reasoning:** As the core contribution of our benchmark, this assesses the critical capability of reasoning across multiple, non-adjacent events. It requires models to construct a "web of causality" by linking distant events, track the evolution of characters and relationships, and achieve a meta-cognitive understanding of the story's structure and themes.

Benchmark Construction

We build our benchmark atop an elaborate, multi-stage pipeline aimed at vetting long-video reasoning (Figure 1). Rather than naive genre filtering, we use a quantitative, AI-based framework to determine the **intrinsic reasoning potential** of every video sourced from sites like YouTube and Bilibili. This winnows out a dataset of videos that are not just long, but are also dense in involved narrative and causal structure. Our criteria for selection are: (1) technical requirements of 30–90 minutes duration and above 720p resolution; (2) thematic inclusion of narrative-heavy content such as **movies**, **documentaries**, and **in-depth news reports**; and (3) exclusion of procedural or descriptive content such as **tutorials**, **product reviews**, and **lectures**.

Analyzability Assessment For each curated video, we assess its **Analyzability**. First, a VLM (Zhang et al. 2025d) generates a dense, time-stamped "visual script". A panel of LLMs (e.g., GPT-4o, DeepSeek, Qwen3) then scores this script from 1-10 across four dimensions: (1) **Narrative Coherence** (logical plot), (2) **Causal Complexity** (rich cause-effect chains), (3) **Visual Information Significance** (necessity of visual cues), and (4) **Entity & Relationship Dynamics** (meaningful character evolution). More details in the supplementary material.

Quantitative Curation To obtain one robust signal from the multi-model, multi-dimensional scores, we utilize a three-stage quantitative procedure.

1. Score Normalization. To balance the intrinsic scoring biases of various LLMs, we use Z-score normalization. For every model i and every dimension d, a raw score $x_{d,i}$ is normalized to a Z-score $z_{d,i}$ based on that model's own mean

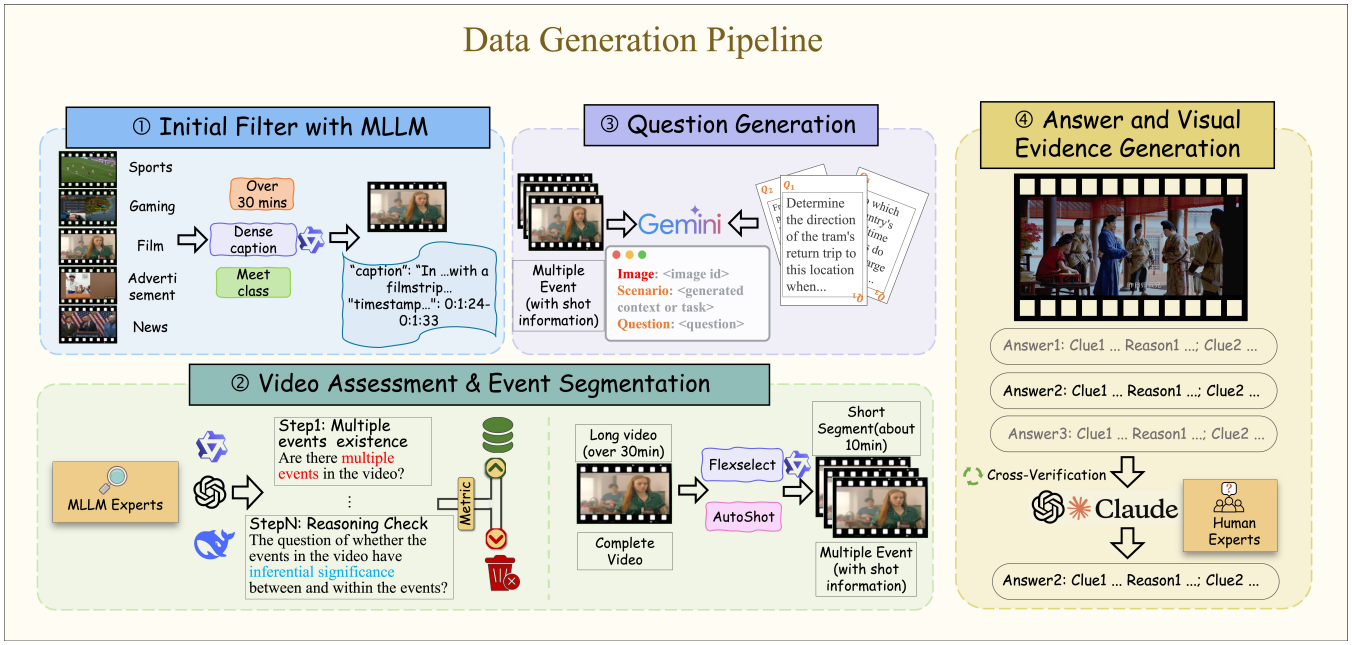

Figure 1: An overview of our multi-stage data curation and annotation pipeline. The process begins with video curation, followed by question generation, and concludes with a two-pass answer annotation, cross-model and expert verification workflow.

$mu_{d,i}$ and standard deviation $sigma_{d,i}$ for that dimension.

$$z_{d,i} = \frac{x_{d,i} - \mu_{d,i}}{\sigma_{d,i}} \tag{1}$$

2. Weighted Score Aggregation. The normalized scores are subsequently combined to calculate the ultimate **Analyzability Score** (S_{final}). It is a weighted sum in which we can place more confidence in more competent models (weight w_i) and favor dimensions most important to our benchmark, i.e., Causal Complexity ($w_{d=\text{causal}} = 0.4$).

$$S_{\text{final}} = \sum_{d \in D} w_d \cdot S_{\text{norm},d} \quad \text{where} \quad S_{\text{norm},d} = \sum_{i=1}^{N} w_i \cdot z_{d,i} \tag{2}$$

3. Disagreement Metric. To manage controversial cases, we measure how much models disagree with each other by computing a **Divergence Score** (D_{video}). It is the arithmetic average of the standard deviations (D_d) of the Z-scores in all M dimensions.

$$D_{\text{video}} = \frac{1}{M} \sum_{d=1}^{M} D_d \quad \text{where} \quad D_d = \sqrt{\frac{\sum_{i=1}^{N} (z_{d,i} - \mu_d)^2}{N}} \tag{3}$$

A large D_{video} value implies high disagreement between the models on the quality of a video. Final filtering is performed by an automated triaging step using the calculated $S_{textfinal}$ and $D_{textvideo}$ metrics, enabling scalable and objective curation. Videos with high analyzability score and strong model agreement ($S_{textfinal} > 1.0$ and $D_{textvideo} < 0.5$) are automatically accepted, and videos with low scores and high agreement ($S_{textfinal} < -1.0$ and $D_{textvideo} < 0.5$) are rejected. Importantly, videos with high model disagreement ($D_{textvideo} > 1.2$) are prioritized for expert review, effectively targeting human resources at the most uncertain cases. Remaining videos are relegated to a secondary queue for regular check.

Semantic Event Segmentation and Annotation We divide videos into narratively meaningful Event Segments. Our pipeline employs AutoShot to first map all visual cuts, and then a VLM detects narrative boundaries by predicting exact start-end shot IDs. These proposals are consolidated into non-overlapping segments, each of which includes a title, timestamps, and a local shot list. Our annotation pipeline is a three-stage workflow. First, Gemini 2.5 Pro produces questions aiming at certain reasoning abilities for every segment. Second, based on the entire video as context, it gives a grounded answer, spatio-temporal cues are extracted, and reasonable distractors for a multiple-choice question are created. Last, cross-validation of both evidence and the ultimate answer is carried out by GPT-4o and Claude-3.5-Sonnet to provide data quality assurance and reduce hallucinations.

Evaluation Metrics

Answer Accuracy (Acc_{ans})

We evaluate for answer correctness in a multiple-choice question (MCQ) format. The model picks one among five choices, and the grading is binary, giving an objective metric of its capacity to arrive at the correct conclusion. The score is defined as:

$$Acc_{\text{MCQ}} = \begin{cases} 1, & \text{if model's choice is correct} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Spatio-Temporal Evidence Quality (S_{clue})

The core of the evaluation measure that we wish to define is described here. We compare the model-generated list of

clues, C_{pred}, with the ground-truth list defined as C_{gt}. The measure compares the model's ability to ground its reasoning in concrete, verifiable events depicted in the video content. The resulting measure S_{clue} is defined as an F1-score that balances precision and recall by computing using both temporal and semantic correspondences.

Matching Score. We first introduce a matching score, S_{match}, between a predicted clue c_p and a ground-truth clue c_g. This score is the product of their temporal overlap and semantic similarity:

$$S_{\text{match}}(c_p, c_g) = \text{IoU}(T_p, T_g) \times \text{Sim}(D_p, D_g) \quad (5)$$

where IoU is the Temporal Intersection over Union of the clue timestamps, and Sim is the cosine similarity of the sentence embeddings of their textual descriptions.

Clue-based Precision, Recall, and F1-Score. With the matching scores, we use an optimal matching algorithm to determine the best one-to-one match between $C_{textpred}$ and C_{textgt}. According to the sum score of such optimal pairs, $sumS_{textmatch}$, we define Clue Precision ($P_{textclue}$) and Clue Recall ($R_{textclue}$):

$$P_{\text{clue}} = \frac{\sum S_{\text{match}}}{|C_{\text{pred}}|} \quad , \quad R_{\text{clue}} = \frac{\sum S_{\text{match}}}{|C_{\text{gt}}|} \quad (6)$$

The overall evidence quality score, $S_{textclue}$, is the harmonic mean (F1-score) of both of these, yielding a single value(0-1) for the overall quality of the given evidence set.

$$S_{\text{clue}} = 2 \times \frac{P_{\text{clue}} \times R_{\text{clue}}}{P_{\text{clue}} + R_{\text{clue}}} \quad (7)$$

Overall Performance Score (S_{overall})

The overall score defined as S_{overall}, is a weighted aggregation of response accuracy with the quality of evidence, based on the argument that genuine reasoning demands not only accurate conclusions but also a well-backed rationale. In the entire assessment, quality of evidence is given 30% and accuracy of response is given 70%, thus encouraging models not to guess randomly but to support their response. The score is calculated as follows:

$$S_{\text{overall}} = \lambda_{\text{ans}} \cdot Acc_{\text{MCQ}} + \lambda_{\text{clue}} \cdot S_{\text{clue}} \quad (8)$$

where we set the weights $\lambda_{\text{ans}} = 0.7$ and $\lambda_{\text{clue}} = 0.3$.

Methodology: DynamicSelect

To address the limitations of static token pruning strategies, which often discard vital long-range context and problem related visual evidence required by the complex tasks in our EV^2-Bench, we propose **DynamicSelect**. Illustrated in Figure 2, we introduce a novel, multi-stage approach for adaptive token compression that intelligently modulates its strategy based on the input video and textual query.

`DynamicSelect` is a lightweight Token Selector that efficiently identifies the most query-relevant visual tokens. Although ranking tokens with a large VLM is accurate, it is computationally prohibitive. We address this via a teacher-student paradigm, distilling a large VLM's ranking ability into a small, efficient student model.

Dynamic Budget Prediction via Policy Gradient

Traditional approaches tend to use a static pre-specified token budget (*e.g.*, a preselected number of tokens, k, to keep). To bypass the limitations, we propose a lightweight *budget predictor*. The module employs a reinforcement learning framework to dynamically predict an optimal budget for every distinct compression task by initially creating a state vector, s, that summarizes its essential features.

Query Semantic Vector (v_{query}), representing the intent of the user. To circumvent computationally costly preprocessing, we compute this vector by passing the text query through our student model to get the last hidden states, $H_{\text{text}} \in \mathbb{R}^{L_{\text{text}} \times D}$, and then apply mean pooling to get one fixed-dimension vector, $v_{\text{query}} \in \mathbb{R}^D$.

Video Metadata Vector (v_{meta}): To make the decision sensitive to the actual physical characteristics of the video, particularly its temporal extent, we include the following important metadata features:

1. d_{orig}: The video's original duration in seconds.
2. f_{orig}: The video's original total frame count.
3. f_{sampled}: The number of frames actually sampled.

The ultimate state representation s is the concatenation of the following features: $s = \text{concat}(v_{\text{query}}, d_{\text{orig}}, f_{\text{orig}}, f_{\text{sampled}})$.

Action Space and Policy Network: To improve training stability, we model the budget choice as a discrete decision process.

1. Action Space (A): Our budget predictor, selects an action a_i from a discrete set of predefined budget values, $A = \{k_1, k_2, \ldots, k_n\}$. For example, actions in $A = \{2048, \ldots, 8192\}$ can correspond to high, medium, and low compression strategies, respectively.

2. Policy Network (π_θ): The 'budget predictor' is a small MLP with weights θ. It takes in a state s and outputs a probability distribution over the action space, $P(a|s;\theta)$, through a `Softmax` activation over its last layer.

Training with Policy Gradient: There is a challenge in that the selection of the top-k tokens is a non-differentiable operation and thus prevents gradient flow from the ultimate task loss to the 'budget predictor'. We get around this by using the Policy Gradient algorithm:

1. Action Sampling: During each training forward pass, an action $k \in A$ is stochastically sampled from the policy distribution $\pi_\theta(a|s)$. This sampled budget, k, then determines the number of tokens for the subsequent selection.

2. Reward Function (R): Upon taking the action k and receiving the importance scores from both the student and teacher models, a reward R is calculated to assess the quality of the selected action. The reward function is designed to balance **ranking accuracy** and **computational efficiency**:

$$R = \rho_{\text{Spearman}}(S_{\text{student}}, S_{\text{teacher}}) - \lambda \cdot \left(\frac{k}{k_{\text{max}}}\right) \quad (9)$$

where ρ_{Spearman} is the Spearman's rank correlation coefficient between the student's and teacher's score vectors (S_{student} and S_{teacher}). This term rewards the student for correctly imitating the teacher's importance ranking. The second term is a penalty proportional to the used budget k,

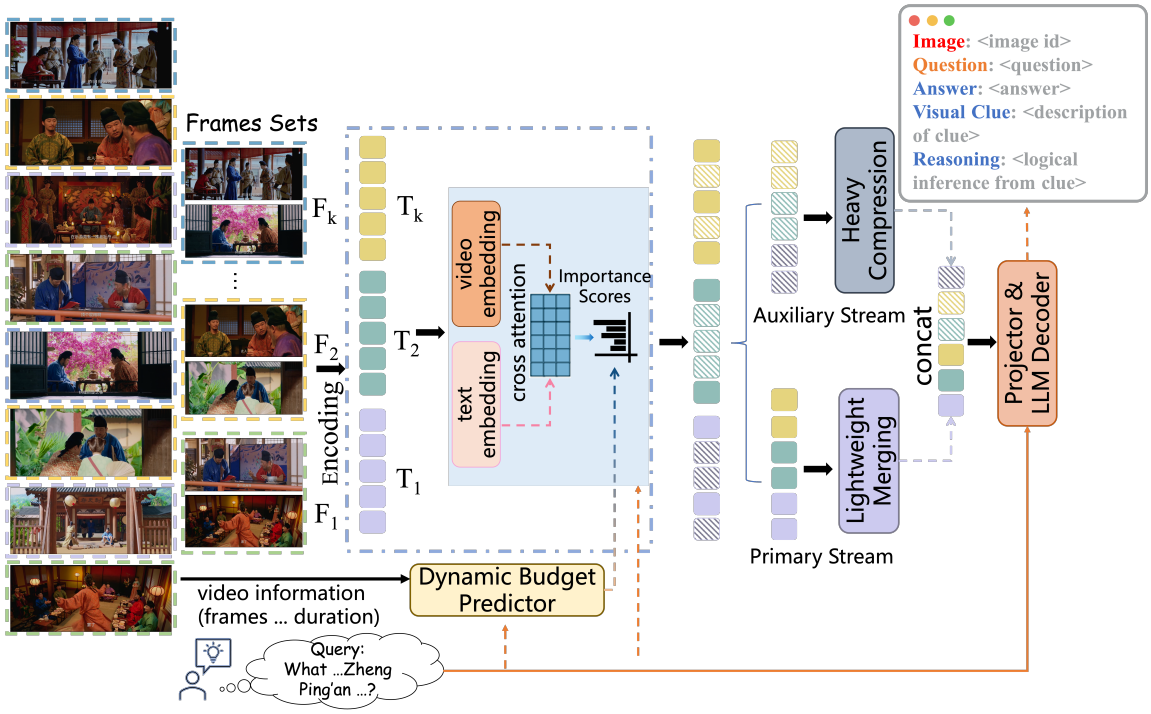

Figure 2: The overall architecture of our DynamicSelect framework. It consists of three stages: Dynamic Budget Prediction for semantically token selection, Hierarchical Token Processing, and Final Sequence Reconstitution.

where $k_{\max}$ is the maximum available budget and λ is a hyperparameter controlling the trade-off. **Loss Function:** The objective for the 'budget predictor' is to maximize the expected reward. Following the policy gradient theorem, the loss function is defined as:

$$\mathcal{L}(\theta) = -\log(\pi_\theta(k|s)) \cdot R_{\text{detached}} \tag{10}$$

Crucially, the reward R is also a fixed scalar weight for the gradient computation, so gradients are not allowed to flow through it. The gradient update is hence solely a function of the log-probability of the action selected. A positive reward R reinforces the choice of action k in state s by increasing its probability $\pi_\theta(k|s)$, and a negative reward will penalize it. The absolute value of the reward $|R|$ naturally scales the gradient update step, thereby facilitating successful training of the 'budget predictor'.

Hierarchical Token Processing

After deciding to apply a dynamic budget represented by k, a formal processing procedure begins to generate scores of significance by using a differential compression approach.

Global Importance Scoring: This step gives a very accurate significance score to every visual token. The student model, known as the 'Token Selector', processes all data input to generate the final hidden states given by $H \in \mathbb{R}^{L \times D}$. The resulting hidden states are then divided into visual (H_{vision}) and textual (H_{text}) parts. A simple cross-attention, working between textual queries and visual keys, is used to compute an importance score s_i per visual token, finally giving rise to a final score vector given by $S \in \mathbb{R}^{L_{\text{vision}}}$.

Hierarchical Compression Strategy: Based on the scores S and budget k, visual tokens are triaged into two streams for differential processing, both utilizing a bipartite merging algorithm detailed below.

Bipartite Token Merging: Inspired by recent work (Bolya et al. 2022; Zhang et al. 2024d) in token merging, our compression algorithm aims to efficiently reduce token count while preserving information by merging similar tokens. Given a set of tokens $T = \{t_1, t_2, \ldots, t_N\}$, $t_i \in \mathbb{R}^D$, the process is as follows:
1. Splitting: The set T is bipartitioned into two disjoint subsets, a set to be merged, A, and a set of merge candidates, B. A simple strategy is to assign $A = \{t_1, \ldots, t_{\lfloor N/2 \rfloor}\}$ and $B = \{t_{\lfloor N/2 \rfloor + 1}, \ldots, t_N\}$.
2. Similarity Matching: We compute a pairwise similarity matrix $K \in \mathbb{R}^{|A| \times |B|}$ between the two sets, where $K_{ij} = \text{sim}(t_i, t_j)$ for $t_i \in A, t_j \in B$, using cosine similarity.
3. Merging: For each token $t_i \in A$, we find its most similar token $t_{j^*} \in B$ where $j^* = \arg\max_j K_{ij}$. We then merge this pair into a new token t'_i. The unmerged tokens from B are preserved and combined with the newly merged tokens to form the final compressed set.

We employ a differential compression strategy. The top-k primary tokens are lightly compressed (5%) to reduce redundancy while preserving high-fidelity details. Concurrently, the remaining auxiliary tokens undergo high-ratio compression (85%) to distill background context. Both sets are then concatenated into the final input sequence, balancing efficiency with informational integrity. Finally, the lightly-merged primary tokens and the condensed auxiliary

Model	Action	Object	Scene	Spatial	Temporal	Event	Overall
GPT-4o	62.1	55.3	61.2	67.0	65.0	71.0	63.6
Gemini 2.5-Pro	76.5	72.1	74.8	78.0	72.0	79.0	75.4
InternVL2.5-38B	50.3	38.2	49.0	**61.0**	40.1	**66.0**	**54.0**
MiniCPM-O	**56.4**	**42.5**	**55.0**	51.0	38.1	44.5	47.3
Qwen2.5-VL-34B	42.1	25.4	51.0	57.0	36.0	54.8	46.7
OWL3	50.3	38.2	49.0	59.0	36.0	44.5	45.9
Qwen2.5-VL-7B	36.0	38.2	**55.0**	59.0	**42.1**	42.1	44.8
InternVL2.5-8B	42.1	31.8	49.0	**61.0**	29.9	46.9	44.3
MiniCPM-V	36.0	40.3	53.0	59.0	38.1	36.6	42.4
TARS-7B	46.2	40.3	37.0	57.0	27.7	41.4	42.0
LLaVA-34B	31.9	33.9	37.0	49.0	29.9	33.4	35.4
LLaVA-7B	36.0	25.4	33.0	37.0	27.9	26.8	28.5
DynamicSelect-7B	43.1	42.4	58.2	48.7	62.1	52.3	51.1

Table 1: Overall performance comparison across all task categories. Best scores in each column are in **bold**.

Token Selector	Dynamic Predictor	Merging	Compression	VideoMME
				65.4
✓				67.4
✓	✓			67.7
✓	✓	✓(5%)		67.9
✓	✓	✓(10%)		67.6
✓	✓	✓(15%)		67.2
✓	✓	✓(5%)	✓(75%)	68.3
✓	✓	✓(5%)	✓(85%)	68.3
✓	✓	✓(5%)	✓(95%)	68.1

Table 2: Ablation study on the effectiveness of key components, focusing on the impact of merging and compression with different compression ratios on model performance.

tokens are concatenated to form the final, optimized input sequence. This composite sequence is then fed to the main LLM for reasoning, ensuring both computational efficiency and a comprehensive understanding of the video content.

Experiments

We evaluated 10 mainstream open-source MLLMs and 2 close-source APIs on EV^2-Bench. Additionally, our model are evaluated on 4 established long-video understanding benchmarks. More details in the supplementary material.

Main Results

Open-source Models Evaluation on EV^2-Bench We report the performance of leading open-source models and our proposed method on the core tasks of EV^2-Bench in Tab. **??**. A comprehensive breakdown is available in Tab. 4 of Appendix. Our primary finding is a stark performance disparity across reasoning dimensions. While most models achieve reasonable accuracy on multiple-choice **Event** comprehension, their performance degrades significantly on tasks requiring precise grounding in spatio-temporal evidence. This is particularly evident in the **Temporal** localization task, where scores for even the best open-source models struggle to surpass 60. This highlights a critical deficiency in the ability of current VLMs to understand the "when" of an event. We hypothesize this weakness stems from their extensive pre-training on static images, indicating that enhancing temporal modeling is a crucial avenue for future work. This challenge is exacerbated by practical constraints; for instance, we limited Qwen2.5-VL to 256 frames(the maximum capacity for a 96GB H20 GPU) which inherently reduces temporal resolution and impairs performance.

To address this, our **DynamicSelect** is designed to process a much denser input of 1024 frames by efficiently identifying and utilizing the most salient information while discarding redundant tokens. As shown in Tab. **??**, this approach not only improves efficiency but also boosts performance, surpassing the strong Qwen2.5-VL baseline across key metrics. Furthermore, we observe an intriguing trend: a model's performance on high-level **Event** reasoning correlates more strongly with its **Spatial** understanding than its **Temporal** accuracy. This suggests that models often compensate for poor temporal grounding by relying heavily on visual scene comprehension to infer narrative progression. By explicitly decoupling and evaluating these capabilities, EV^2-Bench reveals that models with weaker visual faculties (e.g., LLaVA-Video) tend to guess rather than reason from evidence. Therefore, it can be seen that decoupling the deep understanding of events by models from the dimensions of time and visual space is very important for evaluation.

DynamicSelect Evaluation on Other Benchmarks. In Tab. 3, we report the performance of DynamicSelect across multiple public long-video understanding benchmarks. Following FlexSelect, our DynamicSelect is also validated for generalization on two open-source multimodal models, InternVL2.5 and Qwen2.5-VL. For InternVL2.5, to ensure the efficiency of token selection, we adopt InternVL2.5-0.5B as the token selector. Compared to InternVL2.5, our method achieves consistent improvements across all long-video benchmarks, with an average gain of 4.8 points. Relative to our baseline, it yields an average improvement of 1.25 points, narrowing nearly half of the performance gap between the 7B and 78B models. This further compresses and delivers more critical information to the model. For Qwen2.5-VL, a foundation model pre-trained on extensive video data, our method brings an average improvement of

Model	Size	VideoMME		MLVU	LongVB	LVBench
		Long	Overall	M-Avg	Val	Test
Proprietary Models						
GPT-4o (OpenAI 2024)	-	65.3	71.9	64.6	66.7	64.4
Gemini-1.5-Pro (Team et al. 2024)	-	67.4	75.0	-	64.0	65.7
Gemini-2.5-Pro (Team et al. 2024)	-	-	87.0	81.2	-	69.2
Seed1.5-VL-8B	-	-	77.9	82.1	74.4	64.6
Eagle2.5-8B	-	-	72.4	77.6	66.4	-
Open-Source VideoLLMs						
Qwen2-VL (Wang et al. 2024a)	7B	53.8	63.3	66.9	55.6	42.4
NVILA (Liu et al. 2025c)	8B	54.8	64.2	70.1	57.7	-
VideoLLaMA3 (Zhang et al. 2025a)	7B	-	66.2	73.0	59.8	45.3
VideoChat-Flash (Li et al. 2025b)	7B	55.6	65.3	74.7	64.7	48.2
Aria (Li et al. 2025a)	3.5B	58.8	67.6	70.6	65.3	-
Oryx-1.5 (Liu et al. 2025b)	34B	59.3	67.3	72.3	62.0	30.8
Video-XL-Pro (Liu et al. 2025a)	3B	-	60.0	70.6	56.7	-
Video-XL2 (Liu et al. 2025a)	8B	-	66.6	74.8	61.0	48.4
ViLAMP (Cheng et al. 2025)	7B	57.8	67.5	72.6	61.2	45.2
LLaVA-Video (Zhang et al. 2024c)	7B	52.9	64.4	68.6	58.2	43.1
LLaVA-Video (Zhang et al. 2024c)	72B	61.9	70.0	71.2	62.4	45.5
InternVL2.5 (Chen et al. 2025)	78B	-	72.1	75.7	63.6	46.5
InternVL2.5 (Chen et al. 2025)	8B	52.8	64.2	68.9	59.5	43.4
Qwen2.5-VL (Bai et al. 2025)	7B	55.6	65.4	70.2	59.5	45.3
FlexSelect(base-InternVL2.5)	8B	57.9	67.2	71.9	61.2	49.9
FlexSelect(base-Qwen2.5-VL)	7B	58.6	67.4	70.3	61.9	50.0
DynamicSelect(base-InternVL2.5)	8B	59.8 ↑1.9	68.3 ↑1.1	72.3 ↑0.4	63.1 ↑1.9	51.5 ↑1.6
DynamicSelect(base-Qwen2.5-VL)	7B	60.2 ↑1.6	68.7 ↑1.3	70.9 ↑0.6	63.7 ↑1.8	51.9 ↑1.9

Table 3: Comprehensive Evaluation Across Diverse Long-Video Benchmarks. DynamicSelect leverages scores to dynamically select semantic tokens from our lightweight token selector, while implementing dual compression on both selected background tokens and unselected high-confidence tokens.

3.7 points over the base model. Compared to FlexSelect, it achieves an average gain of 1.4 points. Evaluations across public benchmarks show that DynamicSelect can extract more critical information from long videos, enhancing both the effectiveness of token selection and token compression.

Ablation Study

Effectiveness of key components. Tab. 2 presents the core components of DynamicSelect and the results of ablation experiments under different compression ratios. On VideoMME, the dynamic token prediction condenses key information while reducing the number of selected tokens, and even contributes to the model performance. When 5% redundancy compression is applied to the selected tokens, it not only further reduces the number of tokens that the model needs to infer but also eliminates overly redundant information to a greater extent. This information simplification actually enhances the model's ability to capture key information in long videos. We conducted experiments under various compression ratios. Since the token selector has already extracted high-quality tokens, at a compression ratio of 5%, it can further reduce the number of tokens and remove some redundant information. Although text queries can be used to select rich semantic tokens, not all crucial tokens can be selected due to the initially set number of tokens, memory limitations, and inference efficiency constraints. Therefore, there are still some high-confidence tokens discarded by the token selector that are worth utilizing. After compressing these tokens at a high ratio, we found that at a compression ratio of 85%, it can not only supplement additional information but also ensure the introduction of fewer inference tokens, thereby improving the model performance without affecting inference efficiency.

Conclusion

During our research, we strived to overcome two major challenges related to understanding long videos: the prohibitively high computational cost required to process long subsegments of videos and the ubiquity of unsupported reasoning across methods. To achieve our goals, we provided a dual contribution. In particular, our new benchmark, EV^2-Bench, provides a strict evaluation criterion by requiring models to provide evidence to justify their output with accurate spatio-temporal visual information. Additionally, our DynamicSelect introduces a novel solution that effectively reduces video clips by selectively retaining semantically informative tokens, hence trying to eradicate redundancies. Our overall experimental evidence demonstrates that DynamicSelect, besides achieving a SOTA accuracy on EV^2-Bench, provides a better tradeoff between accuracy and computation cost. We believe this work will encourage the development of more faithful and efficient VLMs, capable of truly comprehending the intricate dynamics of long video.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62376270).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; Gu, L.; Wang, X.; Li, Q.; Ren, Y.; Chen, Z.; Luo, J.; Wang, J.; Jiang, T.; Wang, B.; He, C.; Shi, B.; Zhang, X.; Lv, H.; Wang, Y.; Shao, W.; Chu, P.; Tu, Z.; He, T.; Wu, Z.; Deng, H.; Ge, J.; Chen, K.; Zhang, K.; Wang, L.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2025. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv:2412.05271.

Cheng, C.; Guan, J.; Wu, W.; and Yan, R. 2025. Scaling Video-Language Models to 10K Frames via Hierarchical Differential Distillation. arXiv:2504.02438.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Fu, T.; Liu, T.; Han, Q.; Dai, G.; Yan, S.; Yang, H.; Ning, X.; and Wang, Y. 2024. Framefusion: Combining similarity and importance for video token reduction on large visual language models. *arXiv preprint arXiv:2501.01986*.

Jin, P.; Takanobu, R.; Zhang, W.; Cao, X.; and Yuan, L. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13700–13710.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Li, D.; Liu, Y.; Wu, H.; Wang, Y.; Shen, Z.; Qu, B.; Niu, X.; Zhou, F.; Huang, C.; Li, Y.; Zhu, C.; Ren, X.; Li, C.; Ye, Y.; Liu, P.; Zhang, L.; Yan, H.; Wang, G.; Chen, B.; and Li, J. 2025a. Aria: An Open Multimodal Native Mixture-of-Experts Model. arXiv:2410.05993.

Li, X.; Wang, Y.; Yu, J.; Zeng, X.; Zhu, Y.; Huang, H.; Gao, J.; Li, K.; He, Y.; Wang, C.; Qiao, Y.; Wang, Y.; and Wang, L. 2025b. VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling. arXiv:2501.00574.

Liu, X.; Shu, Y.; Liu, Z.; Li, A.; Tian, Y.; and Zhao, B. 2025a. Video-XL-Pro: Reconstructive Token Compression for Extremely Long Video Understanding. arXiv:2503.18478.

Liu, Z.; Dong, Y.; Liu, Z.; Hu, W.; Lu, J.; and Rao, Y. 2025b. Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution. arXiv:2409.12961.

Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; Li, X.; Fang, Y.; Chen, Y.; Hsieh, C.-Y.; Huang, D.-A.; Cheng, A.-C.; Nath, V.; Hu, J.; Liu, S.; Krishna, R.; Xu, D.; Wang, X.; Molchanov, P.; Kautz, J.; Yin, H.; Han, S.; and Lu, Y. 2025c. NVILA: Efficient Frontier Visual Language Models. arXiv:2412.04468.

OpenAI. 2024. Hello GPT-4o. Accessed: 2024-05-20.

Shen, X.; Xiong, Y.; Zhao, C.; Wu, L.; Chen, J.; Zhu, C.; Liu, Z.; Xiao, F.; Varadarajan, B.; Bordes, F.; Liu, Z.; Xu, H.; Kim, H. J.; Soran, B.; Krishnamoorthi, R.; Elhoseiny, M.; and Chandra, V. 2024. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. arXiv:2410.17434.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; and et al. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Gu, X.; Huang, S.; Xu, B.; Dong, Y.; et al. 2024b. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857.

Xing, L.; Huang, Q.; Dong, X.; Lu, J.; Zhang, P.; Zang, Y.; Cao, Y.; He, C.; Wang, J.; Wu, F.; et al. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9127–9134.

Zeng, Z.; Chen, P.; Liu, S.; Jiang, H.; and Jia, J. 2023. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *arXiv preprint arXiv:2312.17080*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; Jin, P.; Zhang, W.; Wang, F.; Bing, L.; and Zhao, D. 2025a. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. arXiv:2501.13106.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024a. Long Context Transfer from Language to Vision. arXiv:2406.16852.

Zhang, Q.; Cheng, A.; Lu, M.; Zhuo, Z.; Wang, M.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2024b. [CLS] Attention is All You Need for Training-Free Visual Token Pruning: Make VLM Inference Faster. *arXiv e-prints*, arXiv–2412.

Zhang, Y.; Liu, X.; Tao, R.; Chen, Q.; Fei, H.; Che, W.; and Qin, L. 2025b. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 5267–5276.

Zhang, Y.; Liu, X.; Zhou, R.; Chen, Q.; Fei, H.; Lu, W.; and Qin, L. 2025c. CCHall: A Novel Benchmark for Joint Cross-Lingual and Cross-Modal Hallucinations Detection in Large Language Models. *arXiv preprint arXiv:2505.19108*.

Zhang, Y.; Lu, Y.; Wang, T.; Rao, F.; Yang, Y.; and Zhu, L. 2025d. FlexSelect: Flexible Token Selection for Efficient Long Video Understanding. *arXiv preprint arXiv:2506.00993*.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024c. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713.

Zhang, Y.; Zhao, Z.; Chen, Z.; Ding, Z.; Yang, X.; and Sun, Y. 2024d. Beyond training: Dynamic token merging for zero-shot video understanding. *arXiv preprint arXiv:2411.14401*.