BEYOND FRAME SELECTION: DECOUPLED ADAPTIVE EVIDENCE SELECTION FOR VIDEO UNDERSTANDING WITH MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

In video understanding, MLLMs often need to reason over visual content whose relevance to a question varies over time. Existing frame selection and visual-token pruning methods typically rely on a unified relevance ranking or fixed selection granularity, overlooking that visual-text and ordinary visual content can exhibit different relevance distributions and contribute differently to different questions. We introduce DAES, a training-free framework for *Decoupled Adaptive Evidence Selection*. DAES uses reflection-based cues to disentangle visual-text and ordinary visual evidence at the patch level, and estimates their query-conditioned relevance through Value-space similarity between query tokens and visual patches in an MLLM. To make the two evidence types directly comparable, we calibrate their relevance against type-specific null distributions. Under a fixed token budget, DAES adaptively selects evidence types and timestamps according to calibrated relevance scores. This evidence-level selection enables broader temporal coverage under a fixed token budget while preserving question-relevant information. For efficient selection, we use a lightweight MLLM as the evidence selector while reserving a larger MLLM for final reasoning. Experiments on subtitle-rendered Video-MME-v2 and LongVideoBench demonstrate consistent improvements in video understanding performance under the same token budget.

1 INTRODUCTION

Multimodal large language models (MLLMs) have shown strong capabilities in understanding both images and videos. Unlike static images, videos distribute useful evidence unevenly over time, while processing more frames rapidly increases the number of visual tokens and computational cost. This creates a fundamental allocation problem: under a limited token budget, the model must decide which moments to observe and which visual evidence to preserve. Existing approaches mainly follow two paradigms: *frame selection*, which identifies query-relevant timestamps, and *visual-token reduction*, which removes redundant spatial or temporal tokens.

However, existing approaches typically assume that all the visual evidence within a video can be measured using a unified relevance criterion. Frame selection usually assigns a single score to each timestamp, while token reduction directly ranks all the visual tokens according to their relevance to the query. Such formulations overlook that different types of visual evidence may contribute to question answering in different ways. This issue is particularly relevant in real-world videos, where textual information such as subtitles, captions, signs, and screen content frequently appears alongside ordinary visual content. Therefore, we distinguish two types of evidence within videos: **Visual Evidence**, which refers to ordinary visual content such as objects, scenes, and actions, and **Textual Evidence**, which corresponds to text appearing in videos. As shown in Fig. 1(a,b), different questions rely on different mixtures of the two evidence types, which also exhibit distinct query-relevance distributions. Treating them with a single relevance score can therefore obscure their different contributions to question answering.

Motivated by these observations, we introduce **Decoupled Adaptive Evidence Selection (DAES)**, a training-free framework that moves beyond frame-level selection to *evidence-level selection*. As illustrated in Fig. 1(c), DAES separately models Visual and Textual Evidence and adaptively allo-

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

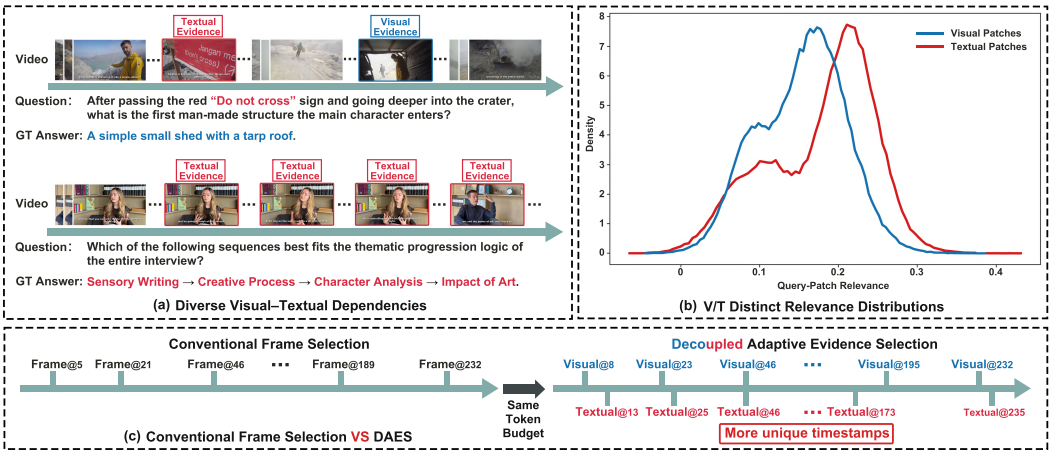

Figure 1: Motivation of DAES. (a) Questions require different mixtures of visual and textual evidence. (b) The two evidence types exhibit distinct query-relevance distributions. (c) DAES selects evidence across timestamps under the same visual-token budget, enabling broader temporal coverage than frame selection.

cates a fixed visual-token budget across evidence types and timestamps. By selecting only question-relevant evidence rather than complete frames, DAES can cover more temporal locations under the same token budget.

To operationalize evidence-level selection, DAES decomposes each frame into Visual and Textual Evidence using reflection-based cues. Since different candidate options may correspond to different relevant evidence, DAES estimates option-conditioned patch relevance in the MLLM Value space. However, the two evidence types exhibit different raw relevance distributions, making their scores not directly comparable. We therefore apply evidence-specific null calibration to map them into a common score space, after which relevant evidence is selected across timestamps and candidate options under a fixed token budget. To make dense evidence evaluation efficient, a lightweight MLLM performs evidence selection while a larger MLLM is reserved for final reasoning.

To evaluate mixed visual–textual evidence, we construct subtitle-rendered versions of Video-MME-v2 and LongVideoBench using temporally aligned subtitles. Under comparable token budgets, DAES substantially outperforms frame-selection baselines, improving over uniform sampling by 8.72 and 8.00 percentage points, respectively. Our main contributions are summarized as follows:

- We identify an overlooked issue in video understanding: Visual Evidence and Textual Evidence exhibit distinct query-relevance characteristics, while existing frame selection and visual-token reduction methods typically model them using a unified relevance criterion.
- We propose **Decoupled Adaptive Evidence Selection (DAES)**, a training-free framework that decouples Visual Evidence and Textual Evidence, calibrates their relevance scores, and adaptively selects evidence under a fixed token budget. By moving beyond frame-level selection to evidence-level selection, DAES enables broader temporal coverage while preserving question-relevant information.
- We evaluate DAES on constructed subtitle-rendered versions of Video-MME-v2 and LongVideoBench, where Visual Evidence and Textual Evidence coexist in realistic video scenarios. Experiments demonstrate that DAES achieves superior performance under the same token budget.

2 Related Work

2.1 Query-Aware Frame Selection

Query-aware frame selection aims to reduce video context by selecting timestamps that are more relevant to the given question. Q-Frame (Zhang et al., 2025) explores question-guided frame selec-

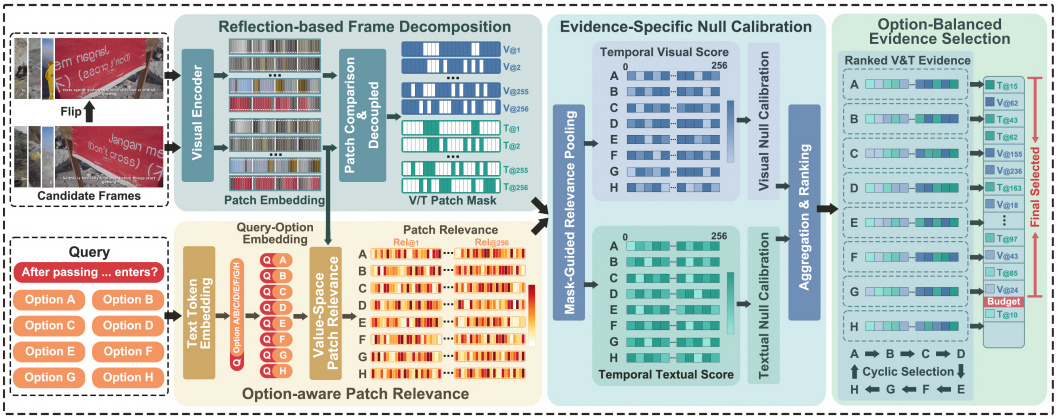

Figure 2: Overview of DAES. Given candidate video frames, a question, and answer options, DAES first applies reflection-based frame decomposition to create patch-level Visual and Textual Evidence masks. For each answer option, option-aware patch relevance estimates query–patch similarity in the MLLM Value space and pools scores within each mask. Evidence-specific null calibration maps the two evidence types to comparable utility scores. Finally, option-balanced evidence selection cycles through answer options while preserving their ranking orders, selecting high-utility evidence until the token budget is reached. The selected patches are arranged by timestamp and passed to the answer MLLM.

tion through vision-language similarity. AKS (Tang et al., 2025) dynamically selects informative frames based on query-conditioned relevance. FOCUS (Zhu et al., 2025) further considers fine-grained query-frame relevance for identifying informative moments. VTR-VLM (Wu et al., 2026) incorporates both questions and answer options to estimate video relevance in multiple-choice video understanding. WFS-SB (Chen et al., 2026) investigates question-aware frame sampling by considering temporal and semantic information.

Despite these advances, existing frame selection methods mainly treat the frame as the fundamental selection unit. The selected timestamps are usually forwarded as complete visual inputs, even when only specific evidence within a frame is relevant to the question. DAES goes beyond frame-level selection by decoupling different evidence types and allowing them to independently select their relevant temporal locations.

2.2 Visual Token Selection and Compression

Recent studies have explored reducing the visual token cost of multimodal models through token pruning, merging, and compression. Token merging methods such as ToMe (Bolya et al., 2022) reduce redundant visual representations by combining similar tokens, while FastV (Chen et al., 2024), SparseVLM (Zhang et al., 2024), and VisionZip (Yang et al., 2024) investigate importance-based token selection for efficient vision-language inference. For video understanding, PruneVid (Huang et al., 2024) further considers temporal and spatial redundancy and performs query-aware visual token pruning.

These methods mainly focus on reducing redundant tokens after visual content has been provided to the model. In contrast, DAES focuses on selecting question-relevant evidence before token allocation. By decoupling Visual Evidence and Textual Evidence, DAES determines not only which tokens are useful, but also which evidence type should be selected at different temporal locations.

3 Method

3.1 Overview

Given a video, a multiple-choice question q, and its candidate options $\mathcal{O} = \{o_i\}_{i=1}^{O}$, DAES performs decoupled evidence-level visual context selection by separating two types of evidence: **Vi-**

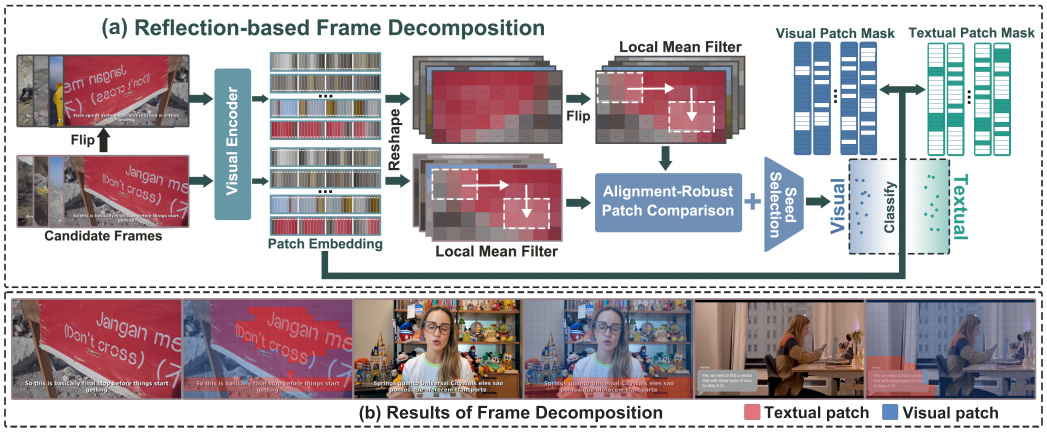

Figure 3: (a) Reflection-based decomposition of video frames into Visual and Textual Evidence. (b) Representative examples of frame decomposition.

sual Evidence and **Textual Evidence**. Under a fixed visual-token budget, DAES adaptively selects question-relevant evidence across timestamps and forwards the corresponding visual tokens to the answer MLLM.

As illustrated in Fig. 2, DAES consists of four stages. First, **Reflection-based Frame Decomposition** separates Visual Evidence and Textual Evidence at the patch level without requiring OCR or external transcripts. Second, **Option-aware Patch Relevance** estimates the relevance between answer hypotheses and visual patches by measuring their interaction in the Value space of an MLLM. Third, **Evidence-Specific Null Calibration** maps the relevance scores of the two evidence types into a common relevance space, enabling direct comparison between Visual Evidence and Textual Evidence. Finally, **Option-Balanced Evidence Retrieval** ranks evidence candidates and cyclically selects evidence across answer options under a fixed token budget.

3.2 Reflection-based Frame Decomposition

Inspired by the reflection sensitivity of vision-language representations observed in MirrorCLIP (Wang et al., 2024), where textual regions exhibit stronger reflection sensitivity while ordinary visual regions remain relatively stable, we extend this property from global vision-language representations to patch-level visual representations in MLLMs for evidence decomposition. As illustrated in Fig. 3, we design a reflection-based patch decomposition mechanism to separate Visual and Textual Evidence at the patch level without relying on external OCR or transcript information.

3.2.1 Alignment-robust reflection response.

For each frame I_t, we encode both the original image and its horizontal reflection $I_t^{\leftrightarrow}$ using the same visual encoder. Different from the cosine similarity used in MirrorCLIP (Wang et al., 2024), which measures reflection sensitivity at the representation level, we use the ℓ_2 feature discrepancy between reflected and original patches to obtain a localized reflection response for patch-level evidence decomposition. Let $Z_t \in \mathbb{R}^{H\times W\times D}$ and $Z_t^{\leftrightarrow} \in \mathbb{R}^{H\times W\times D}$ denote their post-merge feature grids. We first reverse the width axis of $Z_t^{\leftrightarrow}$ to align reflected patches with their corresponding locations in the original image and apply a 3×7 local mean filter $\mathcal{S}(\cdot)$ to reduce patch-level noise. To tolerate minor horizontal misalignment after reflection, we further search over neighboring offsets $\delta \in \{-1, 0, 1\}$. The reflection response at location $p = (h, w)$ is

$$\rho_{t,p} = \min_{\delta\in\{-1,0,1\}} \frac{\|\mathcal{S}(Z_t)_p - \mathcal{S}(\mathrm{Align}(Z_t^{\leftrightarrow}))_{p+\delta}\|_2}{\|\mathcal{S}(Z_t)_p\|_2 + \|\mathcal{S}(\mathrm{Align}(Z_t^{\leftrightarrow}))_{p+\delta}\|_2 + \epsilon}, \tag{1}$$

where $p+\delta = (h, w+\delta)$ denotes a horizontal offset, $\mathrm{Align}(\cdot)$ denotes horizontal realignment of the reflected features for patch-wise comparison with the original image.

3.2.2 SEEDED PROTOTYPE DECOMPOSITION.

Reflection response provides a useful cue for identifying textual regions, but it may also introduce noisy responses. We therefore use it only to discover high-confidence seeds and perform patch decomposition through feature-space prototype comparison. Since many video frames may not contain any textual content, we further introduce a robust peak statistic to identify frames with reliable textual cues. For frame t,

$$z_t^{\text{peak}} = \frac{Q_{0.95}(\rho_t) - \text{med}(\rho_t)}{1.4826\ \text{MAD}(\rho_t) + \epsilon}. \tag{2}$$

where $Q_{0.95}(\rho_t)$ denotes the 95th percentile of the patch-wise reflection responses in frame t, $\text{med}(\rho_t)$ is their median, and $\text{MAD}(\rho_t) = \text{med}(|\rho_t - \text{med}(\rho_t)|)$ is the median absolute deviation. The factor 1.4826 provides a robust scale normalization. Frames satisfying $z_t^{\text{peak}} \geq 2.75$ contribute their top 5% response patches as textual seeds, while the bottom 40% patches of every frame form visual seeds. Denote the two seed sets by $\mathcal{Y}_T$ and $\mathcal{Y}_V$, respectively.

Let x_p be the ℓ_2-normalized, smoothed feature of patch p. We first identify feature dimensions that are discriminative between the two seed groups. For each channel d, we compute a Fisher score

$$F_d = \frac{(\mu_{T,d} - \mu_{V,d})^2}{\sigma_{T,d}^2 + \sigma_{V,d}^2 + \epsilon}, \tag{3}$$

where $\mu_{T,d}$ and $\mu_{V,d}$ are the mean feature values of textual and visual seeds on channel d, and $\sigma_{T,d}$ and $\sigma_{V,d}$ are the corresponding standard deviations. We retain the 64 highest-scoring channels.

In this discriminative subspace, we construct textual and visual prototypes c_T and c_V as the normalized centroids of the corresponding seed features. Each patch is then assigned a soft textuality score:

$$P_T(p) = \text{Sigmoid}\left(\frac{x_p^\top c_T - x_p^\top c_V}{\text{Std}_{p' \in \mathcal{P}_t}(x_{p'}^\top c_T - x_{p'}^\top c_V) + \epsilon} \right), \tag{4}$$

where $\mathcal{P}_t$ denotes all patches in frame t. Textual and visual seeds are assigned fixed probabilities of 1 and 0, respectively. The resulting evidence masks are defined as

$$M_t^T = \{p \in \mathcal{P}_t : P_T(p) \geq 0.5\}, \qquad M_t^V = \mathcal{P}_t \setminus M_t^T. \tag{5}$$

3.3 OPTION-AWARE PATCH RELEVANCE

Different candidate options may rely on different evidence from the video. DAES therefore maintains option-specific relevance rather than a single question-level relevance map. Given the candidate option set $\mathcal{O} = \{o_i\}_{i=1}^O$, we form an option-conditioned query h_i by concatenating the question q with each option o_i.

3.3.1 VALUE-SPACE TOKEN–PATCH SIMILARITY

Motivated by recent explorations of Value-space interactions for visual relevance estimation (Wang et al., 2026), we measure the relevance between option-conditioned query tokens and visual patches in the MLLM Value space. At layer ℓ, under grouped query attention, let $v_{i,k,g}^\ell$ and $v_{t,p,g}^\ell$ denote the Value-projection vectors of query token k and visual patch p at key–value head g, respectively. Their relevance is computed as the mean per-head cosine similarity:

$$C_{i,k,t,p}^\ell = \frac{1}{G} \sum_{g=1}^{G} \frac{\left\langle v_{i,k,g}^\ell, v_{t,p,g}^\ell \right\rangle}{\left\| v_{i,k,g}^\ell \right\|_2 \left\| v_{t,p,g}^\ell \right\|_2}. \tag{6}$$

3.3.2 SPATIALLY DISCRIMINATIVE TOKEN AGGREGATION

Different query tokens may contribute differently to evidence selection. We estimate their relative importance from their token–patch relevance patterns. For each query token, we measure its rele-

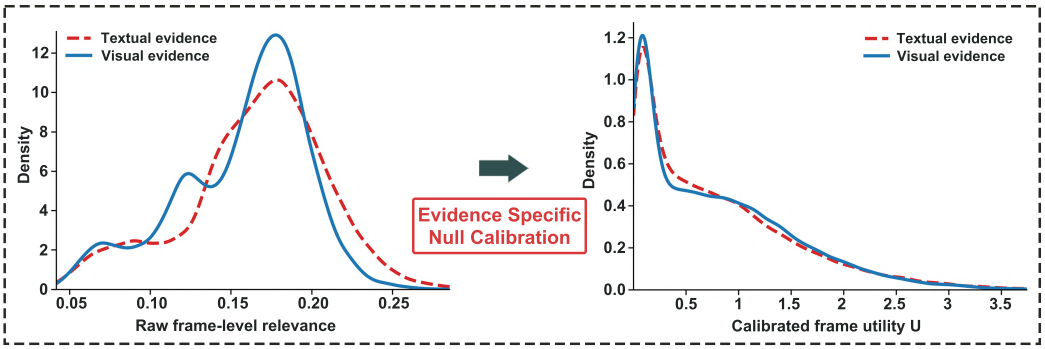

Figure 4: Evidence-specific null calibration aligns the relevance distributions of Visual and Textual Evidence into a comparable utility space.

vance concentration by comparing the top $r = 20\%$ patch responses with the median response:

$$d^{\ell}_{i,k,t} = \max\big(\mathrm{TopMean}_r\left(\{C^{\ell}_{i,k,t,p}\}_{p\in\mathcal{P}_t}\right) - \mathrm{med}_{p\in\mathcal{P}_t}\, C^{\ell}_{i,k,t,p},\, 0\big)\,,$$

$$\pi^{\ell}_{i,k,t} = \frac{1 + d^{\ell}_{i,k,t}/(\bar{d}^{\ell}_{i,t} + \epsilon)}{\sum_{k'\in\mathcal{K}_i}\left[1 + d^{\ell}_{i,k',t}/(\bar{d}^{\ell}_{i,t} + \epsilon)\right]}. \tag{7}$$

Here, $\mathcal{K}_i$ denotes the content tokens of the option-conditioned query h_i, $\mathrm{TopMean}_r$ averages the top r fraction of patch responses, and $\bar{d}^{\ell}_{i,t}$ is the mean concentration score over $k \in \mathcal{K}_i$.

The layer-specific patch relevance is then obtained by aggregating token-level relevance:

$$R^{\ell}_{i,t,p} = \sum_{k\in\mathcal{K}_i} \pi^{\ell}_{i,k,t} C^{\ell}_{i,k,t,p}. \tag{8}$$

To reduce layer-specific variations, we average the relevance maps over $\mathcal{L} = \{7, 11, 15\}$:

$$R_{i,t,p} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} R^{\ell}_{i,t,p}. \tag{9}$$

Finally, mask-guided pooling converts the patch relevance map into separate scores for Visual and Textual Evidence. For evidence type $b \in \{V, T\}$,

$$S^{b}_{i,t} = \mathrm{TopMean}_{20\%}\left(\{R_{i,t,p} : p \in M^{b}_{t}\}\right). \tag{10}$$

where M^{b}_{t} denotes the corresponding evidence mask.

3.4 Evidence-Specific Null Calibration

As shown in Fig. 4, Visual and Textual Evidence follow different raw relevance distributions after mask-guided pooling. We therefore calibrate them separately using evidence-specific empirical null distributions, mapping their scores into a shared space for direct comparison.

To construct the null distributions, we pair videos with unrelated queries to reduce semantic alignment. Applying the same relevance estimation and mask-guided pooling to these mismatched pairs, we collect the resulting scores for each evidence type $b \in \{V, T\}$ into an empirical null set $\mathcal{F}^{b}_{0}$.

Given an observed score $S^{b}_{i,t}$, we compute its empirical upper-tail probability and calibrated utility:

$$p^{b}_{i,t} = \frac{1 + \sum_{x\in\mathcal{F}^{b}_{0}} \mathbf{1}[x \geq S^{b}_{i,t}]}{|\mathcal{F}^{b}_{0}| + 1},$$
$$U^{b}_{i,t} = -\log_{10} p^{b}_{i,t}. \tag{11}$$

where $\mathbf{1}[x \geq S^{b}_{i,t}]$ equals 1 if the condition holds and 0 otherwise. A larger $U^{b}_{i,t}$ indicates stronger query-specific relevance beyond the corresponding null baseline. This calibration preserves within-type ranking while enabling direct comparison between Visual and Textual Evidence. Details of the null distribution construction are provided in Appendix A.2.

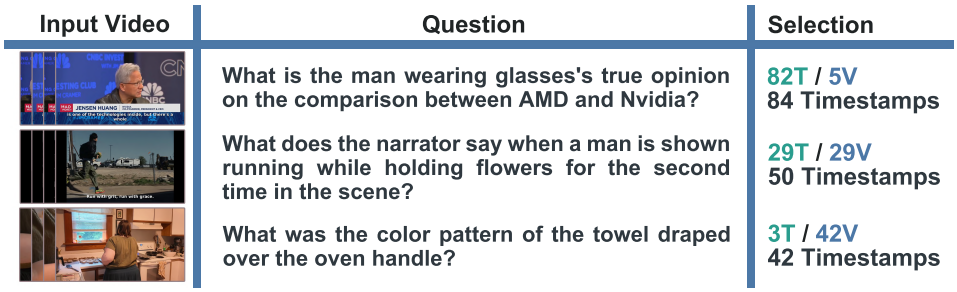

Figure 5: Examples of question-adaptive Visual/Textual Evidence selection. DAES adjusts the amount of selected Textual (T) and Visual (V) Evidence according to the information required by different questions.

Table 1: Results on the Video-MME-v2 with Qwen3.5-9B as the answerer. Frame-selection baselines use 32 full frames, while DAES uses the same token budget.

Method	Acc. (%)	Nonlin.	L1	L2	L3	Consistency	Coherence
Uniform32	32.16	14.83	18.34	13.84	13.48	15.22	14.09
AKS	35.00	17.11	26.34	14.65	13.48	18.77	14.03
Q-Frame	34.66	16.92	27.86	14.52	12.27	17.94	15.04
WFS-SB	36.00	17.32	27.58	16.30	12.07	18.82	14.54
VTR-VLM	36.25	17.33	26.42	17.42	12.87	19.88	13.81
DAES (0.8B)	39.34	20.94	31.84	17.77	16.85	21.74	19.46
DAES (2B)	**40.88**	**22.28**	**33.90**	**19.63**	**17.40**	**23.66**	**19.72**

3.5 Option-Balanced Evidence Selection

After calibration, Visual and Textual Evidence candidates from all candidate options are merged into a unified pool and ranked by their calibrated utilities. To balance coverage across candidate options, DAES selects candidates cyclically while preserving the ranking within each option until the visual-token budget is exhausted.

The selected evidence masks are converted into visual tokens, ordered by timestamp, and passed to the final reasoning MLLM. Selecting evidence rather than complete frames reduces the token cost per timestamp, enabling broader temporal coverage under the same budget. As shown in Fig. 5, DAES also adapts the allocation between Visual and Textual Evidence to each question.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate DAES on two long-video question-answering benchmarks. **Video-MME-v2** (Fu et al., 2026) contains 800 videos and 3,200 questions in the evaluation split, with four questions associated with each video. **LongVideoBench (LVB)** (Wu et al., 2024) evaluates long-context video understanding over videos of varying durations. We use its validation set containing 753 videos and 1,337 multiple-choice questions.

Metrics. For Video-MME-v2, we report question-level accuracy together with the official group-based metrics. **NonLin.** applies non-linear scoring to groups of related questions, emphasizing jointly correct predictions rather than isolated correct answers. **Consistency** measures stable performance across related questions targeting the same underlying capability, while **Coherence** evaluates correctness across dependent reasoning steps. We also report the official difficulty-level scores **L1–L3**. For LongVideoBench, we report overall accuracy together with accuracy on short ($\leq$ 2 min), medium (2–10 min), and long ($\geq$ 10 min) videos.

Table 2: LongVideoBench accuracy (%) across different video durations.

Method	Short ($\leq$ 2 min)	Medium (2–10 min)	Long ($\geq$10 min)	Overall
Uniform32	**82.27**	62.93	56.36	65.37
AKS	80.06	70.98	58.83	68.29
Q-Frame	74.52	69.76	62.72	68.06
WFS-SB	80.61	70.98	61.66	69.63
VTR-VLM	82.18	68.14	65.08	71.18
DAES (0.8B)	79.22	72.93	64.84	71.20
DAES (2B)	81.72	**75.61**	**66.43**	**73.37**

Table 3: Diagnostic results on Video-MME-v2. (a) Decomposition quality using rendered subtitle regions as partial positive annotations. (b) Accuracy on videos without rendered subtitles.

(a) Decomposition quality

Area Coverage (%)		Patch Recall (%)		
Micro	**Macro**	> 0	$\geq 25\%$	$\geq 50\%$
86.10	79.62	83.59	87.57	92.77

(b) Original videos

Method	Acc. (%)
Uniform32	30.00
DAES (2B)	**34.56**

Implementation. We use Qwen3.5-9B (Qwen Team, 2026) as the answer model for all methods and uniformly sample 256 candidate timestamps per video. When subtitle annotations are available, each sampled frame is rendered only with the subtitle active at that timestamp, without accumulating subtitles from adjacent timestamps. All methods use the same rendered candidate pool and receive no subtitle strings, transcripts, OCR outputs, or ground-truth answers as additional text input. For the main experiments, DAES uses a visual-token budget of $B = 11{,}700$, matching the cost of approximately 32 full frames used by the frame-selection baselines. Further implementation and baseline details are provided in the appendix.

4.2 Results and Analysis

Main results. Tables 1 and 2 summarize the main results. On Video-MME-v2, DAES achieves 40.88% accuracy, improving over Uniform32 by 8.72 points while obtaining the best results across all official group-based metrics. On LongVideoBench, DAES reaches 73.37% overall accuracy, with particularly strong performance on medium and long videos (75.61% and 66.43%). Notably, the 0.8B selector still achieves 39.34% and 71.20% on the two benchmarks, showing that effective evidence selection can be performed with a substantially smaller model.

Decomposition quality. Since rendered subtitles provide partial positive annotations rather than exhaustive text masks, we evaluate coverage and recall without treating non-subtitle regions as negatives. As shown in Table 3 (a), DAES achieves 86.10% micro and 79.62% macro subtitle-area coverage. Patch recall further increases from 83.59% with any subtitle overlap to 92.77% when subtitles occupy at least half of the patch, showing reliable identification of text-dominant patches.

Answerer generalization. Table 4 evaluates the transferability of DAES across different answerer models. Using evidence selected by the Qwen3.5-2B selector, DAES improves Qwen3.5-27B from 36.31% to 47.09%, yielding a 10.78-point gain over Uniform32. We further transfer the selected evidence to InternVL3.5-8B (Wang et al., 2025); since the two models use different visual patch grids, the transfer is performed through pixel-level patch mapping. DAES improves InternVL3.5-8B from 28.16% to 35.41%, a gain of 7.25 points. These results show that DAES generalizes across both model scales and architectures.

4.3 Ablation Studies

Component ablation. Table 5(a) evaluates the main components of DAES. Replacing V/T decomposition with full-frame selection reduces accuracy from 40.88% to 38.26%, demonstrating the benefit of separate Visual/Textual Evidence selection. Removing null calibration or cyclic evidence

Table 4: Results of answerer generalization on Video-MME-v2.

Method	Acc. (%)	Nonlin.	L1	L2	L3	Consistency	Coherence
Uniform32 + Qwen3.5-27B	36.31	18.36	21.49	17.18	17.36	18.02	18.98
DAES + Qwen3.5-27B	**47.09**	**28.44**	**39.83**	**27.40**	**22.56**	**29.91**	**25.72**
Uniform32 + InternVL3.5-8B	28.16	12.46	15.78	10.68	11.79	12.32	12.72
DAES + InternVL3.5-8B	**35.41**	**18.14**	**27.76**	**16.19**	**13.93**	**18.38**	**17.70**

Table 5: Component and token-budget ablations on Video-MME-v2.

(a) Component ablation

Variant	Acc. (%)
DAES (base)	**40.88**
w/o V/T decomposition (Full32)	38.26
w/o calibration	39.88
w/o cyclic evidence selection	39.91
Q+all-options relevance	39.00
Q-only relevance	38.09

(b) Token budget

Budget	Acc. (%)
20F ($B = 7313$)	39.66
26F ($B = 9506$)	40.16
32F ($B = 11700$)	40.88
38F ($B = 13894$)	40.72
44F ($B = 16088$)	**41.22**

selection reduces accuracy to 39.88% and 39.91%, respectively. Replacing option-aware relevance with Q+all-options or question-only relevance yields 39.00% and 38.09%, respectively, highlighting the importance of maintaining option-specific evidence relevance.

Original-video setting. As shown in Table 3(b), DAES still improves accuracy from 30.00% to 34.56% on the original videos. Compared with the rendered setting, Textual Evidence decreases from 42.13 to 19.15, while Visual Evidence increases from 28.77 to 33.25, and selected timestamps drop from 59.34 to 44.19. The remaining Textual Evidence likely comes from naturally occurring visual text.

Effect of token budget. Table 5(b) studies the sensitivity of DAES to the token budget. Increasing the budget from 7,313 to 11,700, corresponding approximately to the cost of 20 to 32 complete frames, improves accuracy from 39.66% to 40.88%. With larger budgets, performance remains stable and reaches the best accuracy of 41.22% at $B = 16{,}088$. However, the improvement beyond $B = 11{,}700$ is relatively small compared with the additional token cost. We therefore use $B = 11{,}700$ as the default setting, which provides a favorable balance between token cost and question-answering performance.

5 Conclusion

We introduced **Decoupled Adaptive Evidence Selection (DAES)**, a training-free framework for efficient video understanding under a fixed token budget. Moving beyond conventional frame-level selection, DAES separates Visual and Textual Evidence at the patch level, estimates their option-conditioned relevance in the Value space of an MLLM, and calibrates the two evidence types using evidence-specific null distributions. The calibrated evidence is then adaptively selected across timestamps and candidate options, allowing the model to preserve question-relevant information while covering a broader temporal context.

Experiments on Video-MME-v2 and LongVideoBench demonstrate substantial improvements over frame-selection baselines under comparable token costs, while ablation studies validate the contributions of the main components. Despite these results, reflection-based decomposition may still miss or misclassify textual regions under challenging visual conditions, which can affect subsequent evidence selection. Moreover, DAES currently considers only two broad evidence types, Visual and Textual Evidence. Future work will explore more accurate textual-region decomposition and extend DAES to richer evidence categories beyond Visual and Textual Evidence. Overall, our results suggest that explicitly modeling different evidence types and performing adaptive evidence selection provides an alternative to conventional frame-level selection for efficient video understanding.

AI USE STATEMENT

In this work, we used generative AI tools to assist with searching for and identifying relevant literature, drafting and debugging research code, and editing the manuscript for clarity, readability, and language quality. All AI-assisted outputs were manually reviewed by the authors. Relevant literature was checked against the original sources, AI-assisted code was inspected and tested before use, and all manuscript edits were reviewed for technical accuracy and consistency with the authors' intended claims. We take full responsibility for the final content of the paper.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ArXiv*, abs/2210.09461, 2022.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 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*, 2024.

Wang Chen, Yu Zeng, Yong Luo, Tianyu Xie, Luo-Jun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based frame selection by detecting semantic boundary for long video understanding. *ArXiv*, abs/2603.00512, 2026.

Chaoyou Fu, Hao Yuan, Yuhao Dong, Yifan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jin-Sen Su, Chengwu Long, Xiaoyao Xie, Yong Xie, Xiawu Zheng, Xuejiao Yang, Haoyu Cao, Yun-Sheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *ArXiv*, abs/2604.05015, 2026.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. *ArXiv*, abs/2412.16117, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, 2025.

Tiancheng Wang, Yuguang Yang, Lin-Lin Yang, Shaohui Lin, Juan Zhang, Guodong Guo, and Baochang Zhang. Clip in mirror: Disentangling text from visual images through reflection. *Advances in Neural Information Processing Systems 37*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. Visually-guided policy optimization for multimodal reasoning. *arXiv preprint arXiv:2604.09349*, 2026.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *ArXiv*, abs/2407.15754, 2024.

Zhirong Wu, Xiaodong Wang, Langling Huang, Teng Xu, and Peixi Peng. A Training-Free Framework for Long Video Understanding via Video-Query-Options Similarity. In *International Conference on Learning Representations*, 2026.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024.

Shaojie Zhang, Jia-Hui Yang, Jianqin Yin, Zhen-Bo Luo, Jian Luan, and Milm Plus. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22056–22065, 2025.

Yuan Zhang, Chunkai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shang-Hang Zhang. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *ArXiv*, abs/2410.04417, 2024.

Zi-Xuan Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. *ArXiv*, abs/2510.27280, 2025.

A Appendix

A.1 Subtitle-Rendered Dataset Construction

We construct subtitle-rendered versions of Video-MME-v2 and LongVideoBench (LVB) using temporally aligned subtitle annotations. For each video, 256 candidate frames are uniformly sampled over the full video duration. The word-level subtitle annotations are grouped into subtitle cues using a temporal gap threshold of 0.7 s, with at most 10 words per cue and a maximum cue duration of 4 s. No additional subtitle hold time is applied. At each candidate timestamp, only the subtitle cue active at that moment is rendered onto the corresponding frame; subtitles between adjacent candidate timestamps are not accumulated. If no cue is active, the original frame is retained.

For Video-MME-v2, subtitles are rendered near the bottom center of the frame using DejaVu Sans Bold at font size 60, with white text and a black shadow offset by 5 pixels. The bottom margin is 65 pixels. For LVB, we use the validation split and render subtitles at font size 40, with a 43-pixel bottom margin and a 3-pixel shadow offset. These rendering parameters differ to accommodate the different frame resolutions of the two benchmarks. For both datasets, the rendered text width is limited to 0.88 of the frame width, with a character budget of 42 characters per line. Some examples of subtitle-rendered datasets are shown in Fig.6

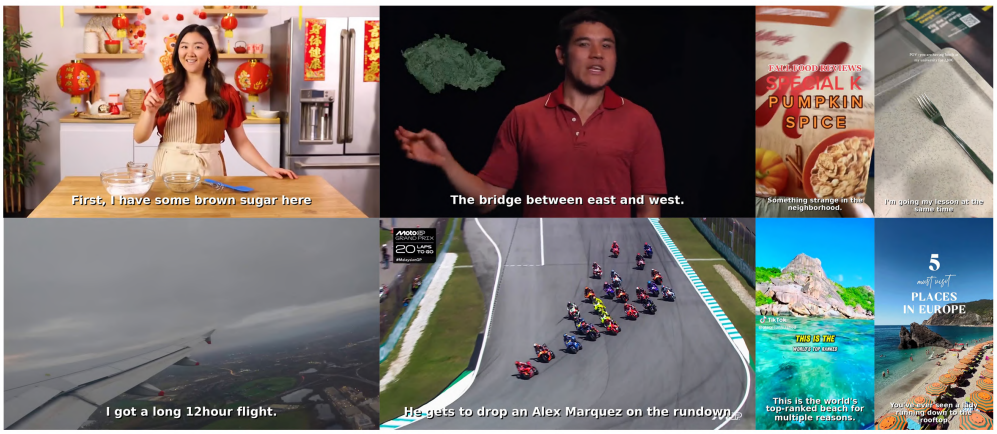

Figure 6: Examples of subtitle-rendered datasets.

A.2 Details of Null Distribution Construction

Calibration data and mismatched pairing. We construct the empirical null distributions from a fixed subset of 400 questions from 100 videos. Each question together with all of its candidate options is paired with a different video using a fixed one-to-one permutation. Same-video pairings are not allowed, and the permutation is generated once with random seed and reused throughout all experiments.

Only the textual query is changed by this mismatch. The target video's 256 candidate frames, visual preprocessing, post-merge token grid, and Visual/Textual Evidence masks remain unchanged. Each candidate option is processed independently using the same option-conditioned query format as in the main method. No ground-truth answer is used when constructing the mismatched pairs.

Evidence-specific null distributions. For each mismatched video–query pair, we apply the same relevance estimation and mask-guided pooling as in the main pipeline. The resulting scores are collected separately for Visual and Textual Evidence to construct $\mathcal{F}_0^V$ and $\mathcal{F}_0^T$. Frames without valid Textual Evidence do not contribute a textual null score. The two evidence types therefore use separate empirical reference distributions throughout calibration.

Cross-fitting. The remaining 2,800 questions use the full frozen 400-question null reference. For the 400 questions used to construct the reference set, we apply deterministic five-fold cross-fitting.

The 400 questions are divided into five fixed folds, and each question is calibrated using null scores constructed only from the other four folds. Thus, a question's own mismatched observations are never included in the reference distribution used to calibrate that question.

The calibration subset, mismatched pairings, and fold assignments are fixed before evidence selection and reused for all experiments. No ground-truth answer, QA correctness signal, or answer-model output is used during null construction or calibration.

A.3 Evidence Token Construction

All evidence masks are defined on the post-merge visual-token grid of the Qwen vision encoder. For a selected Visual or Textual Evidence candidate, we retain only the visual tokens covered by its evidence mask rather than forwarding the complete frame.

Evidence candidates are considered in the option-balanced cyclic order described in the main paper. If the evidence corresponding to a candidate has already been selected by another option, the candidate is skipped to avoid duplicate selection. When both Visual and Textual Evidence at the same timestamp are selected, their masks are merged, recovering the complete frame at that timestamp.

Selection continues until adding the next non-duplicate candidate would exceed the visual-token budget. We use $B = 11{,}700$ in the main experiments. Finally, the selected timestamps are sorted in temporal order, and the corresponding visual tokens are passed to the final answer MLLM.

A.4 Details of Baseline Reproduction

Common evaluation protocol. For a controlled comparison, all frame-selection baselines operate on the same frozen pool of 256 uniformly sampled subtitle-rendered RGB frames. We adapt each released selector to this common candidate pool and set the output budget to 32 unique frames. The selected frames are temporally ordered and passed as complete visual inputs to the same Qwen3.5-9B answer MLLM, using identical answer-model preprocessing, QA prompting, and decoding settings. No reflection-based decomposition, patch-level evidence selection, or null calibration from DAES is applied to the baselines.

AKS. We reproduce AKS using its official implementation with the BLIP image–text matching scorer. Following the released implementation, the question alone is used as the textual query, without candidate options. Each candidate frame I_t is scored by the positive-class probability of the BLIP image–text matching head:

$$s_t = \mathrm{softmax}\left(\mathrm{BLIP}_{\mathrm{ITM}}(I_t, q)\right)_1 . \tag{12}$$

Frame selection follows the official recursive procedure with min–max score normalization and mean–standard-deviation based segment splitting. We retain the released settings $t_1 = 0.8$, $t_2 = -100$, maximum recursion depth 5, and sampling ratio 1. The final frame budget is set to 32 unique candidates for the common evaluation protocol.

Q-Frame. We reproduce Q-Frame using its released Long-CLIP-based query-aware selector. The question alone is used as the textual query, without candidate options. Let f_I and f_T denote the Long-CLIP image and text encoders. For each candidate frame I_t, we compute

$$\ell_t = f_I(I_t)^\top f_T(q), \qquad t = 1, \ldots, 256. \tag{13}$$

The similarities are converted into a sampling distribution using temperature $\tau = 0.8$:

$$p_t = \frac{\exp(\ell_t/\tau)}{\sum_{j=1}^{256} \exp(\ell_j/\tau)}. \tag{14}$$

We then apply Gumbel–Max based ranking and retain the top 32 unique candidates. The random generator is deterministically seeded per question for reproducibility. We use the fixed-frame configuration and disable Multi-Resolution Adaptation (MRA), so that all selected frames are forwarded as complete visual inputs. The selected frames are finally reordered by their original video timestamps before answer generation.

WFS-SB. We reproduce WFS-SB using its released wavelet-based selection pipeline with the supported BLIP feature extractor. The textual query concatenates the question and all candidate options. Query–frame relevance scores over the common 256-frame pool are processed by wavelet-based multi-resolution analysis to identify semantic boundaries and partition the video into temporal segments. The frame budget is then allocated across the resulting segments, followed by Maximal Marginal Relevance (MMR) selection within each segment to balance query relevance and visual diversity.

We use the released settings with wavelet type *db4*, decomposition level 3, height factor 0.5, prominence factor 0.05, minimum-distance ratio 0.02, and minimum absolute distance 5. The segment scoring weights are

$$(w_{\text{duration}}, w_{\text{mean}}, w_{\max}, w_{\text{var}}) = (0.4, 0.2, 0.3, 0.1), \tag{15}$$

with strictness factor 1.2, temperature 1.0, and $\lambda = 0.5$. We set the final frame budget to 32 unique candidates for the common evaluation protocol.

VTR-VLM. We reproduce VTR-VLM using its released allocation procedure with the PE-Core-G14-448 (PE-G/14) vision–language encoder. The 256 candidate frames are divided into 16 consecutive temporal segments of 16 frames each. Frame-level visual embeddings are ℓ_2-normalized, and each segment representation is obtained by averaging the embeddings within that segment.

For each candidate option o_i, VTR-VLM constructs a question–option query and computes its similarity to every segment. The segment relevance is defined as the maximum similarity across candidate options:

$$r_j = \max_i \left\langle e^{\text{text}}_{q,o_i}, e^{\text{seg}}_j \right\rangle . \tag{16}$$

We retain the released allocation procedure with overall allocation parameter 8, per-segment budget 4, and candidate allocation levels $[2, 4, 8]$. This configuration directly produces 32 unique selected frames without selecting a larger set followed by post-hoc truncation. The selected indices are mapped back to their timestamps in the common candidate pool.

Downstream evaluation. For all four baselines, the selected RGB frames are retained as complete visual inputs without patch-level pruning. The downstream answer model and its input processing are therefore controlled across methods, while each baseline retains its own frame-scoring and temporal-selection strategy.

A.5 Details of Rendered-Subtitle Decomposition Evaluation

Evaluation setup. We evaluate the spatial quality of the text–visual decomposition using rendered subtitles as a positive-only proxy for textual regions. The evaluation uses the same frozen 256-frame candidate pool as the main experiments. We use 50 videos for development and another 50 non-overlapping videos as a held-out set. All decomposition parameters are fixed before evaluation, and no question, candidate option, ground-truth answer, or QA prediction is used in this diagnostic.

Subtitle support on the visual-token grid. For each sampled frame, the binary pixel mask generated together with the rendered subtitle is projected through the same spatial preprocessing and patch-merging geometry as the corresponding RGB frame. This produces a continuous coverage map

$$C_t(u, v) \in [0, 1], \tag{17}$$

where $C_t(u, v)$ denotes the fraction of post-merge visual token (u, v) covered by the rendered subtitle. The resulting mask is therefore defined on the same visual-token grid as the predicted Textual Evidence mask, allowing direct token-level comparison. This proxy annotates only the rendered subtitles and does not cover other text naturally present in the video.

Frozen decomposition. We apply the same reflection-based decomposition as in the main experiments without further tuning. The decomposition uses a 3×7 local aggregation window, horizontal alignment offsets $\{-1, 0, +1\}$, the top 5% and bottom 40% reflection-response tokens as textual and visual seeds, respectively, and the 64 channels with the highest Fisher scores for prototype construction. The final binary Textual Evidence mask is obtained by thresholding the textuality probability at 0.5.

Coverage metrics. Let $M_t(u, v) \in \{0, 1\}$ denote the predicted Textual Evidence mask. For each frame containing rendered subtitle support, we compute

$$A_t = \sum_{u,v} C_t(u, v), \qquad I_t = \sum_{u,v} C_t(u, v) M_t(u, v), \tag{18}$$

where A_t is the total subtitle support and I_t is the portion covered by the predicted mask. We report

$$\text{MicroCoverage} = \frac{\sum_t I_t}{\sum_t A_t}, \qquad \text{MacroCoverage} = \frac{1}{|\mathcal{P}|} \sum_{t \in \mathcal{P}} \frac{I_t}{A_t}, \tag{19}$$

where $\mathcal{P}$ contains frames with non-zero subtitle support. Micro coverage weights frames by subtitle area, whereas macro coverage gives equal weight to each subtitle-positive frame.

We additionally report patch recall for tokens with different amounts of subtitle coverage. Specifically, positive tokens are defined by $C_t(u, v) > 0$, $C_t(u, v) \geq 0.25$, and $C_t(u, v) \geq 0.5$, corresponding to any, at least 25%, and at least 50% subtitle coverage, respectively. Patch recall measures the fraction of these tokens included in the predicted Textual Evidence mask.

Positive-only interpretation. Rendered subtitles provide only partial positive annotations rather than complete text-segmentation ground truth. A predicted Textual Evidence token outside the rendered subtitle region may correspond to natural scene text, signs, logos, or interface text and therefore cannot be treated as a false positive. We consequently report coverage and recall rather than precision or IoU.

A.6 Details of Answerer Generalization

Evaluation protocol. To evaluate whether the selected evidence generalizes beyond the main Qwen3.5-9B answerer, we reuse exactly the same frozen DAES selection results with Qwen3.5-27B and InternVL3.5-8B. For every question, the candidate-frame pool, selected timestamps, and corresponding evidence masks remain unchanged. No relevance estimation, reflection-based decomposition, null calibration, or evidence selection is recomputed for the new answerers.

All answerers receive the same question and candidate options and use deterministic decoding. The evaluation is conducted on the same Video-MME-v2 set of 800 videos and 3,200 questions. Subtitles remain part of the rendered RGB frames and are not provided as additional text input.

Qwen3.5-27B. Since Qwen3.5-27B uses the same visual-token geometry as the Qwen3.5 models used in the main pipeline, the frozen post-merge evidence masks can be reused directly. For each selected timestamp, the 27B model first extracts the complete post-merge visual-token sequence. The frozen evidence mask is directly applied to the 27B visual-token sequence, and only the selected visual tokens are passed to the language model.

We additionally verify that the post-merge visual grid produced by the 27B processor matches the grid used during evidence selection. Thus, the only model change in this experiment is the answerer checkpoint, while the selected timestamps and visual evidence remain fixed.

InternVL3.5-8B. InternVL3.5-8B uses a different vision preprocessing pipeline and visual-token grid, so Qwen patch indices cannot be transferred directly. We instead map the frozen Qwen evidence masks through normalized image coordinates.

For a Qwen post-merge grid of size $H_Q \times W_Q$, each selected cell (u, v) is represented by its normalized image region

$$B^Q_{u,v} = \left[\frac{v}{W_Q}, \frac{u}{H_Q}, \frac{v+1}{W_Q}, \frac{u+1}{H_Q} \right]. \tag{20}$$

InternVL3.5-8B processes each selected frame as a single 448×448 tile, represented by a 16×16 visual-token grid. For each InternVL token region B^I_j, we compute the fraction of its area overlapping the selected Qwen regions:

$$\gamma_j = \frac{\text{area}\left(B^I_j \cap \bigcup_{(u,v):M^Q(u,v)=1} B^Q_{u,v} \right)}{\text{area}(B^I_j)}. \tag{21}$$

An InternVL visual token is retained when $\gamma_j \geq 0.5$. If no token in a selected frame reaches this threshold, the token with the largest overlap is retained. The mapping is performed independently for each selected timestamp.

After InternVL extracts its visual features, only the tokens retained by this spatial mapping are forwarded to the language model. No relevance score or reflection response is recomputed in the InternVL feature space.

Token-budget accounting. For Qwen3.5-27B, the original Qwen post-merge token accounting is preserved because the evidence masks are reused directly. InternVL3.5-8B has a different visual-token geometry, so its retained token count is determined by the spatial mapping above and is not expected to exactly match the Qwen token count. This experiment therefore evaluates whether the selected visual evidence transfers across answerer architectures, rather than enforcing identical model-specific token counts.

A.7 Details of Component Ablations

Common setting. All component ablations use the same 256 subtitle-rendered candidate frames, Qwen3.5-2B selector, Qwen3.5-9B answerer, and deterministic QA setting as the main experiments. Relevance is computed from layers $\{7, 11, 15\}$ with the same top-20% spatial aggregation. The visual-token budget is fixed to $B = 11{,}700$.

The full DAES model uses four components: option-specific relevance, Visual/Textual Evidence decomposition, evidence-specific null calibration, and option-balanced cyclic evidence selection. In each ablation, we change only one component and keep the others unchanged.

Without V/T decomposition. This variant removes the Visual/Textual Evidence decomposition. We still use the same 256 candidate frames and the same relevance scores, but selection is performed at the frame level instead of the evidence level. We select 32 complete frames and pass all visual tokens from these frames to the answerer.

Without null calibration. This variant keeps the Visual/Textual Evidence masks, option-specific relevance, and cyclic evidence selection. The only change is that evidence is ranked using the raw relevance scores instead of the calibrated scores. All other settings, including the visual-token budget, remain unchanged.

Without cyclic evidence selection. This variant keeps option-specific relevance and null calibration, but removes the cyclic selection across candidate options. Evidence candidates from all options are placed in one list and ranked by their calibrated scores. If the same evidence has already been selected through another option, the duplicate is skipped. The same visual-token budget is used.

Q+all-options relevance. This variant replaces the separate question–option queries with one query that contains the question and all candidate options. This produces one shared relevance map instead of one relevance map for each option. All candidate options then use this same relevance map, while calibration and evidence selection remain unchanged.

Q-only relevance. This variant computes relevance using only the question text and does not use any candidate option when estimating relevance. As a result, all candidate options again use the same relevance map. The following calibration and evidence-selection steps are unchanged.