REASON FIRST, RETRIEVE WHAT MATTERS: HYPOTHESIS-GUIDED KEYFRAME SELECTION FOR LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long video understanding demands that models recover and reason over sparse, temporally distributed evidence within limited frame budgets. Existing keyframe selection (KFS) methods usually retrieve query-relevant frames from a densely sampled candidate set before inference. However, these methods still have two key limitations. First, the global image–text matching strategy introduces unnecessary computation overhead when a small set of uniformly sampled frames already provides sufficient evidence. Moreover, they may miss evidence that distinguishes competing answers, while single-pass frame selection provides no opportunity to correct selection errors. To address the issues, we propose **HypoKFS**, a training-free and plug-and-play framework that couples keyframe selection with adaptive multi-step inference. Specifically, HypoKFS first reasons over sparsely sampled frames to construct an initial hypothesis space. Then the resulting answer distribution determines whether further frame selection is required. When additional frame selection is needed, the retained hypotheses guide discriminative frame selection, followed by renewed inference and hypothesis refinement. This iterative process focuses additional visual computation on unresolved hypotheses while avoiding unnecessary global frame matching. Experimental results on the Video-MME, LongVideoBench, and MLVU demonstrate consistent improvements across different MLLM backbones. HypoKFS improves accuracy by up to 6.7 percentage points and achieves 72.9% accuracy on Video-MME while applying image–text matching to only 1.3% of the video frames.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have substantially advanced video understanding and reasoning over temporally structured visual content Zhang et al. (2024); Bai et al. (2025b). As these models are extended from short clips to long videos Fu et al. (2025); Wu et al. (2024); Zhou et al. (2024); Yang et al. (2025a), the amount of visual content grows rapidly while task-relevant evidence often remains sparse over long temporal contexts Wu et al. (2024); Tang et al. (2025); Liu et al. (2025). Selecting informative frames under a limited visual budget has therefore become an important challenge for long-video understanding Tang et al. (2025); Zhang et al. (2025); Zhu et al. (2026).

A common strategy is *uniform sampling* Fu et al. (2025); Wu et al. (2024); Liu et al. (2025), as illustrated in Fig. 1(a), where a fixed number of frames are uniformly sampled before MLLM inference. Despite its simplicity, uniform sampling is query-agnostic and may miss task-relevant evidence under a limited visual budget Liu et al. (2025); Tang et al. (2025). Recent keyframe selection (KFS) methods therefore introduce more targeted frame-selection strategies Tang et al. (2025); Zhang et al. (2025); Sun et al. (2025); Zhu et al. (2026). As illustrated in Fig. 1(b), these methods typically construct a denser candidate set and select informative frames before MLLM inference using signals such as query relevance, temporal coverage, diversity, or adaptive exploration.

Despite this progress, performing keyframe selection entirely before inference does not exploit the predictive information already available from the MLLM. For some queries, inference over only a small number of uniformly sampled frames already assigns high confidence to one answer, making

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.

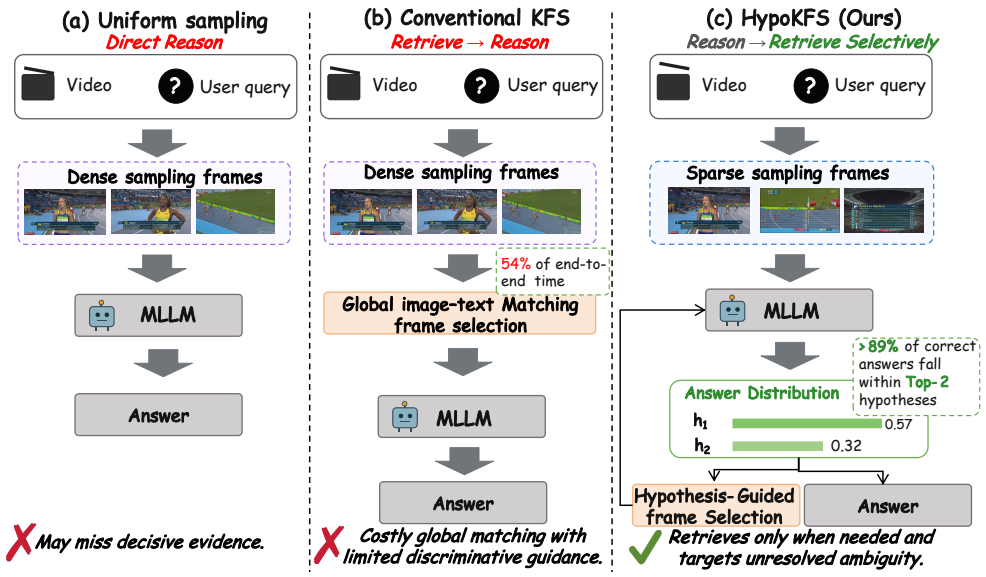

Figure 1: **Evolution of visual evidence selection for VideoQA. (a)** Uniform sampling uses fixed, query-agnostic frames. **(b)** Conventional KFS performs global image–text matching and frame selection before MLLM inference. **(c)** HypoKFS first infers from sparsely sampled frames and uses the resulting hypothesis space to guide subsequent frame selection.

additional global image–text matching unnecessary. For more ambiguous queries, multiple hypotheses remain plausible and further visual evidence is required. In such cases, query relevance alone is insufficient for effective frame selection, since a highly relevant frame may provide similar support for competing hypotheses and therefore offer limited discriminative value. Moreover, most existing KFS methods perform frame selection only once before MLLM inference. If informative evidence is missed during this stage, the selection cannot be revised according to subsequent model predictions. These limitations motivate coupling keyframe selection with the evolving answer distribution rather than treating it as a fixed preprocessing stage.

Fig. 1(c) provides two observations supporting this perspective. First, with only uniformly sampled sparse frames, 89.4% of correct answers on Video-MME are already covered by the top-2 hypotheses produced by initial MLLM inference. This indicates that inexpensive initial inference can provide a compact and informative hypothesis space for subsequent frame selection. Second, global image–text matching in conventional KFS accounts for approximately 54% of the end-to-end inference time, making unconditional matching over a dense candidate set costly. Together, these observations suggest that frame matching need not be performed before inference for every query. Instead, the model can first estimate the current answer distribution and invoke additional frame selection only when further visual evidence is required.

Motivated by these observations, we propose **HypoKFS**, a training-free and plug-and-play framework for hypothesis-guided iterative keyframe selection. HypoKFS first performs low-cost inference over uniformly sampled sparse frames to estimate the current answer distribution and construct a compact hypothesis space. Confident predictions are returned directly without additional frame matching, while ambiguous cases trigger hypothesis-guided frame selection that emphasizes both hypothesis relevance and discriminative utility. The selected frames are then incorporated into renewed inference, allowing the updated hypothesis space to guide subsequent frame selection. This iterative process progressively resolves competing hypotheses and terminates after a bounded number of rounds, with the highest-probability hypothesis returned as the final answer.

Our contributions are summarized as follows:

- **Resolving answer ambiguity through iterative frame selection.** We propose **HypoKFS**, a training-free and plug-and-play framework that iteratively selects discriminative frames according to the current hypothesis space, allowing competing answers to be progressively resolved while avoiding unnecessary frame matching for confident predictions.

- **Reducing redundant computation in keyframe selection.** We reformulate conventional single-stage keyframe selection as a hypothesis-guided iterative process, substantially reducing unnecessary global frame matching and providing a new perspective for efficient long-video understanding.
- **Consistent gains with lower matching cost.** Experiments on Video-MME, LongVideoBench, and MLVU demonstrate consistent improvements across different MLLM backbones while substantially reducing frame–hypothesis matching computation.

2 Related Work

2.1 Long Video Understanding

Long video understanding introduces additional challenges due to the large number of visual tokens and long-range temporal dependencies. Compared with short-video settings, long-form videos require models to reason over extended temporal contexts, substantially increasing computational and memory costs. Existing approaches mainly address this problem through memory mechanisms, visual token compression, or long-context modeling. For example, MovieChat Song et al. (2023) employs memory banks to maintain long-term context, while Video-Streaming Qian et al. (2024) and MA-LMM He et al. (2024) compress video features through compact temporal representations. LongVLM Weng et al. (2025) and LongVILA Chen et al. (2024) further extend the temporal context that can be processed by video MLLMs. More recent methods increasingly focus on reducing visual-token redundancy. LongVU Shen et al. (2025) performs adaptive spatiotemporal compression using inter-frame redundancy and cross-modal signals, while DyCoke Tao et al. (2025) dynamically compresses visual tokens and the corresponding KV cache during inference. PVC Yang et al. (2025b) progressively compresses visual tokens across frames, and VoCo-LLaMA Ye et al. (2025) learns compact visual representations that better align with the LLM's visual understanding process. Although these approaches substantially improve the scalability of video MLLMs, they mainly optimize the representation of already sampled visual content. HypoKFS instead focuses on determining which additional frames should be processed according to the model's current inference state.

2.2 Frame Selection

Frame selection reduces temporal redundancy by retaining a compact set of informative frames before video reasoning. Early methods such as AdaFrame Wu et al. (2019) learn adaptive sampling policies based on video content, while recent Video-LLM approaches increasingly incorporate query information into the selection process. Existing methods can be broadly divided into learning-based and training-free approaches. FrameVoyager Yu et al. (2024) and the M-LLM-based selector of Hu et al. Hu et al. (2025) learn frame-selection policies from task supervision or importance signals. In contrast, training-free methods such as BOLT Liu et al. (2025), AKS Tang et al. (2025), and Q-Frame Zhang et al. (2025) select frames using query relevance, model reasoning, or adaptive visual representations without additional training. Recent approaches further explore complementary selection criteria: MaxInfo Li et al. (2026b) selects representative frames by maximizing diversity in the embedding space, DIG Li et al. (2026a) adapts the selection strategy according to query type, and DAFS Wang et al. (2026) derives frame-level relevance from cross-modal attention at test time. Despite these advances, most existing methods determine the selected frames before the final answer inference. HypoKFS instead couples frame selection with the evolving answer distribution, allowing subsequent selections to be guided by the unresolved hypothesis space.

3 Method

3.1 Overview

Given a long video V and a user query q, HypoKFS answers q using a frozen multimodal large language model (MLLM) under a limited visual computation budget. We denote the candidate answer hypotheses by $\mathcal{H}$, containing C alternatives $h_1, \ldots, h_C$. HypoKFS first performs inference over a small number of uniformly sampled frames to estimate the answer distribution and construct a compact hypothesis space. This initial prediction also determines whether further visual evidence is

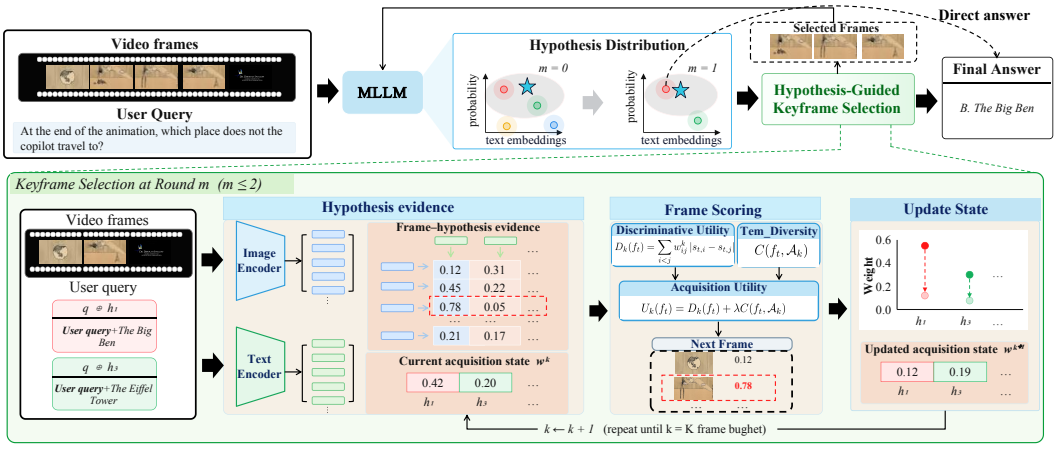

Figure 2: **Overview of HypoKFS.** Initial inference from sparse frames constructs a compact hypothesis space and determines whether additional computation is required. HypoKFS then iteratively selects discriminative frames and updates the hypothesis space through renewed inference for at most M rounds.

required. For queries requiring additional evidence, HypoKFS performs iterative keyframe selection conditioned on the current hypothesis space. Within each round, candidate frames are evaluated according to their ability to distinguish the remaining plausible hypotheses, while the acquisition state is progressively updated to avoid repeatedly emphasizing already covered hypotheses. The selected frames are then passed to the same frozen MLLM for renewed inference. The resulting answer distribution updates the hypothesis space and determines whether another selection round is needed. The process runs for at most M rounds; if the stopping criterion remains unsatisfied after the final round, the hypothesis with the highest probability is returned as the final answer.

3.2 Initial Hypothesis Space Construction

We uniformly sample N sparse frames from V and denote this initial frame set by $\mathcal{F}_{\text{init}}$. Together with the query q, these frames are fed into the frozen MLLM. For each hypothesis h_i, we extract the probability of its first answer token and normalize the probabilities over all candidate hypotheses:

$$\mathbf{p}^0 = [p_1^0, \ldots, p_C^0], \qquad p_i^0 = P(h_i \mid \mathcal{F}_{\text{init}}, q). \tag{1}$$

Here, $P(h_i \mid \mathcal{F}_{\text{init}}, q)$ denotes the normalized first-token probability associated with hypothesis h_i, rather than the sequence probability of the complete answer. The distribution $\mathbf{p}^0$ therefore represents the relative support assigned by the MLLM to different answer hypotheses under the initial sparse visual context.

We sort the hypotheses in descending order of their probabilities and retain the smallest set whose cumulative probability mass reaches the coverage threshold ρ:

$$\mathcal{H}_\rho = \{h_{(1)}, \ldots, h_{(K_\rho)}\}, \qquad K_\rho = \min\left\{k : \sum_{i=1}^{k} p_{(i)}^0 \geq \rho\right\}. \tag{2}$$

The resulting $\mathcal{H}_\rho$ retains the hypotheses accounting for most of the current answer probability while excluding low-probability alternatives from subsequent frame matching. Its size K_ρ is query-dependent: a concentrated answer distribution preserves fewer hypotheses, whereas an ambiguous prediction retains more competing alternatives. After renewed inference in later rounds, the same construction is applied to the updated first-token distribution.

Selective computation. The maximum probability in the current answer distribution is used as the routing score. If it reaches the routing threshold τ, HypoKFS directly returns the top-ranked hypothesis and bypasses additional frame matching. Otherwise, the query enters the keyframe-selection stage. The two thresholds serve different purposes. The routing threshold τ determines

whether additional visual computation is required, whereas the coverage threshold ρ determines which plausible hypotheses participate in subsequent matching and discrimination. We use ρ no smaller than τ, ensuring that queries routed to frame selection retain multiple competing hypotheses. The same routing criterion is applied after every renewed inference. If it remains unsatisfied after round M, HypoKFS returns the highest-probability hypothesis from the final answer distribution.

3.3 HYPOTHESIS-GUIDED KEYFRAME SELECTION

For each routed query, we construct a denser candidate set containing N_c frames sampled from the video and denote it by $\mathcal{F}$. Frame selection is conditioned on the hypothesis space $\mathcal{H}_\rho$ obtained from the current inference round. For clarity, we omit the outer inference-round index in the following formulation.

Frame–hypothesis evidence modeling. For each retained hypothesis h_i, we concatenate the query q with its corresponding answer hypothesis to form $q \oplus h_i$. A frozen vision–language matching model evaluates the compatibility between candidate frame f_t and hypothesis h_i as

$$s_{t,i} = \mathrm{sim}\left(\phi_v(f_t), \phi_t(q \oplus h_i)\right), \tag{3}$$

where ϕ_v and ϕ_t denote the visual and textual encoders, respectively, and sim denotes cosine similarity between their normalized representations. The resulting scores form the frame–hypothesis evidence matrix $\mathbf{S}$ with N_c rows and K_ρ columns, corresponding to candidate frames and retained hypotheses. Each frame–hypothesis similarity is computed when the corresponding hypothesis first participates in selection and is cached thereafter, allowing the same similarities to be reused if the hypothesis remains active in later inference rounds.

Frame scoring. At the beginning of each selection round, the current answer probabilities over $\mathcal{H}_\rho$ are renormalized to initialize the acquisition weights w_i^0, where w_i^k denotes the acquisition importance of hypothesis h_i at selection step k. The discriminative utility of candidate frame f_t is defined as

$$D_k(f_t) = \sum_{i<j} w_{ij}^k \left| s_{t,i} - s_{t,j} \right|, \qquad w_{ij}^k = \frac{w_i^k w_j^k}{\sum_{u<v} w_u^k w_v^k}, \tag{4}$$

where (i, j) denotes a retained hypothesis pair and (u, v) enumerates all unordered pairs in $\mathcal{H}_\rho$ for normalization. The pair weight w_{ij}^k emphasizes competitions involving currently plausible hypotheses, while $|s_{t,i} - s_{t,j}|$ measures how differently frame f_t supports the two hypotheses. To avoid repeatedly selecting temporally nearby evidence, we further combine hypothesis discrimination with temporal diversity:

$$U_k(f_t) = \underbrace{D_k(f_t)}_{\text{hypothesis discrimination}} + \lambda \underbrace{C(f_t, \mathcal{A}_k)}_{\text{temporal diversity}}, \qquad t_k^* = \arg\max_{t:\, f_t \notin \mathcal{A}_k} U_k(f_t), \tag{5}$$

where $\mathcal{A}_k$ contains the frames selected before step k, λ controls the contribution of temporal diversity, and $C(f_t, \mathcal{A}_k)$ is the minimum temporal distance from f_t to the selected frames, normalized by the temporal span of the candidate sequence. The diversity term is omitted when $\mathcal{A}_k$ is empty, and $f_{t_k^*}$ is selected according to the resulting acquisition utility.

Acquisition state update. After selecting frame $f_{t_k^*}$, we update the acquisition state as

$$w_i^{k+1} = \frac{w_i^k \exp(-\alpha s_{t_k^*, i})}{\sum_j w_j^k \exp(-\alpha s_{t_k^*, j})}, \tag{6}$$

where $s_{t_k^*, i}$ is the similarity between the selected frame and hypothesis h_i, α controls the suppression strength, and the denominator normalizes the updated weights over all retained hypotheses. Hypotheses receiving stronger evidence are down-weighted more strongly, reducing the importance of their associated pairs and shifting subsequent selection toward less-covered competitions. The discrimination score is then recomputed using the updated acquisition weights and the cached evidence matrix $\mathbf{S}$, and this *score–select–update* procedure continues until the per-round budget of K keyframes is reached.

Iterative inference and termination. The acquisition state $\mathbf{w}^k$ is updated within each selection round, whereas the MLLM answer distribution is refreshed only after the acquired keyframes are restored to temporal order and passed back to the frozen MLLM together with the query. The updated distribution reconstructs $\mathcal{H}_\rho$ and, when additional evidence is still required, initializes the next round of frame selection. If the updated prediction satisfies the stopping criterion, HypoKFS returns the top-ranked hypothesis; otherwise, the procedure continues for at most M rounds. If the criterion remains unsatisfied after the final round, the hypothesis with the highest probability in the final answer distribution is returned. Neither the MLLM nor the vision–language matching model is updated during inference.

3.4 Complexity

Let r denote the fraction of queries routed to keyframe selection and $\bar{K}_\rho$ the average number of distinct retained hypotheses requiring frame–hypothesis matching across all rounds. For a candidate set of N_c frames, the average matching complexity over all queries is $\mathcal{O}(rN_c\bar{K}_\rho)$. Iterative inference does not require repeated matching for hypotheses that have already appeared, since their corresponding similarities in $\mathbf{S}$ are cached and reused in later rounds; only newly introduced hypotheses require additional matching. Within each round, discrimination scoring, diversity-aware selection, and acquisition-state updates operate entirely on the cached similarities and introduce no additional forward passes through the vision–language matching model. Compared with conventional KFS methods that perform global image–text matching over the candidate set for every query, HypoKFS restricts matching to routed queries and their retained hypotheses, thereby avoiding substantial unnecessary matching computation. The resulting reduction in frame-matching computation is reported in Fig. 3.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation. We evaluate HypoKFS on three long-video understanding benchmarks: Video-MME Fu et al. (2025), LongVideoBench Wu et al. (2024), and MLVU Zhou et al. (2024). These benchmarks contain multiple-choice VideoQA questions covering diverse video durations and reasoning types. Following the standard evaluation protocols, we report multiple-choice accuracy on all three benchmarks.

Implementation. We evaluate HypoKFS with Qwen3-VL Bai et al. (2025b) and LLaVA-Video Zhang et al. (2024), covering two representative open video MLLMs and enabling direct comparison with existing training-free frame-selection methods such as AKS Tang et al. (2025) and Q-Frame Zhang et al. (2025). All MLLMs remain frozen, and HypoKFS requires no additional training or parameter updates. For the main results, we consider two per-inference frame budgets, 32 and 64. In each setting, the number of uniformly sampled frames used for initial inference is kept identical to the keyframe budget in each subsequent selection round. We allow at most two rounds of iterative frame selection.Candidate frames are sampled at 1 fps, and CLIP ViT-B/32 Radford et al. (2021) serves as the default frozen vision–language encoder for frame–hypothesis matching. Unless otherwise specified, we set the hypothesis coverage threshold $\rho = 0.8$, routing threshold $\tau = 0.8$, suppression coefficient $\alpha = 1.0$, and temporal diversity weight $\lambda = 0.1$. All experiments are conducted on NVIDIA H20-3e GPUs with 141 GB of memory, using the same hardware configuration for efficiency and runtime comparisons.

4.2 Main Results

Table 1 compares HypoKFS with uniform sampling and existing training-free keyframe selection methods across three long-video benchmarks. HypoKFS consistently improves both MLLM backbones on all benchmarks. With Qwen3-VL, HypoKFS improves LongVideoBench from 56.3 to 60.4 ($+4.1$), MLVU from 64.1 to 70.0 ($+5.9$), and Video-MME from 69.1 to 72.9 ($+3.8$). Similar improvements are observed with LLaVA-Video-7B, where the gains reach $+5.9$, $+6.7$, and $+3.5$ points, respectively. The consistent improvements across two different MLLM backbones suggest

Table 1: Performance comparison on LongVideoBench (LVB), MLVU, and Video-MME (V-MME). For HypoKFS, Frames / Budget reports the average total number of processed frames and the per-inference frame budget, respectively.

Model	Frames	Params	LVB	MLVU	V-MME
Video MLLMs					
GPT-5 OpenAI (2025)	–	–	72.6	77.3	81.8
Gemini-2.5-Pro Gemini Team, Google (2025)	–	–	77.6	–	84.8
Qwen2.5-VL-72B Bai et al. (2025a)	256	72B	60.7	74.6	73.3
InternVL3-8B Zhu et al. (2025)	32	8B	62.3	71.4	66.5
ViLAMP Cheng et al. (2025)	1fps	7B	67.5	–	61.2
Keye-VL-1.5 Yang et al. (2025a)	64	8B	73.0	–	66.0
Training-Free Plug-and-Play Frame Selection					
Qwen3-VL Bai et al. (2025b)	128	8B	56.3	64.1	69.1
+ AKS Tang et al. (2025)	128	8B	58.4	65.4	69.3
+ Q-Frame Zhang et al. (2025)	128	8B	59.1	66.6	70.3
+ FOCUS Zhu et al. (2026)	128	8B	59.7	67.2	70.7
+ **HypoKFS**	82.3/64	8B	**60.4** ↑ 4.1	**70.0** ↑ 5.9	**72.9** ↑ 3.8
LLaVA-Video-7B Zhang et al. (2024)	64	7B	58.9	71.1	64.4
+ AKS Tang et al. (2025)	64	7B	62.5	71.2	65.1
+ Q-Frame Zhang et al. (2025)	64	7B	63.0	72.2	65.7
+ FOCUS Zhu et al. (2026)	64	7B	63.5	72.7	65.4
+ **HypoKFS**	48.2 / 32	7B	**64.8** ↑ 5.9	**77.8** ↑ 6.7	**67.9** ↑ 3.5

that the benefit of HypoKFS does not depend on a particular model architecture or visual input budget. HypoKFS also compares favorably with existing plug-and-play frame selectors under the same backbone. On Video-MME with Qwen3-VL, HypoKFS achieves 72.9, compared with 69.3 for AKS and 70.3 for Q-Frame. Under LLaVA-Video-7B, HypoKFS reaches 64.8 on LongVideoBench and 67.9 on Video-MME, consistently improving over the corresponding uniform-sampling baseline and existing frame selectors. These results indicate that the advantage does not arise merely from replacing uniform sampling with a generic relevance-based retrieval stage. Instead, conditioning evidence acquisition on the hypothesis space produced by initial inference provides more decision-relevant visual evidence for resolving the remaining answer ambiguity.

4.3 Analysis of hypothesis-guided Selective Retrieval

Fig. 3 analyzes HypoKFS from three complementary perspectives: retrieval efficiency, hypothesis concentration, and selective computation. As shown in Fig. 3(a), HypoKFS achieves 72.9% accuracy with only a 1.3% matching ratio, providing a favorable accuracy–computation trade-off compared with conventional keyframe selectors. This efficiency stems from selective activation: queries with concentrated initial answer distribution bypass candidate-frame matching, whereas additional visual computation is reserved for queries that remain ambiguous.

Fig. 3(b) examines whether the initial answer distribution provides a sufficiently informative search space for subsequent evidence acquisition. The cumulative ground-truth coverage increases rapidly with the number of retained hypotheses, with the top-2 hypotheses already covering 89.4% of correct answers. This observation supports the hypothesis-space construction in Eq. 2: HypoKFS can focus visual acquisition on a compact set of plausible alternatives and target evidence that distinguishes competing hypotheses rather than evidence that is merely relevant to the query.

Fig. 3(c) further studies *when* additional evidence should be acquired by varying the routing threshold τ. Increasing τ activates retrieval more aggressively and consequently raises the matching ratio. Accuracy improves steadily as τ increases toward 0.8. Beyond $\tau = 0.8$, however, the matching ratio continues to increase while accuracy improves only marginally, revealing diminishing returns from additional retrieval. We therefore use $\tau = 0.8$ as the default operating point, achieving 72.9% accuracy with a 1.3% matching ratio.

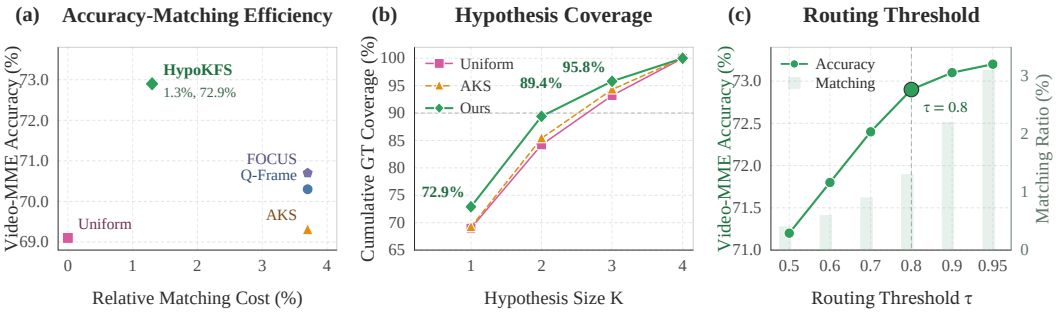

Figure 3: Analysis of hypothesis-guided selective retrieval on Video-MME. **(a)** Accuracy–retrieval efficiency. **(b)** Cumulative top-K hypothesis coverage. **(c)** Accuracy–computation trade-off under different thresholds τ.

Table 2: Ablation studies on Video-MME. **(a)** Component ablation across video durations. **(b)** Hyperparameter sensitivity to α and λ under keyframe budgets $K = 64/128/256$, where each cell reports the corresponding accuracies in this order.

(a) Component Ablation

Components			Video-MME		
Route	Discrim.	Update	Short	Medium	Long
–	–	–	78.6	68.7	60.0
✓	–	–	79.5	69.5	60.8
–	✓	–	80.2	70.4	61.8
✓	✓	–	81.2	71.4	62.9
✓	✓	✓	**82.3**	**72.5**	**63.9**

(b) Hyperparameter Sensitivity

α \ λ	0.0	0.1	0.2
0.0	72.0/72.3/72.5	72.2/72.5/72.7	72.0/72.3/72.5
0.5	72.3/72.6/72.8	72.6/72.9/73.1	72.4/72.7/72.9
1.0	72.5/72.8/73.0	**72.9/73.3/73.5**	72.6/73.0/73.2
1.5	72.4/72.7/72.9	72.7/73.1/73.3	72.5/72.8/73.0
2.0	72.1/72.4/72.6	72.4/72.7/72.9	72.2/72.5/72.7

4.4 Ablation Studies

Table 2(a) disentangles the contributions of the three main components of HypoKFS. Selective routing determines *when to retrieve* by bypassing candidate-frame matching when the initial answer distribution is sufficiently concentrated. Introducing routing alone already improves performance across all video durations, showing that the initial MLLM inference provides a useful signal for computation allocation. Hypothesis discrimination instead determines *what to retrieve* by prioritizing frames that separate plausible hypotheses rather than frames with generic query relevance. It provides larger gains than routing alone, particularly on medium and long videos, where the larger temporal search space makes evidence localization more challenging. Combining routing and hypothesis discrimination further improves the results to 81.2, 71.4, and 62.9 on short, medium, and long videos, respectively, demonstrating that selective computation and hypothesis-guided evidence acquisition provide complementary benefits. Finally, incorporating the acquisition-state update yields the full HypoKFS performance of 82.3, 72.5, and 63.9. By progressively down-weighting hypotheses already supported by selected evidence, the update redirects subsequent acquisition toward unresolved competitions and reduces repeated selection of redundant evidence.

Table 2(b) studies the sensitivity to the keyframe budget K, suppression coefficient α, and temporal diversity weight λ. Increasing K from 64 to 128 consistently improves accuracy across different (α, λ) settings, while increasing the budget further to 256 yields smaller additional gains. This trend indicates diminishing returns beyond $K = 64$. For the acquisition-state update, $\alpha = 0$ disables evidence-dependent suppression and consistently underperforms intermediate values, confirming the benefit of redirecting acquisition after each selected frame. Meanwhile, a moderate temporal diversity weight improves performance over $\lambda = 0$, whereas a larger value can overemphasize temporal coverage at the expense of hypothesis discrimination.

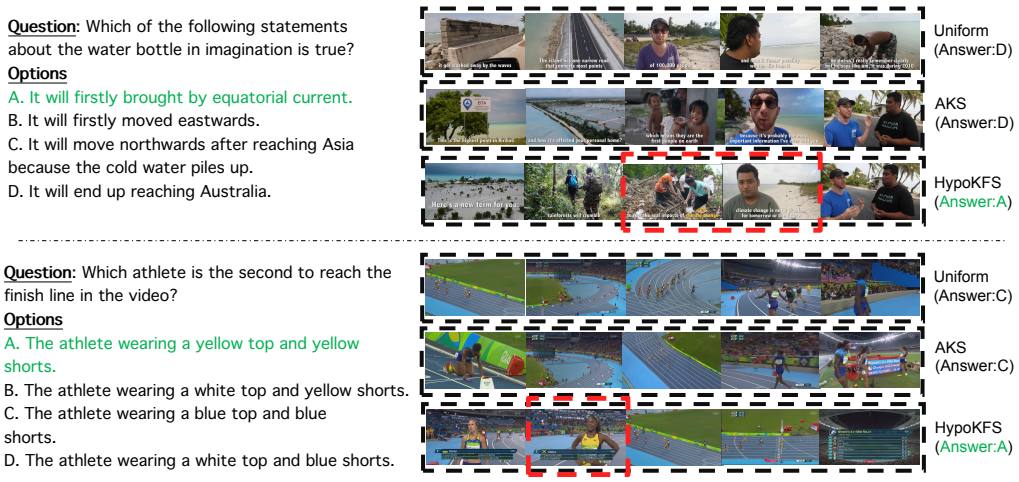

Figure 4: **Qualitative comparison on Video-MME.** Uniform sampling may miss sparse evidence, while relevance-based retrieval can favor query-related yet non-discriminative frames. HypoKFS instead retrieves evidence that resolves competing answer hypotheses.

4.5 Qualitative Analysis

Fig. 4 presents two representative examples from Video-MME. In the first example, answering the question requires identifying the visual evidence that distinguishes different descriptions of the water bottle's trajectory. Uniform sampling and AKS both predict option D, whereas HypoKFS retrieves frames that better distinguish the competing hypotheses and correctly selects option A. In the second example, the model must identify the second athlete reaching the finish line from fine-grained appearance cues. Uniform sampling and AKS predict option C, while HypoKFS focuses on frames in which the relevant athlete can be distinguished and correctly selects option A. These examples illustrate that query relevance alone may retrieve visually related frames without resolving the answer ambiguity, whereas HypoKFS explicitly targets evidence that discriminates among the current hypotheses.

5 Conclusion

Summary. We presented **HypoKFS**, a training-free and plug-and-play framework for hypothesis-guided iterative keyframe selection. HypoKFS first reasons over sparsely sampled frames to construct a compact hypothesis space, which determines whether additional visual evidence is required and guides subsequent frame selection. For ambiguous queries, discriminative frame scoring and iterative hypothesis refinement focus computation on unresolved alternatives, while confident predictions bypass unnecessary global frame matching. Experiments on Video-MME, LongVideoBench, and MLVU demonstrate consistent accuracy improvements across different MLLM backbones while substantially reducing frame–hypothesis matching computation.

Limitations and Outlook. The current formulation is naturally suited to tasks where a compact set of answer hypotheses can be reliably constructed, as in multiple-choice VideoQA. In principle, the same framework can be extended to open-ended settings by constructing hypotheses from model-generated candidate responses, although the quality and stability of this hypothesis construction become more important in such cases. HypoKFS also relies on an additional frozen vision–language encoder for frame–hypothesis matching, despite substantially reducing the amount of matching required through selective routing and cross-round reuse. Finally, subsequent frame selection depends on the quality of the current hypothesis space, and hypotheses assigned very low probability in early inference may receive limited attention. Future work may explore more flexible hypothesis construction and recovery mechanisms for broader long-video reasoning tasks.

6 AI USE STATEMENT

In this work, we used generative AI tools to assist with literature discovery, research ideation, implementation discussion, experimental planning, and writing and editing of the manuscript. Generative AI tools were also used to assist with mathematical derivations and presentation of the method. All AI-assisted content, references, methodological suggestions, implementations, and derivations were independently reviewed and verified by the authors. All reported experimental results were produced and verified by the authors. We take full responsibility for the final content of this work.

REFERENCES

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

Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint*, 2025b.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, and Zhijian Liu. Longvila: Scaling long-context visual language models for long videos. 2024.

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

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

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind, 2025.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13504–13514, 2024.

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

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

Pengyi Li, Irina Abdullaeva, Alexander Gambashidze, Andrey Kuznetsov, and Ivan Oseledets. Maxinfo: A training-free key-frame selection method using maximum volume for enhanced video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7198–7207, 2026b.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3318–3327, 2025.

OpenAI. Gpt-5 system card. OpenAI, 2025.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024.

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

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tianbo Ye, Yang Lu, and Jenq Neng Hwang. Moviechat: From dense token to sparse memory for long video understanding. *ArXiv*, abs/2307.16449, 2023.

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

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

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18992–19001, 2025.

Yilin Wang, Xiangxi Zheng, Dongxing Mao, Linjie Li, Zhengyuan Yang, Ping Yu, Rui Yan, Yuan Yao, and Alex Jinpeng Wang. Efficient frame selection for long videos at test time with attention-based mllm selectors. *arXiv preprint arXiv:2607.15689*, 2026.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding vialarge language models. In *European Conference on Computer Vision*, 2025.

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

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1278–1287, 2019.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025a.

Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, and Jifeng Dai. Pvc: Progressive visual token compression for unified image and video processing in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24939–24949, 2025b.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. Voco-llama: Towards vision compression with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29836–29846, 2025.

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

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

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongqiang Xiong, Bo Zhang, Tiejun Huang, Zheng Liu, et al. Mlvu: Benchmarking multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *The Fourteenth International Conference on Learning Representations*, 2026.

A Appendix

A.1 Implementation Details

Multiple-Choice Inference and Hypothesis Construction. For multiple-choice VideoQA, we instruct the MLLM to predict the option label corresponding to the best answer. At the first decoding position, we extract the logits associated with the candidate answer labels (e.g., A, B, C, and D) and apply a softmax only over these candidate-label logits. This produces the answer distribution $\mathbf{p} = [p_1, \ldots, p_C]$ used throughout HypoKFS. Therefore, p_i denotes the normalized first-token probability of the option label corresponding to hypothesis h_i, rather than the sequence probability of the complete answer.

The option label and its semantic content are used for different purposes. The MLLM answer distribution is constructed from the first-token logits of the option labels, whereas frame–hypothesis matching uses the complete semantic content of each answer option. Specifically, for hypothesis h_i, we concatenate the question q with the textual content of its corresponding answer option to form $q \oplus h_i$. The resulting hypothesis-conditioned text is matched against candidate video frames using the frozen vision–language encoder. This preserves the standard multiple-choice output format for MLLM inference while providing semantically informative text for frame–hypothesis matching.

Video-MME Evaluation Prompt. For Video-MME Fu et al. (2025), we follow the official multiple-choice evaluation format. The instruction used for inference is:

```
Select the best answer to the following multiple-choice
question
based on the video.  Respond with only the letter (A, B, C,
or D)
of the correct option.
Question:  [Question]
A. [Option A]
B. [Option B]
C. [Option C]
D. [Option D]
The best answer is:
```

The same answer format is used for both initial inference and renewed inference in subsequent rounds, allowing the candidate-label logits to be extracted consistently throughout the iterative process.

Inference Procedure. HypoKFS begins by uniformly sampling a sparse set of frames and performing an initial forward pass through the frozen MLLM. The first-token logits of the candidate answer labels are normalized to construct the current answer distribution. If its maximum probability reaches the routing threshold τ, the highest-probability hypothesis is returned without additional

Table 3: Scaling performance under different frame settings on Video-MME using Qwen3-VL.

Model	Frames	Params	Video-MME
Qwen3-VL	128	8B	69.1
+ HypoKFS	128	8B	**72.9**
Qwen3-VL	256	8B	70.2
+ HypoKFS	256	8B	**73.8**
Qwen3-VL	1 fps	8B	70.8
+ HypoKFS	1 fps	8B	**74.3**

frame matching. Otherwise, the smallest probability-ranked hypothesis set whose cumulative probability reaches ρ is retained for subsequent keyframe selection.

For a routed query, candidate frames are matched against the retained hypothesis-conditioned texts. Each frame–hypothesis similarity is computed when the corresponding hypothesis first participates in selection and is cached for later reuse. Within each selection round, HypoKFS repeatedly scores the remaining candidate frames, selects the frame with the largest acquisition utility, and updates the acquisition state. After the per-round frame budget is reached, the selected frames are restored to temporal order and passed to the same frozen MLLM for renewed inference. The updated first-token answer distribution is then used to reconstruct the hypothesis space and determine whether another selection round is required. In the main experiments, we perform at most two keyframe-selection rounds. If the routing criterion remains unsatisfied after the final round, the hypothesis with the highest probability in the final answer distribution is returned.

A.2 Frame-Budget Scaling

We further examine whether the improvement of HypoKFS persists as the available visual input increases. We evaluate Qwen3-VL under three frame settings on Video-MME: 128 uniformly sampled frames, 256 uniformly sampled frames, and dense sampling at 1 fps. The results are reported in Table 3.

Increasing the frame density improves the Qwen3-VL baseline from 69.1 with 128 frames to 70.8 at 1 fps, confirming that denser temporal coverage can recover additional visual information. HypoKFS consistently improves all three settings, with gains of $+3.8$, $+3.6$, and $+3.5$ percentage points at 128 frames, 256 frames, and 1 fps, respectively. The improvement remains stable as the frame density increases, indicating that HypoKFS is not restricted to low-frame settings. Instead, hypothesis-guided frame selection remains complementary to increasing the visual input budget by prioritizing frames that better distinguish the remaining answer hypotheses.

A.3 Analysis of the Selection Mechanism

Why Routed Queries Retain Competing Hypotheses. The routing threshold τ and hypothesis-coverage threshold ρ serve different purposes. A query enters keyframe selection only when its maximum answer probability is smaller than τ. Since we use $\rho \geq \tau$, a single hypothesis cannot reach the required cumulative probability mass for a routed query:

$$\max_i p_i < \tau \leq \rho. \tag{7}$$

It follows that the smallest hypothesis set whose cumulative probability reaches ρ must contain at least two hypotheses, i.e., $K_\rho \geq 2$. Pairwise discrimination is therefore well defined whenever keyframe selection is activated.

Interpretation of Pairwise Competition Weights. Recall that the pair weight at selection step k is

$$w_{ij}^k = \frac{w_i^k w_j^k}{\sum_{u<v} w_u^k w_v^k}. \tag{8}$$

Since the acquisition weights are non-negative, the pair weights are also non-negative and satisfy $\sum_{i<j} w_{ij}^k = 1$. The discrimination score

$$D_k(f_t) = \sum_{i<j} w_{ij}^k \, |s_{t,i} - s_{t,j}| \tag{9}$$

can consequently be interpreted as a weighted average of the pairwise separation induced by candidate frame f_t. A frame receives a large score when it provides substantially different support for hypothesis pairs that currently carry large acquisition weights. Conversely, a frame that assigns similar scores to competing hypotheses contributes little to resolving the current ambiguity, even when it is broadly relevant to the question.

Effect of the Acquisition-State Update. After selecting frame $f_{t_k^*}$, the acquisition state is updated according to

$$w_i^{k+1} = \frac{w_i^k \exp(-\alpha s_{t_k^*,i})}{\sum_j w_j^k \exp(-\alpha s_{t_k^*,j})}. \tag{10}$$

For any two retained hypotheses h_i and h_j, their relative acquisition weight after the update satisfies

$$\frac{w_i^{k+1}}{w_j^{k+1}} = \frac{w_i^k}{w_j^k} \exp\left(-\alpha \left[s_{t_k^*,i} - s_{t_k^*,j}\right]\right). \tag{11}$$

If the selected frame has a larger similarity to h_i than to h_j, the relative acquisition weight assigned to h_i decreases. The update therefore reduces repeated emphasis on hypotheses that have already received stronger visual support and redirects subsequent frame selection toward less-covered hypothesis pairs. Importantly, these acquisition weights are used only within the frame-selection process and do not modify the MLLM answer distribution.

Cross-Round Similarity Reuse. Frame–hypothesis similarities are cached across inference rounds. Once the similarities between a retained hypothesis and all candidate frames have been computed, they are reused whenever that hypothesis remains active in a later round. If renewed inference introduces a previously unseen hypothesis, only the similarities associated with that hypothesis need to be computed. Discrimination scoring, temporal-diversity scoring, and acquisition-state updates operate directly on the cached similarities and therefore require no additional forward passes through the frozen vision–language matching model.