BROAD COVERAGE, FOCUSED RESOLUTION: HYPOTHESIS-DISCRIMINATIVE EVIDENCE ACQUISITION FOR LONG VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long video question answering requires selecting evidence that determines the answer under a limited visual budget. Uniform sampling can miss brief events, while existing selection methods mainly organize evidence by query relevance and temporal coverage. This paper focuses on selecting a compact and complementary evidence set based on the distinctions among candidate answers. We propose HDEA, a two stage evidence acquisition framework driven by competition among answer candidates. Broad Competition Coverage considers all candidate pairs, selects complementary discriminative evidence, and builds an evidence core through temporally constrained submodular coverage. Focused Competition Resolution uses the core prediction to reweight candidate pairs, selects supplementary evidence while preserving the core, and fuses the predictions from the core and two nested expanded visual contexts. The two stages follow one principle: broad coverage before answering and focused resolution after answering. HDEA requires no additional training, subtitles, or external language models. On VideoMME, LongVideoBench, and MLVU, HDEA reaches 67.44%, 65.22%, and 74.25%, respectively, and achieves competitive performance among representative long video methods. Across four frozen video MLLMs and three benchmarks, HDEA demonstrates strong cross model transferability, outperforming uniform sampling in all 12 model and benchmark combinations. These results support candidate answer competition as a common structure for evidence acquisition and refinement in long video question answering. Code is available at `https://anonymous.4open.science/r/HDEA`.

1 INTRODUCTION

Long video question answering requires multimodal large language models (MLLMs) to identify relevant visual evidence from videos that may last several hours. Dense frame processing quickly exceeds practical context and computation budgets, while uniform sampling can miss brief events. Recent methods improve evidence acquisition through question-aware temporal coverage (Tang et al., 2025), relevance and diversity modeling (Sun et al., 2025), multiresolution frame selection (Zhang et al., 2025a), and progressive or query-adaptive search (Zhu et al., 2026; Li et al., 2026a). Constructing a compact visual context that is sufficient for answering has therefore become a central problem in long video question answering.

Multiple choice question answering provides a natural decision structure for evidence acquisition. Each candidate answer gives one possible interpretation of the video. Different candidates form a set of competitions that must be resolved. Local video evidence often serves different discriminative roles. One event may distinguish two actions, while another event may distinguish a different pair of states or behaviors. Complementarity among local evidence is therefore important when the context capacity is limited. A useful visual context should contain evidence that covers different discrimination needs among the candidate answers.

Figure 1 examines set complementarity through normalized retrieval discrimination coverage. For each question, we measure the fraction of the full candidate clip set's discriminative score retained by the selected clips. This diagnostic removes temporal constraints to compare set based selection

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.

with independent clip ranking; its complete definition is provided in Appendix D.3. Pairwise Top K ranks clips by their individual discrimination scores, while Broad Coverage selects each new clip according to its marginal contribution to the current set. The two strategies select the same clip at $K = 1$. As the budget increases, Broad Coverage accumulates complementary pair discrimination more rapidly. At $K = 4$, it retains 99.1%, 97.6%, and 99.0% of the full candidate set's score on VideoMME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2025), respectively, giving an equal benchmark average of 98.6%.

This set structure motivates different evidence priorities before and after the initial prediction. Before the first inference, all candidate pairs receive equal priority, and the initial visual budget is used to cover their discrimination needs. After the first inference, the core prediction provides a distribution over candidate answers. This distribution guides supplementary evidence acquisition by assigning different priorities to candidate pairs. We summarize this process with one principle: **broad coverage before answering, focused resolution after answering.**

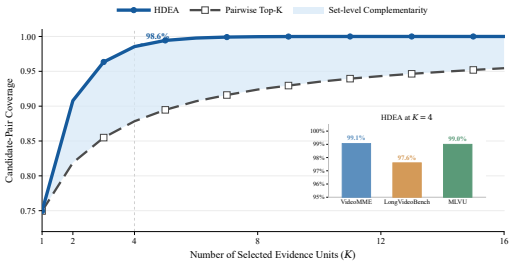


Figure 1: **Set complementarity in retrieval discrimination coverage.** Broad Coverage achieves higher normalized coverage than Pairwise Top K as the number of selected clips increases.

Following this principle, we propose HDEA (Hypothesis-Discriminative Evidence Acquisition). **Broad Competition Coverage** selects complementary clips across candidate pairs, constructs an evidence core, and performs the first inference. **Focused Competition Resolution** reweights candidate pairs using the core prediction, selects supplementary evidence outside the core, and constructs nested expanded visual contexts. Predictions from the core and the expanded contexts are then combined to produce the final answer.

The two stages share the same candidate pair representation and frame level retrieval scores. Broad Competition Coverage maintains the discrimination already covered for each pair and selects new clips by their marginal contribution. The resulting objective is monotone submodular and admits a constant factor greedy approximation under fixed temporal constraints (Fisher et al., 1978). Focused Competition Resolution uses the core probabilities to rank supplementary evidence according to the current priorities of candidate pairs.

We evaluate HDEA on VideoMME, LongVideoBench, and MLVU, obtaining 67.44%, 65.22%, and 74.25%, respectively. We use Qwen3-VL (Bai et al., 2025) as the primary answer model and further evaluate InternVL3 (Zhu et al., 2025), mPLUG-Owl3 (Ye et al., 2025), and LLaVA-Video (Zhang et al., 2025b). Our contributions are as follows.

1. We propose Broad Competition Coverage, which represents local evidence in a candidate pair space and selects complementary clips through temporally constrained submodular coverage.
2. We propose Focused Competition Resolution, which uses the core prediction to reweight candidate pairs and acquire supplementary evidence while preserving the established evidence core.
3. We evaluate HDEA on three long video benchmarks and four frozen video MLLMs. HDEA achieves competitive performance among the compared methods across all benchmarks and shows strong cross model transferability across the 12 model and benchmark combinations.

2 Related Work

2.1 Long Video Context Modeling and Compression

Long videos produce visual sequences that grow rapidly with duration, which increases context length and inference cost. Existing methods address this problem through memory, context extension, multi rate temporal modeling, and visual token compression. MovieChat compresses con-

tinuous visual information with short term and long term memory (Song et al., 2024). MA-LMM stores previous visual content in an external memory bank for streaming long video processing (He et al., 2024). TimeChat introduces time aware visual encoding to improve temporal understanding and reasoning in long videos (Ren et al., 2024). LongVA transfers the long context capability of a language model to visual sequences and extends the amount of visual context that the model can process (Zhang et al., 2024). SF-LLaVA-1.5 combines slow and fast visual streams with joint image and video training for token efficient long video understanding (Xu et al., 2025). LongVU uses frame redundancy and text relevance for adaptive spatial and temporal compression (Shen et al., 2025). VideoChat-Flash uses hierarchical token aggregation to reduce the long video context at multiple temporal scales (Li et al., 2026b). These approaches mainly improve context capacity or reduce visual redundancy, without explicitly organizing evidence around answer distinctions.

2.2 Long Video Frame Selection

Frame selection methods retain frames or clips that are useful for the current question under a fixed visual budget. Frame-Voyager learns to select frame combinations that are useful for downstream question answering (Yu et al., 2025). AKS jointly considers query relevance and temporal coverage and performs adaptive sampling without additional training (Tang et al., 2025). Q-Frame selects video frames according to the query and adjusts their visual resolution (Zhang et al., 2025a). MDP^3 combines relevance, diversity, and temporal order and uses a determinantal point process to model the selected frame set (Sun et al., 2025). FOCUS formulates keyframe selection as a combinatorial pure exploration problem to identify informative temporal regions (Zhu et al., 2026). DIG changes its frame acquisition strategy for global and localized queries (Li et al., 2026a). VQOS further uses candidate answer information during video filtering and combines similarities among the video, query, and answer options with adaptive sampling and dynamic resolution allocation (Wu et al., 2026). These methods improve selection through relevance, diversity, temporal coverage, or answer conditioned similarity, without modeling complementarity across candidate distinctions.

3 Method

HDEA organizes visual evidence according to competition among candidate answers and builds the visual context in two stages. Broad Competition Coverage selects complementary clips, builds an evidence core, and performs the first inference. Focused Competition Resolution uses the core prediction to select supplementary evidence, expands the visual context while preserving the core, and fuses three predictions. Figure 2 provides an overview of the HDEA framework.

3.1 Visual Evidence Representation

Given a long video V, a question q, and a set of m candidate answers $\mathcal{C} = \{c_1, \ldots, c_m\}$, HDEA selects a set of source frames E as input to a frozen answer model F. Here c_a denotes the text of candidate a, and E_b denotes a frame set with budget b. The frames are presented in their original temporal order.

Two frame clips. We sample the video at one frame per second and obtain the candidate sequence $\mathcal{F}_{\text{cand}} = (f_1, \ldots, f_T)$, where T is the number of candidate frames. When $T > 32$, adjacent candidate frames form nonoverlapping two frame clips:

$$u_i = (f_{2i-1}, f_{2i}), \qquad i = 1, \ldots, N, \qquad N = \left\lfloor \frac{T}{2} \right\rfloor. \tag{1}$$

All clips form $\mathcal{U} = \{u_1, \ldots, u_N\}$. If T is odd, the last unmatched frame is not used to form a two frame clip. When $T \leq 32$, we directly keep all candidate frames as the core. We do not add or duplicate frames.

Matching scores. We use a frozen SigLIP retriever (Zhai et al., 2023) to encode video frames and text. Let $\phi(f_t)$ be the visual representation of frame f_t, $\psi(q)$ be the representation of the question, and $\psi(q, c_a)$ be the representation of the question together with candidate c_a. We compute inner

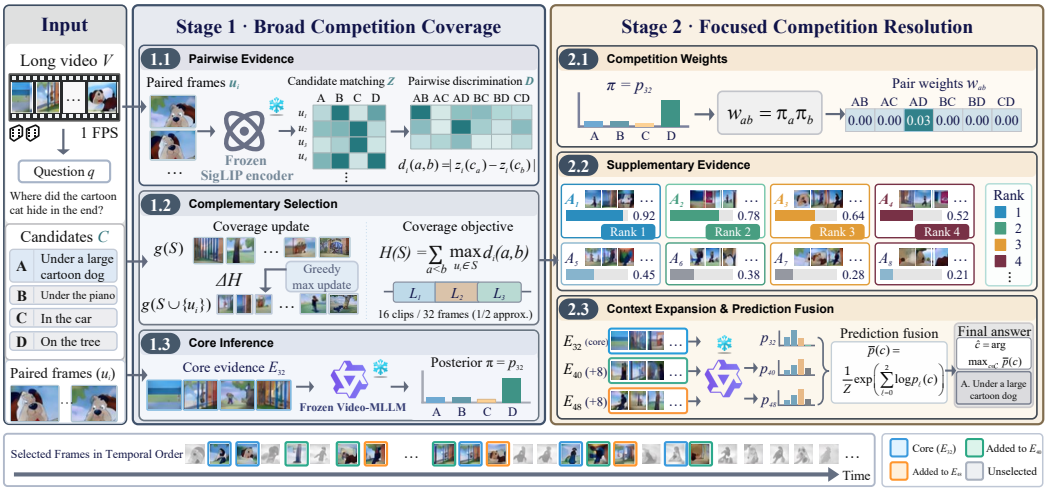


Figure 2: **Overview of HDEA.** Broad Competition Coverage selects complementary evidence in the candidate pair space and builds the evidence core. Focused Competition Resolution uses the core prediction to update candidate pair priorities, adds targeted evidence, and fuses the predictions from E_{32}, E_{40}, and E_{48}.

product matching scores and take the maximum over the two frames in each clip:

$$r_i = \max_{t \in \{2i-1, 2i\}} \langle \phi(f_t), \psi(q) \rangle, \tag{2}$$

$$z_i(c_a) = \max_{t \in \{2i-1, 2i\}} \langle \phi(f_t), \psi(q, c_a) \rangle, \qquad a = 1, \ldots, m. \tag{3}$$

The score r_i measures the relevance between clip u_i and the question and is used for temporal partitioning. The score $z_i(c_a)$ measures how well the clip matches the question conditioned on candidate c_a.

Candidate pair discrimination. For each candidate pair (c_a, c_b), we measure retrieval discrimination by the absolute difference between their candidate conditioned matching scores:

$$d_i(a, b) = |z_i(c_a) - z_i(c_b)|, \qquad 1 \le a < b \le m. \tag{4}$$

The indices a and b refer to candidate answers, and $a < b$ ensures that each unordered pair is counted once. The absolute difference measures the separation between two candidate-conditioned matching scores in the frozen retrieval representation. A larger $d_i(a, b)$ indicates that clip u_i is more discriminative for that candidate pair. We retain the full vector $\mathbf{d}_i = (d_i(a, b))_{a<b}$ so that Broad Competition Coverage can track the discrimination covered for each candidate pair during selection.

3.2 Broad Competition Coverage

Before the first inference, all candidate pairs have equal priority. HDEA uses complementarity among clips to construct an evidence core under temporal region constraints and a global budget.

Complementary coverage. For a selected clip set $S \subseteq \mathcal{U}$, we keep the highest discrimination score for each candidate pair and denote it by $g_{ab}(S)$. The total coverage is

$$g_{ab}(S) = \max_{u_i \in S} d_i(a, b), \tag{5}$$

$$H(S) = \sum_{1 \le a < b \le m} g_{ab}(S), \tag{6}$$

with $g_{ab}(\varnothing) = 0$. A new clip contributes to a candidate pair only when its score exceeds the current coverage. This makes the selection depend on what the current evidence set already contains.

Selection constraints. We recursively partition the video timeline using the query relevance score r_i (Tang et al., 2025). Let $\mathcal{L}_\ell$ be the candidate clips in temporal region ℓ, and let κ_ℓ be the maximum number of clips that can be selected from this region. The core budget is $B_0 = 16$ clips, which corresponds to 32 source frames. If the question contains a valid explicit temporal reference, the corresponding required clips form a set R; otherwise $R = \varnothing$.

The feasible family and the selection objective are

$$\mathcal{I} = \{S \subseteq \mathcal{U} : |S \cap \mathcal{L}_\ell| \le \kappa_\ell \, \forall \ell, \quad |S| \le B_0\}, \tag{7}$$

$$S^\star \in \arg\max_{\substack{S \in \mathcal{I} \\ R \subseteq S}} H(S). \tag{8}$$

The temporal regions and capacities are fixed before selection, and R is required to be feasible. Required clips consume both regional capacity and the global budget.

Starting from $S = R$, each step adds the feasible clip with the largest coverage gain. For clip u_i, the gain is

$$\Delta_H(u_i \mid S) = H(S \cup \{u_i\}) - H(S) \tag{9}$$

$$= \sum_{1 \le a < b \le m} \max\{0, d_i(a,b) - g_{ab}(S)\}. \tag{10}$$

After a clip is added, the pair coverage values are updated. Ties are resolved by higher query relevance and then by earlier time. Selection stops when the budget is reached or no feasible clip remains. If all feasible clips have zero marginal gain, selection continues under the same tie rule, using query relevance and temporal order to allocate the remaining budget.

Theoretical property. As the selected set grows, the best score for each candidate pair does not decrease, while the gain of a fixed new clip cannot increase. Therefore H is a nonnegative monotone submodular function. The regional capacities define a partition matroid. Adding the global budget corresponds to truncation and preserves the matroid structure. After fixing a feasible required set R, the remaining problem is a matroid contraction. The feasible greedy solution S_g satisfies the classical guarantee $H(S_g) \ge \frac{1}{2} H(S^\star)$ (Fisher et al., 1978).

This guarantee applies to the coverage objective H under fixed retrieval scores and temporal constraints. The complete proof is provided in Appendix B. The source frames selected by S_g are merged and sorted by time to form the evidence core $E^{(0)} = E_{32}$. The frozen answer model F performs the first inference on this core, and its candidate probabilities guide the next stage.

3.3 Focused Competition Resolution

The core prediction assigns a probability to every candidate answer. Focused Competition Resolution uses these probabilities to prioritize candidate pairs and rank supplementary evidence outside the core. It preserves the selected core frames, constructs two nested expanded visual contexts, and combines the resulting predictions.

Competition weights. Let π_a be the probability of candidate c_a under the core:

$$\pi_a = p_F(c_a \mid E^{(0)}, q, \mathcal{C}), \qquad \sum_{a=1}^{m} \pi_a = 1. \tag{11}$$

We define the competition weight of a candidate pair as the product of its core probabilities:

$$w_{ab} = \pi_a \pi_b, \qquad 2 \sum_{1 \le a < b \le m} w_{ab} = 1 - \sum_{a=1}^{m} \pi_a^2. \tag{12}$$

The right hand side is the probability that two independent samples from the core prediction give different answers. Thus, w_{ab} gives higher priority to pairs whose candidates both retain probability mass under the core prediction.

Supplementary evidence scoring. From the candidate frames outside the core, we construct eight groups of four frames in temporal order, denoted by $\mathcal{A} = \{A_1, \ldots, A_8\}$. The groups do not overlap with each other or with the core. This construction is used when enough candidate frames remain. The exact construction and fallback rules are given in Appendix A.

For a frame f in a supplementary group, define $z_f(c_a) = \langle \phi(f), \psi(q, c_a) \rangle$. We first compute the absolute matching difference for each candidate pair on every frame, then average within the four frame group. We combine this score with the competition weights:

$$\bar{d}_{A_j}(a, b) = \frac{1}{|A_j|} \sum_{f \in A_j} |z_f(c_a) - z_f(c_b)|, \tag{13}$$

$$v_\pi(A_j) = \sum_{1 \le a < b \le m} w_{ab}\, \bar{d}_{A_j}(a, b). \tag{14}$$

Here $|A_j| = 4$. The competition weights prioritize candidate pairs under the core prediction, while the group scores measure their discrimination in the new evidence. Their weighted sum determines the priority of each supplementary group.

For any group A, extend $\bar{d}_A$ symmetrically and set $\bar{d}_A(a, a) = 0$. If two candidate indices are sampled independently from $\boldsymbol{\pi}$, then

$$\mathbb{E}_{a,b \sim \boldsymbol{\pi}}[\bar{d}_A(a, b)] = 2 v_\pi(A). \tag{15}$$

Therefore, ranking by v_π is equivalent to ranking supplementary groups by their expected discrimination under the core prediction.

Expanded visual contexts. Sort the eligible supplementary groups by decreasing value as $A_{(1)}, A_{(2)}, \ldots$. The expanded visual context at level ℓ contains the core and the first K_ℓ groups:

$$E^{(\ell)} = E^{(0)} \cup \bigcup_{j=1}^{K_\ell} A_{(j)}, \qquad \ell \in \{1, 2\}, \tag{16}$$

$$E^{(0)} \subseteq E^{(1)} \subseteq E^{(2)}. \tag{17}$$

We use $K_1 = 2$ and $K_2 = 4$. Starting from the 32 frame core, these choices add 8 and 16 frames and produce E_{40} and E_{48}. Both expansions use the same ordering obtained from the core prediction. All frames are sorted by their original video time before they are given to the answer model.

Prediction fusion. The same frozen answer model produces a candidate distribution $p_\ell(c) = p_F(c \mid E^{(\ell)}, q, \mathcal{C})$ for $\ell = 0, 1, 2$. The core distribution p_0 is reused from the first inference, so only E_{40} and E_{48} require additional calls. We fuse the three distributions with equal weights in log probability space:

$$\bar{p}(c) = \frac{1}{Z} \exp\left(\frac{1}{3} \sum_{\ell=0}^{2} \log p_\ell(c) \right), \qquad \hat{c} = \arg\max_{c \in \mathcal{C}} \bar{p}(c), \tag{18}$$

where Z is the normalization constant. HDEA returns the candidate with the highest fused probability as the final answer.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate HDEA on three long video multiple choice question answering benchmarks: VideoMME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2025). VideoMME contains short, medium, and long duration groups. LongVideoBench includes videos up to one hour long. For MLVU, we use the official development split and report M-AVG, the unweighted mean accuracy over seven task families. The three benchmarks contain 2,700, 1,337, and 2,174 questions, corresponding to 900, 753, and 1,122 videos.

Table 1: Performance comparison with representative long video methods. Accuracy is reported in percent. LVB denotes LongVideoBench.

Method	Backbone	Frames	VideoMME	LVB	MLVU
GPT-4V (Tech. 2023)	GPT-4V	256	59.90	61.30	–
Gemini-1.5-Flash (Tech. 2024)	Gemini-1.5	256	70.30	61.60	–
Gemini-1.5-Pro (Tech. 2024)	Gemini-1.5	256	**75.00**	64.00	–
Video-LLaVA (EMNLP 2024)	7B	8	39.90	39.10	–
MiniCPM-V 2.6 (arXiv 2024)	8B	64	60.90	54.90	–
InternVL2 (arXiv 2024)	40B	16	61.20	59.70	–
LLaVA-Video-72B (arXiv 2024)	72B	64	70.60	63.90	–
InternVL3 (arXiv 2025)	8B	–	66.30	58.80	71.40
LongVA (arXiv 2024)	Qwen2-7B	128	54.30	–	56.30
LongVU (ICML 2025)	Qwen2-7B	1 FPS	60.60	–	65.40
Frame-Voyager (ICLR 2025)	LLaVA-OV-7B	8	57.50	–	65.60
SF-LLaVA-1.5 (COLM 2025)	SF-LLaVA-7B	128	63.90	62.50	71.50
AKS (CVPR 2025)	LLaVA-Video-7B	64	64.60	62.10	–
Q-Frame (ICCV 2025)	Qwen2-VL-7B	4+8+32	58.30	58.40	65.40
MDP^3 (ICCV 2025)	Ovis2-8B	8	63.90	62.70	73.90
FOCUS (ICLR 2026)	LLaVA-Video-7B	64	65.40	63.50	72.70
DIG (CVPR 2026)	Qwen2.5-VL-7B	32	–	60.43	67.20
VQOS-GO (ICLR 2026)	Qwen2.5-VL-7B	768	66.70	63.40	72.30
HDEA (Ours)	Qwen3-VL-8B	48	67.44	**65.22**	**74.25**

Answer models. Qwen3-VL-8B (Bai et al., 2025) is our primary answer model. We also evaluate InternVL3-8B (Zhu et al., 2025), mPLUG-Owl3-7B (Ye et al., 2025), and LLaVA-Video-7B (Zhang et al., 2025b) to test transfer across models. All our experiments use frozen answer models in a zero shot setting without subtitles. Each model uses its native visual processor. Candidate probabilities are obtained from restricted option token logits. Focused Competition Resolution always uses the core prediction produced by the same answer model.

Implementation. HDEA constructs a 32 frame evidence core, expands it to 40 and 48 frames, and fuses the three predictions. The full pipeline uses at most 48 source frames per call and requires three calls to the frozen answering model. This configuration is fixed across all three benchmarks and four answering models. Detailed budget and inference cost analyses are provided in Appendix G.

4.2 Performance Analysis

Table 1 compares HDEA with representative long video methods. The upper block lists reference video MLLMs. The lower block focuses on long video context construction and frame selection methods. We keep the answer model and visual input setting reported by each method.

HDEA reaches 65.22% on LongVideoBench, the highest value in the full table. On VideoMME and MLVU, it reaches 67.44% and 74.25%. These are the highest values among the long video context construction and frame selection methods listed in the lower block.

4.3 Cross Model Transfer

We keep the HDEA configuration fixed and apply it to four frozen video MLLMs. Figure 3 compares Uniform32, Broad Competition Coverage, and the full HDEA pipeline on all three benchmarks.

Both Broad Competition Coverage and full HDEA outperform their corresponding Uniform32 baselines in all 12 model and benchmark combinations. The HDEA configuration is fixed across models, and the selector requires no retraining. Complete results are provided in Appendix E.

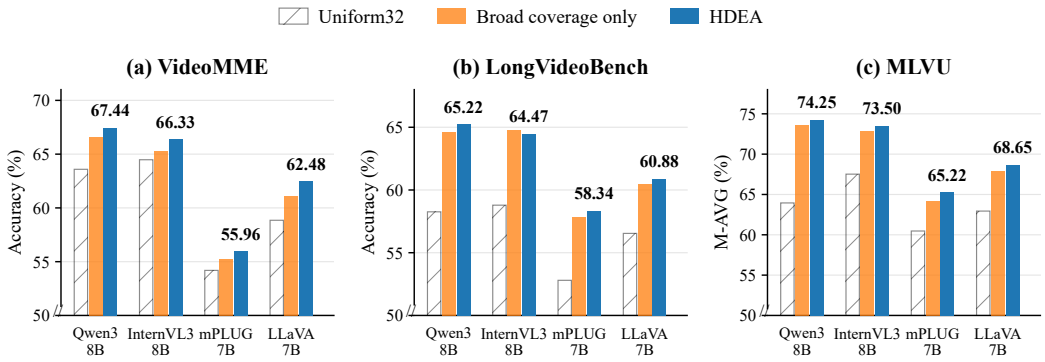


Figure 3: **Cross model transfer on VideoMME, LongVideoBench, and MLVU.** For each answer model, we compare Uniform32, Broad Competition Coverage, and the full HDEA pipeline. Uniform32 and Broad Competition Coverage use one 32 frame answer model call. HDEA uses the complete two stage pipeline.

Table 2: Controlled comparison for Broad Competition Coverage. All methods use 32 frames and one answer model call.

Dataset	Uniform	AKS	OptionPool	Pairwise	Broad Coverage
VideoMME	63.59	62.74	64.81	62.22	**66.56**
LongVideoBench	58.26	61.63	64.17	56.32	**64.62**
MLVU	63.95	70.91	73.23	63.13	**73.62**

4.4 Analysis of the Two Stages

Broad Competition Coverage. We fix Qwen3-VL-8B and compare frame selection strategies with 32 frames and one answer model call. Uniform32 provides the uniform sampling baseline. AKS32 keeps the original relevance and temporal coverage selection rule. We also define two controlled variants. OptionPool32 averages the normalized text representations of all candidate conditioned queries, normalizes the result again, and uses it for clip scoring. PairwisePointwise32 averages the candidate pair discrimination scores into one independent clip score and then performs pointwise selection.

Broad Competition Coverage gives the highest value on all three benchmarks. Compared with PairwisePointwise32, which uses the same candidate pair discrimination scores, Broad Competition Coverage improves accuracy by 4.33, 8.30, and 10.48 percentage points. This result shows that the contribution of a new clip depends on the evidence already selected. Compared with OptionPool32, the gains are 1.74, 0.45, and 0.38 points.

Focused Competition Resolution. Starting from the same evidence core, we compare the core prediction, the predictions from E_{40} and E_{48}, and full HDEA. These comparisons separate the effect of supplementary evidence from the effect of prediction fusion. Uniform48 and one pass Broad Coverage with 48 frames provide matched maximum frame budget references.

With Qwen3-VL-8B, both E_{40} and E_{48} improve over the core prediction on all three benchmarks. Across the 12 model and benchmark combinations in Figure 3, full HDEA improves over Broad Competition Coverage in 11 cases, with an average gain of 0.70 points. The only decrease is 0.30 points for InternVL3-8B on LongVideoBench. The additional effect of prediction fusion varies across models and benchmarks; complete results are provided in Appendix F.

At the same maximum input of 48 frames, E_{48} improves over Uniform48 by 2.00, 4.71, and 7.53 points on VideoMME, LongVideoBench, and MLVU. Compared with Broad Coverage 48, preserving the core and adding targeted evidence gives higher point estimates on VideoMME and

Table 3: Ablation of Focused Competition Resolution and matched 48 frame references with Qwen3-VL-8B. The first four rows share the same evidence core. Calls count all answer model calls needed to produce the reported output.

Configuration	Max frames	Calls	VideoMME	LVB	MLVU
HDEA	48	3	67.44	65.22	74.25
Broad Coverage (E_{32})	32	1	66.56	64.62	73.62
Expanded prediction (E_{40})	40	2	67.44	65.07	73.77
Expanded prediction (E_{48})	48	2	67.52	65.15	74.35
Uniform48	48	1	65.52	60.43	66.83
Broad Coverage 48	48	1	66.63	64.55	74.87

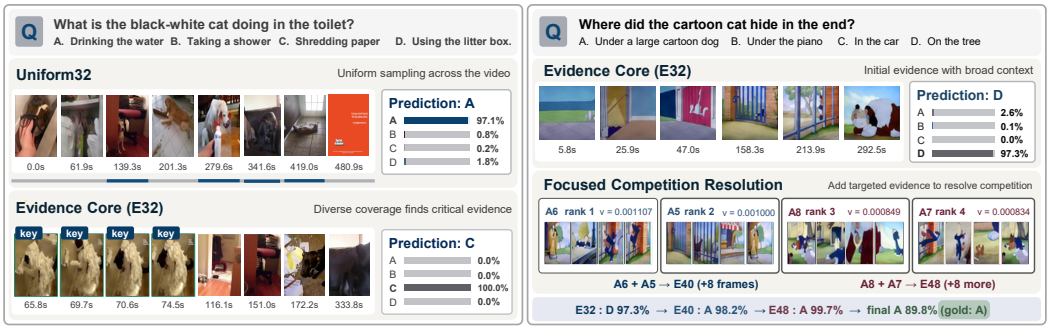


Figure 4: **Qualitative analysis of the two stage evidence acquisition process.** Broad Competition Coverage recovers a brief event missed by uniform sampling, while Focused Competition Resolution adds targeted evidence that corrects the core prediction.

LongVideoBench and a lower point estimate on MLVU. Duration and task level breakdowns are provided in Appendix H.

Qualitative analysis. Figure 4 shows two real examples. In the VideoMME example, Uniform32 predicts A with 97.1% probability. Broad Competition Coverage captures the paper shredding event around 65.8 to 74.5 seconds, which uniform sampling misses, and changes the E_{32} prediction to the correct answer C. In the MLVU example, E_{32} first predicts D with 97.3% probability. The A–D pair then receives the highest competition priority. Focused Competition Resolution selects supplementary groups A6, A5, A8, and A7. The support for the correct answer A rises to 98.2% in E_{40} and 99.7% in E_{48}. The two examples illustrate evidence complementarity in the first stage and targeted supplementary evidence in the second stage. Detailed candidate probabilities and supplementary group rankings are provided in Appendix I.

5 Conclusion

We presented HDEA, a two stage long video evidence acquisition framework driven by competition among candidate answers. Broad Competition Coverage considers all candidate pairs and uses temporally constrained submodular coverage to select complementary evidence and build an evidence core. Focused Competition Resolution uses the core prediction to update the priority of candidate competitions, adds targeted evidence while preserving the core, and fuses predictions from nested visual contexts. The two stages form one process: broad coverage before answering and focused resolution after answering. HDEA requires no additional training, subtitles, or external language models. Across VideoMME, LongVideoBench, and MLVU and four frozen video MLLMs, HDEA improves over the corresponding uniform sampling baseline in all 12 model and benchmark combinations. Competition structure analysis, matched budget controls, budget sweeps, and qualitative examples further show the roles of complementary core construction and focused evidence acquisition. These results support candidate answer competition as a common structure for evidence acquisition and refinement under a limited visual budget.

AI USE STATEMENT

Generative AI tools were used for language translation and editing, figure layout assistance, assistance with mathematical derivations, and feedback on the authors' methodology, experiments, and interpretation of results. The research ideas, method design, experimental design, and scientific conclusions were developed by the authors. All AI-assisted outputs were reviewed by the authors. The authors take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses publicly available long video question answering benchmarks and existing multimodal models. We do not collect new human subject data or introduce new data collection procedures. All benchmark data are used under their released research conditions. Our method operates only at inference time by selecting visual evidence for existing answering models.

REPRODUCIBILITY STATEMENT

We provide the complete method formulation and theoretical analysis in Section 3 and Appendix B. Implementation details and evaluation protocols are provided in Appendices A and C. Anonymous source code is publicly available at `https://anonymous.4open.science/r/HDEA`. All models and benchmarks used in our experiments are publicly available, and additional experimental results and reproducibility records are included in the supplementary appendices.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Marshall L. Fisher, George L. Nemhauser, and Laurence A. Wolsey. An analysis of approximations for maximizing submodular set functions—ii. In *Polyhedral Combinatorics*, volume 8 of *Mathematical Programming Studies*, pp. 73–87. 1978.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. 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.

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.

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.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, 2026b.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra.

Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *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, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp^3: 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025.

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

Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. SlowFast-LLaVA-1.5: A family of token-efficient video large language models for long-form video understanding. In *Conference on Language Modeling*, 2025.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, 2025.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11941–11952, 2023.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 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*, pp. 22056–22065, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning with Synthetic Data. *Transactions on Machine Learning Research*, 2025b.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, 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 *International Conference on Learning Representations*, 2026.

A Implementation of Two Stage Evidence Acquisition

A.1 From Video Frames to Discriminative Evidence

Given a video V, a question q, and a candidate answer set $\mathcal{C} = \{c_1, \ldots, c_m\}$, we first sample the video at approximately one frame per second (1 FPS), producing a temporally ordered candidate frame sequence $\mathcal{F}_{\text{cand}} = (f_1, \ldots, f_T)$. Here, T is the number of candidate frames, m is the number of candidate answers, and c_a denotes the text of candidate answer indexed by a. Each candidate frame corresponds to a source frame in the original video.

A frozen SigLIP-So400M patch14-384 model encodes images and text separately. Video features are cached per video, while question features and candidate conditioned text features are computed per question. Let $\phi(f_t)$ denote the visual encoding, $\psi(q)$ the question encoding, and $\psi(q, c_a)$ the joint encoding of the question and candidate answer. The corresponding frame level matching scores are $r_t^f = \langle \phi(f_t), \psi(q) \rangle$ and $z_t^f(c_a) = \langle \phi(f_t), \psi(q, c_a) \rangle$. Both stages reuse these frame level scores.

When $T > 32$, adjacent candidate frames are grouped into two frame clips $u_i = (f_{2i-1}, f_{2i})$, yielding $N = \lfloor T/2 \rfloor$ nonoverlapping clips collected in $\mathcal{U}$. Within each clip, we take the maximum matching value for the question and for each candidate answer, then compute the absolute difference between candidate answers:

$$\begin{aligned} r_i &= \max\{r_{2i-1}^f, r_{2i}^f\}, \\ z_i(c_a) &= \max\{z_{2i-1}^f(c_a), z_{2i}^f(c_a)\}, \\ d_i(a,b) &= |z_i(c_a) - z_i(c_b)|, \qquad 1 \le a < b \le m. \end{aligned} \tag{A.1}$$

r_i is used to partition temporal regions, and $d_i(a, b)$ measures the matching difference between candidate answers c_a and c_b in the retrieval representation of the clip. The condition $a < b$ ensures that each unordered answer pair is counted once. Scores for all answer pairs form $\mathbf{d}_i = (d_i(a,b))_{a<b}$. For a four choice question, this vector contains six components corresponding to $(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)$.

Maximum pooling is applied independently to each candidate answer, so the maximum matching values for two candidate answers may come from different frames within the same clip. The implementation retains frame wise encodings, pools matching scores first, and then computes answer pair differences. If T is odd, the final unpaired frame is excluded from the two frame clips but remains available for later evidence supplementation. When $T \le 32$, all candidate frames directly form the evidence core, clip selection is skipped, and no frames are added or duplicated.

A.2 Complementary Selection in Broad Competition Coverage

Broad Competition Coverage considers all candidate answer pairs and selects clips that complement the discriminative information already covered by the current set. We recursively partition the timeline according to question relevance r_i, producing disjoint temporal regions. Let $\mathcal{L}_\ell$ denote the candidate clip set in region ℓ, and let κ_ℓ denote the maximum number of clips that may be selected from that region. The regions and capacities are fixed before selection. The total core budget is $B_0 = 16$ clips, corresponding to 32 distinct source frames. Valid temporal references in the question define a required clip set R; when no such requirement exists, $R = \varnothing$. The set R must satisfy the feasibility constraints in Eq. equation 8 of the main paper and counts toward both regional capacities and the total budget.

The algorithm starts from $S = R$ and maintains the current coverage value $g_{ab}(S) = \max_{u_i \in S} d_i(a,b)$ for each answer pair, with zero coverage for the empty set. At each iteration, the coverage gain of every new clip is computed according to Eq. equation 10 in the main paper. The feasible clip with the largest gain is added, after which the coverage values for all answer pairs are updated. Ties in gain are resolved by higher question relevance and then by earlier temporal position. Selection continues until the budget is reached or no feasible clip remains. A zero maximum gain does not terminate selection before either condition is met.

The selected clips are expanded into frames and sorted by their original video timestamps to form the evidence core $E^{(0)} = E_{32}$. Here, E_b denotes a visual context with frame budget b; when fewer

candidate frames are available, the actual number of frames may be smaller than the nominal budget. The frozen answer model F performs the first inference on the core and produces candidate answer probabilities $\pi_a = p_F(c_a \mid E^{(0)}, q, \mathcal{C})$, which form the core prediction distribution $\boldsymbol{\pi}$. The next stage uses this distribution to determine where supplementary evidence should be acquired.

A.3 Supplementary Evidence in Focused Competition Resolution

Focused Competition Resolution acquires new evidence outside the core. We remove all source frames already included in the core from the complete candidate frame sequence while preserving the temporal order of the remaining frames, yielding $\mathcal{R} = (\tilde{f}_1, \ldots, \tilde{f}_n)$, where n is the number of remaining frames. When $n \geq 32$, the remaining sequence is divided by frame count into eight contiguous partitions of nearly equal size, with any remainder assigned to earlier partitions. From each partition, we select four approximately evenly spaced frames, including the first and last positions, to form a supplementary group A_j. The resulting groups do not share source frames with one another or with the core.

For each supplementary group, we first compute frame level absolute matching differences and then average them over the four frames. The core prediction assigns priorities to answer pairs through competition weights $w_{ab} = \pi_a \pi_b$, producing the supplementary value

$$\begin{aligned} \bar{d}_{A_j}(a,b) &= \frac{1}{4} \sum_{f \in A_j} |z_f(c_a) - z_f(c_b)|, \\ v_\pi(A_j) &= \sum_{a<b} w_{ab}\, \bar{d}_{A_j}(a,b). \end{aligned} \tag{A.2}$$

where $z_f(c_a) = \langle \phi(f), \psi(q, c_a) \rangle$ reuses the frame level matching scores defined above. Both two frame clips and four frame supplementary groups represent evidence in the candidate answer pair space. The former computes matching differences after pooling within a clip, while the latter averages frame wise absolute differences. Taking the absolute value before averaging prevents frames with opposite matching directions from canceling each other during aggregation.

Supplementary groups are sorted in descending order of $v_\pi(A_j)$, with ties resolved in favor of earlier temporal partitions. By default, the top $K_1 = 2$ groups and the top $K_2 = 4$ groups are merged with the core to form E_{40} and E_{48}, respectively. Both expansions use the same core prediction and the same supplementary group ranking, thereby preserving the nested relation $E_{32} \subseteq E_{40} \subseteq E_{48}$.

When the remaining candidate frames are insufficient to construct eight four frame supplementary groups, no partial expansion is performed. The core is retained unchanged, so $E_{40} = E_{48} = E_{32}$. Short videos therefore do not use duplicated frames to satisfy a nominal frame budget.

A.4 Complete Inference and Prediction Fusion

Hypothesis-Discriminative Evidence Acquisition (HDEA) proceeds through core construction, prediction guided evidence supplementation, and final fusion. The first inference completed during Broad Competition Coverage provides $p_0 = \boldsymbol{\pi}$. Focused Competition Resolution then sends E_{40} and E_{48} separately to the same answer model to obtain p_1 and p_2. All visual contexts are presented to the model in the original temporal order. Following Eq. equation 18 in the main paper, the candidate log probabilities from the three contexts are averaged with equal weight, normalized, and used to select the candidate with the highest probability.

All four answer models are frozen and run in bfloat16 precision with their native visual processors. Candidate answer probabilities are computed from logits over restricted option tokens. Each model uses its own core prediction to determine competition weights, and answer probabilities are not shared across models. By default, HDEA uses at most 48 frames in any single model input and invokes the answer model three times. The core probabilities are reused directly and are not recomputed for fusion.

The ablations in the main paper correspond to different output points in this procedure. Broad Coverage (E_{32}) directly returns the first prediction and requires one answer model call. Expanded

prediction (E_{40}) and expanded prediction (E_{48}) return the prediction from the corresponding expanded visual context and each requires two calls, one for the core and one for the target expansion. Uniform48 and Broad Coverage selection with 48 frames directly construct a single 48 frame context and answer once. Their context construction strategies are compared with focused evidence supplementation in Appendix F.3.

A.5 Computational Complexity

The two stages reuse frame level retrieval scores, and the main selection cost comes from the set selection procedure in Broad Competition Coverage. Let the embedding dimension be D and the number of candidate answer pairs be $P = m(m-1)/2$. Once image and text embeddings are available, computing candidate conditioned similarities requires $O(TmD)$ operations, while constructing clip level discriminative representations requires $O(NP)$. A direct implementation of greedy selection examines at most N clips per iteration and therefore requires $O(B_0NP)$ operations over at most B_0 iterations. Maintaining the current coverage values requires $O(P)$ memory.

Focused Competition Resolution scores eight four frame supplementary groups, requiring $O(32P)$ operations for the discriminative scores, followed by sorting eight groups. These complexity terms cover evidence selection only. The actual visual token counts and runtime of the answer models are reported in Appendix G.

B Theoretical Proofs

B.1 Monotonicity and Submodularity of the Coverage Objective

Broad Competition Coverage records the evidence already covered for each answer pair through the maximum discriminative score. We show that this construction gives diminishing returns as the selected set grows.

Fix nonnegative discriminative scores $d_i(a, b)$. For each answer pair, define $g_{ab}(S) = \max_{u_i \in S} d_i(a, b)$ and $g_{ab}(\varnothing) = 0$. This function is nonnegative, normalized, and monotone. Let $d_u(a, b)$ denote the discriminative score of clip u. For any $X \subseteq Y$ and $u \notin Y$,

$$\begin{aligned} g_{ab}(X \cup \{u\}) - g_{ab}(X) &= \max\{0, d_u(a,b) - g_{ab}(X)\} \\ &\geq \max\{0, d_u(a,b) - g_{ab}(Y)\} \\ &= g_{ab}(Y \cup \{u\}) - g_{ab}(Y). \end{aligned} \tag{B.1}$$

Thus, each g_{ab} is submodular. A nonnegative weighted sum of finitely many submodular functions remains submodular, so $H(S) = \sum_{a<b} g_{ab}(S)$ is a normalized, nonnegative, monotone submodular objective. Summing the marginal gains over all answer pairs yields Eq. equation 10 in the main paper.

B.2 Feasible Sets under Temporal Constraints

The disjoint temporal regions form a partition of the candidate clip set. The regional capacity constraints $|S \cap \mathcal{L}_\ell| \leq \kappa_\ell$ define a partition matroid. Adding the total budget constraint $|S| \leq B_0$ is equivalent to rank truncation of this matroid, so the feasible family $\mathcal{I}$ in Eq. equation 8 of the main paper still satisfies the hereditary and exchange properties of a matroid.

When a fixed and feasible required set $R \in \mathcal{I}$ is present, define the following family over the unselected clips:

$$\mathcal{I}_R = \{X \subseteq \mathcal{U} \setminus R : R \cup X \in \mathcal{I}\}, \qquad H_R(X) = H(R \cup X) - H(R). \tag{B.2}$$

$\mathcal{I}_R$ is the feasible family of the matroid obtained by contracting R, and H_R remains a normalized, nonnegative, monotone submodular function. In the implementation, subtracting the regional capacity and total budget already occupied by required clips corresponds exactly to solving this residual problem.

Table C.1: Evaluation scale and primary metrics for the three benchmarks.

Benchmark	Questions	Physical videos	Metric
VideoMME	2,700	900	Accuracy; results are also reported for short, medium, and long duration groups.
LongVideoBench	1,337	753	Accuracy.
MLVU	2,174	1,122	Unweighted macro average (M-AVG) across seven task accuracies.

B.3 Approximation Guarantee for Greedy Selection

We first consider the case without required clips. Let the greedy sequence be $G_i = \{u_{g_1}, \ldots, u_{g_i}\}$ and let the final set be $G = G_r$. The algorithm stops when the budget is exhausted or no feasible element remains, so G is a basis of the constrained matroid. By monotonicity, an optimal independent set can also be extended to a basis O without decreasing the objective value.

By the matroid exchange property, O can be ordered as $u_{o_1}, \ldots, u_{o_r}$ such that $G_{i-1} \cup \{u_{o_i}\}$ is feasible at step i. The greedy rule gives

$$\Delta_H(u_{o_i} \mid G_{i-1}) \leq \Delta_H(u_{g_i} \mid G_{i-1}). \tag{B.3}$$

Combining monotonicity and submodularity yields

$$\begin{aligned} H(O) &\leq H(G \cup O) \\ &\leq H(G) + \sum_{u \in O \setminus G} \Delta_H(u \mid G) \\ &\leq H(G) + \sum_{i=1}^{r} \Delta_H(u_{o_i} \mid G_{i-1}) \\ &\leq H(G) + \sum_{i=1}^{r} \Delta_H(u_{g_i} \mid G_{i-1}) = 2H(G). \end{aligned} \tag{B.4}$$

Therefore, $H(G) \geq \frac{1}{2} H(O)$, which gives the classical greedy guarantee for monotone submodular maximization under a matroid constraint (Fisher et al., 1978).

With a required set, the same result applies to $(\mathcal{I}_R, H_R)$. Let X_g be the greedy solution to the contracted problem, let $S^\star$ be the optimal solution to the original problem subject to containing R, and define $S_g = R \cup X_g$. Then

$$\begin{aligned} H(S_g) - H(R) &\geq \tfrac{1}{2}\big[H(S^\star) - H(R)\big], \\ H(S_g) &\geq \tfrac{1}{2} H(S^\star) + \tfrac{1}{2} H(R) \geq \tfrac{1}{2} H(S^\star). \end{aligned} \tag{B.5}$$

This establishes the greedy approximation guarantee stated in the main paper. The approximation guarantee applies to the coverage objective under fixed retrieval scores, while the effect of the selected evidence on question answering accuracy is evaluated empirically.

C Evaluation Setup and Result Statistics

C.1 Benchmarks and Metrics

All three benchmarks use the same complete question lists as the main paper, with different methods paired by question identifier (ID) and candidate answer order. Missing, duplicate, and failed records are checked before evaluation. Table C.1 reports the number of questions, the number of physical videos, and the primary metric for each benchmark. MLVU uses its official development set.

VideoMME and LongVideoBench use accuracy as the primary metric. For MLVU, we follow the official evaluation and report M-AVG, the unweighted mean of accuracy across its seven task categories. The per question micro average on MLVU is reported only as an auxiliary result. Accuracy is expressed in percent and gains are expressed in percentage points. All gains are computed from unrounded metrics and then rounded to two decimal places, so a displayed gain may differ by 0.01 percentage points from the difference between two displayed table entries.

In paired comparisons, "Corrections" count questions that are incorrect under the reference configuration but correct under the compared configuration, while "Regressions" count the opposite changes. For MLVU, these counts are computed per question and therefore do not directly correspond to changes in the task balanced M-AVG.

D Controlled Comparisons for Broad Competition Coverage

D.1 Candidate Answer Information and Set Selection

This section supplements the Broad Competition Coverage experiments in Section 4.4 of the main paper by examining how candidate answer information is used for clip scoring and how maintaining set level coverage during selection affects performance. All five configurations use Qwen3-VL-8B, a 32 frame budget, and one answer model call. OptionPool32, PairwisePointwise32, and Broad Coverage (E_{32}) share the same two frame candidate set, temporal constraints, required clips, and budget. AKS32 follows its original frame selection rule.

OptionPool32 averages the normalized embeddings of all candidate conditioned texts with equal weight and then normalizes the resulting vector into a single query representation. Let $e_a = \psi(q, c_a)/\|\psi(q, c_a)\|$. Then $e_{\text{pool}} = (\sum_a e_a)/\|\sum_a e_a\|$. The implementation sums embeddings in the fixed candidate order and then uses the similarity between this pooled query and each frame to construct an independent clip score.

PairwisePointwise32 uses the same answer pair discriminative scores as Broad Competition Coverage, but averages them into an independent score for each clip:

$$s_i = \frac{2}{m(m-1)} \sum_{a<b} d_i(a, b). \tag{D.1}$$

Broad Competition Coverage instead maintains the current coverage value for every answer pair and selects clips according to their additional coverage. The comparison therefore directly tests the set complementarity design used in the main method.

D.2 Performance Gains from Broad Competition Coverage

Broad Competition Coverage improves over PairwisePointwise32 by 4.33, 8.30, and 10.48 percentage points on the three benchmarks, respectively. Relative to OptionPool32, the gains are 1.74, 0.45, and 0.38 percentage points.

PairwisePointwise32 performs worse than OptionPool32 on all three benchmarks, with decreases of 2.59, 7.85, and 10.10 percentage points on VideoMME, LongVideoBench, and MLVU, respectively. The corresponding numbers of corrections and regressions are 242/312, 83/188, and 161/375. Together with Table D.1, these results show that averaging candidate pair scores into an independent score for each clip loses the set level complementarity captured by Broad Competition Coverage.

D.3 Candidate Answer Pair Coverage Diagnostic

Figure 1 analyzes the set complementarity of Broad Competition Coverage through normalized candidate answer pair coverage. This diagnostic omits the temporal region and explicit temporal reference constraints used during formal inference, allowing the analysis to isolate how much retrieval discrimination is retained as the number of selected clips increases.

For question q, let $\mathcal{U}_q$ denote the complete candidate clip set and $S_{q,K}$ the set of K selected clips. We measure the fraction of candidate pair discrimination retained by the selected set as

Table D.1: Gains and prediction changes of Broad Competition Coverage (E_{32}) relative to each reference configuration. Gains on MLVU correspond to M-AVG, while corrections and regressions are counted per question.

Benchmark	Reference configuration	Gain (pp)	Corrections	Regressions
VideoMME	Uniform32	+2.96	275	195
VideoMME	AKS32	+3.81	238	135
VideoMME	OptionPool32	+1.74	222	175
VideoMME	PairwisePointwise32	+4.33	297	180
LongVideoBench	Uniform32	+6.36	153	68
LongVideoBench	AKS32	+2.99	124	84
LongVideoBench	OptionPool32	+0.45	87	81
LongVideoBench	PairwisePointwise32	+8.30	172	61
MLVU	Uniform32	+9.67	351	140
MLVU	AKS32	+2.71	149	87
MLVU	OptionPool32	+0.38	132	120
MLVU	PairwisePointwise32	+10.48	352	126

$$C_q(S_{q,K}) = \frac{\sum_{a<b} \max_{u_i \in S_{q,K}} d_i(a,b)}{\sum_{a<b} \max_{u_i \in \mathcal{U}_q} d_i(a,b)}.$$

The metric is first averaged equally over questions within each benchmark and then equally over VideoMME, LongVideoBench, and MLVU. It measures retrieval discrimination coverage rather than human annotated evidence coverage or question answering accuracy.

If the denominator in $C_q(S_{q,K})$ is no larger than float64 machine precision, the normalized coverage for that question is treated as missing and excluded from the average. This condition does not occur on any of the three benchmarks, so all questions contribute to Figure 1. At $K = 4$, the coverage values are 99.1% on VideoMME, 97.6% on LongVideoBench, and 99.0% on MLVU, giving an equal benchmark average of 98.6%.

E Cross Model Transfer

Section 4.3 of the main paper examines how HDEA transfers across different answer models. This section provides the complete results for four frozen video multimodal large language models (MLLMs) on three long video benchmarks. We compare Uniform32, Broad Competition Coverage, and the complete HDEA pipeline to examine whether evidence acquisition driven by candidate answer competition transfers consistently across different answer models.

As shown in Figure E.1, Broad Competition Coverage outperforms Uniform32 in all 12 combinations formed by four answer models and three benchmarks, with gains ranging from 0.78 to 9.67 percentage points. Qwen3-VL-8B improves by 2.96, 6.36, and 9.67 percentage points on VideoMME, LongVideoBench, and MLVU, respectively. InternVL3-8B, mPLUG-Owl3-7B, and LLaVA-Video-7B also achieve positive gains on all three benchmarks. These results show that the evidence selection strategy in Broad Competition Coverage transfers across different frozen answer models.

Complete HDEA preserves this cross model consistency. Relative to Uniform32, HDEA improves all 12 combinations, with gains ranging from 1.78 to 10.30 percentage points. Qwen3-VL-8B improves by 3.85, 6.96, and 10.30 percentage points on the three benchmarks, while the other three answer models also maintain positive gains across all benchmarks. Relative to Broad Competition Coverage, complete HDEA further improves 11 of the 12 combinations, with an average gain of 0.70 percentage points. The only decrease is 0.30 percentage points for InternVL3-8B on LongVideoBench.

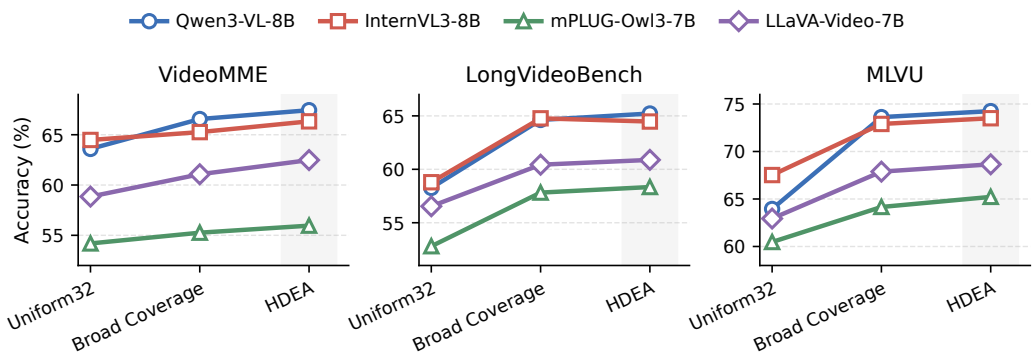


Figure E.1: **Cross model transfer of HDEA across different answer models.** The three panels correspond to VideoMME, LongVideoBench, and MLVU, respectively, and compare Uniform32, Broad Competition Coverage, and complete HDEA across four frozen video MLLMs. Broad Competition Coverage and HDEA outperform their corresponding Uniform32 baselines in all 12 model and benchmark combinations.

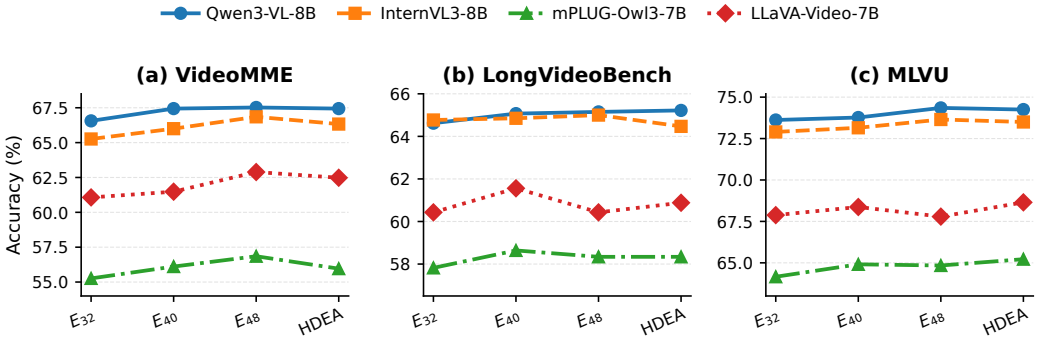


Figure F.1: **Prediction evolution during Focused Competition Resolution.** The three panels correspond to VideoMME, LongVideoBench, and MLVU. Each curve represents a frozen video MLLM. The horizontal axis shows the evidence core E_{32} produced by Broad Competition Coverage, the two expanded visual contexts E_{40} and E_{48}, and complete HDEA after fusing the predictions from all three contexts.

These results show that Broad Competition Coverage constructs effective evidence cores across different answer models, while complete HDEA further uses each answer model's own core prediction to guide focused evidence supplementation. The entire evidence acquisition process requires no additional training for a specific answer model and demonstrates consistent cross model transfer.

F Complete Results for Focused Competition Resolution

F.1 From Core Predictions to Expanded Predictions

This section supplements the Focused Competition Resolution experiments in Section 4.4 of the main paper. For each model and question, all configurations share the same broad coverage core and use the model's own core prediction to determine the ranking of supplementary groups. We follow the prediction trajectory $E_{32} \rightarrow E_{40} \rightarrow E_{48} \rightarrow$ HDEA. The E_{40} and E_{48} configurations directly output the prediction from the corresponding expanded visual context, while HDEA fuses the core prediction with the two expanded predictions.

As shown in Figure F.1, expanding from E_{32} to E_{40} improves all 12 combinations formed by four answer models and three benchmarks. Relative to the core prediction, E_{48} improves 10 combi-

Table F.1: Gains and prediction changes of HDEA relative to Broad Competition Coverage (E_{32}).

Answer model	Model size	Benchmark	Gain (pp)	Corrections	Regressions
Qwen3-VL	8B	VideoMME	+0.89	78	54
		LongVideoBench	+0.60	29	21
		MLVU	+0.64	55	40
InternVL3	8B	VideoMME	+1.07	77	48
		LongVideoBench	-0.30	24	28
		MLVU	+0.60	41	28
mPLUG-Owl3	7B	VideoMME	+0.70	86	67
		LongVideoBench	+0.52	24	17
		MLVU	+1.06	67	47
LLaVA-Video	7B	VideoMME	+1.41	73	35
		LongVideoBench	+0.45	27	21
		MLVU	+0.77	44	33

nations, matches one, and decreases one. These results show that adding supplementary evidence while preserving the original evidence core generally improves answer model predictions.

Complete HDEA then fuses the candidate answer distributions from E_{32}, E_{40}, and E_{48}. Relative to the broad coverage core, complete HDEA improves 11 of the 12 combinations by an average of 0.70 percentage points. The only decrease occurs for InternVL3-8B on LongVideoBench and is 0.30 percentage points. Relative to directly using E_{48}, fusion improves 4 combinations, leaves 1 unchanged, and decreases 7. The main performance gains in Focused Competition Resolution therefore come from acquiring supplementary evidence on top of the preserved core, while prediction fusion integrates decisions across visual contexts and its effect varies across answer models and benchmarks.

F.2 Prediction Changes across Complete Focused Competition Resolution

Table F.1 reports corrections and regressions of HDEA relative to the core prediction.

Except for InternVL3-8B on LongVideoBench, corrections exceed regressions in the remaining 11 combinations. This pattern agrees with the overall change of complete HDEA relative to the core prediction shown in Figure F.1.

F.3 Core Expansion versus Direct 48 Frame Selection

The two stage design in the main paper preserves the established evidence core and then uses the core prediction to acquire targeted supplementary evidence. To distinguish this process from simply increasing the frame count, we compare three context construction strategies under the same maximum input of 48 frames. Expanded prediction (E_{48}) preserves E_{32} and adds four supplementary groups. Broad Coverage selection with 48 frames reruns broad coverage selection from scratch. Uniform48 samples frames uniformly over time. The first strategy requires two answer model calls, one for the core and one for the expansion, while the other two require one call each.

Expanded prediction (E_{48}) improves over Uniform48 by 2.00, 4.71, and 7.53 percentage points on the three benchmarks, respectively. Relative to direct Broad Coverage selection, preserving the core and adding supplementary evidence gives higher scores on VideoMME and LongVideoBench and a lower score on MLVU. The two methods differ in both core preservation and the selection rule for additional evidence, so this comparison characterizes the complete context construction strategy.

Table F.2: Paired comparison of 48 frame context construction strategies. Gain is defined as the compared configuration minus the reference configuration.

Benchmark	Comparison	Gain (pp)	Corrections	Regressions
VideoMME	Expanded prediction (E_{48}) vs. Broad Competition Coverage (48 frames)	+0.89	127	103
	Expanded prediction (E_{48}) vs. Uniform48	+2.00	228	174
	Broad Competition Coverage (48 frames) vs. Broad Competition Coverage (E_{32})	+0.07	94	92
LongVideoBench	Expanded prediction (E_{48}) vs. Broad Competition Coverage (48 frames)	+0.60	62	54
	Expanded prediction (E_{48}) vs. Uniform48	+4.71	129	66
	Broad Competition Coverage (48 frames) vs. Broad Competition Coverage (E_{32})	-0.07	40	41
MLVU	Expanded prediction (E_{48}) vs. Broad Competition Coverage (48 frames)	-0.52	93	105
	Expanded prediction (E_{48}) vs. Uniform48	+7.53	293	124
	Broad Competition Coverage (48 frames) vs. Broad Competition Coverage (E_{32})	+1.26	86	60

Table F.3: Supplementary group ranking diagnostic on MLVU. Gain is defined as competition weighted ranking minus temporal ranking and is reported in percentage points. The primary metric is M-AVG, while corrections and regressions are counted per question.

Output	Temporal ranking	Competition weighted ranking	Gain (pp)	Corrections	Regressions
E_{40}	74.26	73.77	-0.49	51	56
E_{48}	74.00	74.35	+0.36	73	58
Fusion of three predictions	74.77	74.25	-0.52	35	40

F.4 Supplementary Group Ranking Diagnostic

On MLVU, we compare competition weighted ranking with temporal ranking while holding the Broad Competition Coverage core, candidate supplementary groups, frame budget, number of answer model calls, prompt, and output rule fixed.

Compared with temporal ranking, competition weighted ranking improves M-AVG at E_{48} and lowers it at E_{40} and after fusion. Its effect varies with the visual budget and output configuration.

G Visual Budget and Answering Cost

To examine the effect of the visual budget, we keep the 32 frame core and supplementary group ranking fixed and vary the number of added groups. We then report answer model calls, visual tokens, and runtime for the core prediction, expanded predictions, and full HDEA. Figure G.1 summarizes

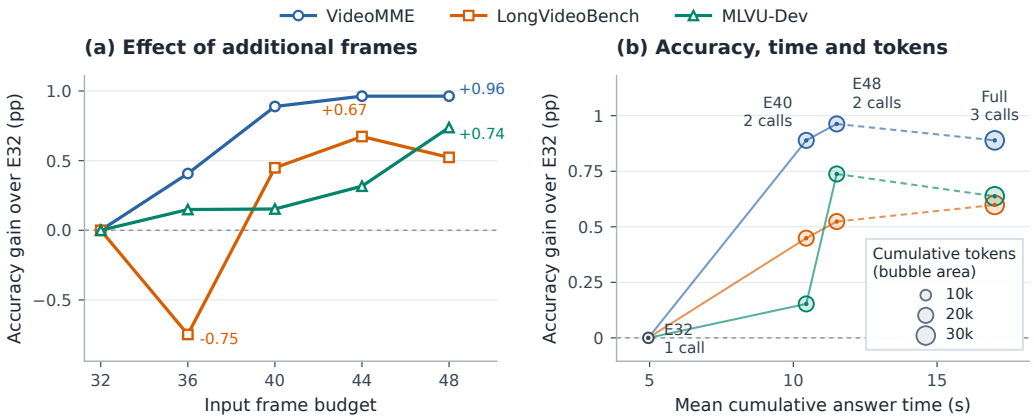


Figure G.1: **Visual budget and answering stage cost.** Results for Qwen3-VL-8B on three long video benchmarks. (a) Performance of expanded predictions under different frame budgets with the 32 frame evidence core and supplementary group ranking fixed. (b) Cumulative answering time and accuracy gain relative to the corresponding core prediction for each configuration. Bubble area represents cumulative visual tokens. MLVU uses M-AVG, and costs are averaged over all 6,211 questions across the three benchmarks.

Table G.1: Fixed core budget sweep with Qwen3-VL-8B. Bold indicates the best result among the evaluated budgets.

Benchmark	32 frames	36 frames	40 frames	44 frames	48 frames	Best gain
VideoMME	66.56	66.96	67.44	**67.52**	**67.52**	+0.96
LongVideoBench	64.62	63.87	65.07	**65.30**	65.15	+0.67
MLVU	73.62	73.77	73.77	73.93	**74.35**	+0.74

the Qwen3-VL-8B results across the three benchmarks, and Table G.2 gives the cost breakdown for all four answer models.

G.1 Number of Supplementary Evidence Groups and Prediction Performance

We fix the 32 frame core produced by Broad Competition Coverage and the supplementary group ranking, then sequentially add the first one, two, three, and four supplementary groups to form visual contexts with 36, 40, 44, and 48 frames. Each budget directly outputs the prediction from the corresponding context without fusing predictions from multiple contexts. This experiment changes only the amount of supplementary evidence.

As shown in Figure G.1(a), VideoMME improves as the budget increases and reaches the same best result at 44 and 48 frames. LongVideoBench decreases at 36 frames, then reaches its highest value at 44 frames before a small decline at 48 frames. MLVU reaches its highest value at 48 frames. The best gains over the core are 0.96, 0.67, and 0.74 percentage points on the three benchmarks, respectively. The benefit of supplementary evidence varies across benchmarks and budgets, and additional frames do not always produce further accuracy gains.

G.2 Answering Cost for Different Outputs

We measure all answer model calls required to obtain each output. We distinguish the number of unique source frames from cumulative frame instances across calls. Because the HDEA contexts are nested, complete HDEA uses at most the frames contained in E_{48} as distinct source frames, while the core frames are processed again in the expanded contexts. Videos with insufficient candidate frames use fewer than the nominal frame budgets.

Table G.2: Answering cost by model and benchmark. Each entry reports visual tokens (k) / time (s); HDEA reports the cumulative cost of three answer model calls. LVB denotes LongVideoBench.

Answer model	Benchmark	Core (1 call)	E_{40} (1 call)	E_{48} (1 call)	HDEA (3 calls)
Qwen3-VL-8B	VideoMME	10.45 / 4.39	10.83 / 4.68	10.66 / 5.55	31.94 / 14.63
	LVB	10.34 / 5.00	10.61 / 4.48	10.36 / 5.15	31.32 / 14.64
	MLVU	8.22 / 5.62	8.98 / 7.13	9.58 / 8.67	26.79 / 21.42
InternVL3-8B	VideoMME	8.16 / 4.03	10.01 / 4.79	11.86 / 5.56	30.04 / 14.38
	LVB	7.28 / 3.08	8.78 / 3.68	10.27 / 4.37	26.33 / 11.12
	MLVU	8.19 / 6.25	10.24 / 8.45	12.29 / 10.29	30.72 / 24.98
mPLUG-Owl3-7B	VideoMME	23.24 / 3.55	28.51 / 4.23	33.79 / 5.21	85.53 / 12.99
	LVB	20.74 / 3.23	24.99 / 3.89	29.25 / 4.20	74.98 / 11.32
	MLVU	23.33 / 5.59	29.16 / 8.13	34.99 / 9.51	87.48 / 23.23
LLaVA-Video-7B	VideoMME	6.69 / 3.63	8.21 / 4.45	9.73 / 5.54	24.64 / 13.62
	LVB	5.97 / 3.32	7.20 / 4.03	8.43 / 4.68	21.60 / 12.03
	MLVU	6.72 / 6.01	8.40 / 8.49	10.08 / 10.72	25.20 / 25.22

Figure G.1(b) reports answering costs averaged over all 6,211 questions across the three benchmarks. The vertical axis shows the accuracy gain over the corresponding core prediction, and bubble area represents cumulative visual tokens.

Experiments were conducted on a server equipped with four NVIDIA A100-PCIE-40GB GPUs, an Intel Xeon processor (x86_64, 32 threads, 2.99 GHz), and 256 GiB of RAM. The server ran Ubuntu 22.04.5 LTS.

Runtime includes video decoding, visual preprocessing, and synchronous forward computation for all answer model calls. SigLIP feature extraction and evidence selection are excluded. Visual token counts are measured after each answer model applies its native visual processing. Table G.2 reports the answering stage cost across answer models and benchmarks.

Broad Coverage requires one answer model call and uses an average of 9.65k visual tokens and 4.95 seconds. Producing E_{40} or E_{48} first requires the core prediction and then one call on the target expanded context, resulting in cumulative visual token counts of 19.79k and 19.87k and cumulative times of 10.45 and 11.51 seconds, respectively. HDEA computes the core prediction and both expanded predictions before fusion, using three answer model calls, 30.00k cumulative visual tokens, and 17.01 seconds.

For Qwen3-VL-8B, E_{48} and E_{40} have similar cumulative visual token counts, while E_{48} achieves higher accuracy on all three benchmarks. Complete HDEA additionally evaluates E_{40} and then fuses the three predictions, while the resulting accuracy change varies across benchmarks. Together with Appendix F, these results show that the main improvement from Focused Competition Resolution comes from supplementary evidence added to the preserved core, while fusion further integrates decisions across different visual contexts.

G.3 Cost Breakdown across Answer Models

Table G.2 reports the measured costs for four answer models on the three benchmarks. The core, E_{40}, and E_{48} columns record the cost of scoring each visual context once, while the HDEA column reports the total cost of all three calls. The total answering cost for expanded prediction (E_{40}) or expanded prediction (E_{48}) is obtained by adding the core column to the corresponding expansion column.

Costs are accumulated from unrounded records and rounded only for reporting. No linear conversion is assumed between source frame count and visual token count because the latter also depends on each model's native visual processing. The reported runtimes describe the measured execution of each model under the experimental setup and should be interpreted within that runtime environment.

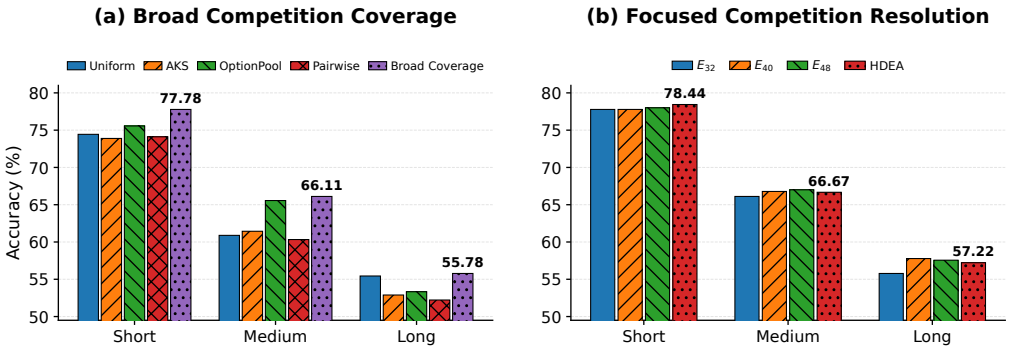


Figure H.1: **Results across VideoMME duration groups.** (a) Comparison of Broad Competition Coverage with Uniform, AKS, OptionPool, and Pairwise. All configurations use 32 frames and one answer model call. (b) Results from expanding the evidence core E_{32} to E_{40} and E_{48}, followed by HDEA prediction fusion.

H Duration and Task Breakdowns

H.1 Results across Video Durations

The short, medium, and long duration groups of VideoMME each contain 900 questions. We examine the controlled frame selection results of Broad Competition Coverage and the changes produced by Focused Competition Resolution as the visual context is expanded from the core. Figure H.1 summarizes the performance of both stages across video durations.

As shown in Figure H.1(a), Broad Competition Coverage reaches 77.78%, 66.11%, and 55.78% on the short, medium, and long duration groups, respectively. Relative to Uniform32, the corresponding gains are 3.33, 5.22, and 0.33 percentage points. Relative to PairwisePointwise32, the gains are 3.67, 5.78, and 3.56 percentage points. The largest improvement appears in the medium duration group. Broad Competition Coverage also maintains an advantage over the other controlled frame selection strategies in the long duration group.

Figure H.1(b) shows how Focused Competition Resolution changes performance across the three duration groups. On short videos, performance increases from 77.78% at E_{32} to 78.44% with complete HDEA. On medium videos, E_{48} reaches 67.00% and complete HDEA reaches 66.67%. On long videos, adding supplementary evidence raises performance from 55.78% at E_{32} to 57.78% at E_{40} and 57.56% at E_{48}, while complete HDEA reaches 57.22%. Core expansion yields the largest gains on the long duration group.

H.2 Results across Tasks

The primary MLVU metric is the equal average of accuracies over seven task categories. To further analyze Broad Competition Coverage across tasks, we first report the absolute accuracy for each task and then use a heatmap to show task level gains of Broad Competition Coverage relative to four reference methods.

Figure H.2 shows the gain distribution across tasks. Relative to Uniform32, Broad Coverage improves five of the seven tasks, matches Anomaly Recognition, and decreases Topic Reasoning by 0.76 percentage points. The gains on Counting, Order, and NeedleQA reach 22.33, 15.06, and 11.83 percentage points, respectively. Relative to PairwisePointwise32, Broad Coverage improves Counting, NeedleQA, and Egocentric Reasoning by 31.07, 14.08, and 13.92 percentage points, respectively, further illustrating the effect of complementary set selection on these tasks.

Task level changes also differ across reference methods. For example, relative to OptionPool32, Broad Coverage improves PlotQA by 3.15 percentage points and decreases Order by 8.11 percentage

Table H.1: Task accuracies, M-AVG, and auxiliary micro average on MLVU.

Task	Uniform 32	AKS 32	OptionPool 32	Pairwise Pointwise32	Broad Coverage (E_{32})
Anomaly Recognition	76.00	71.00	69.50	77.50	76.00
Counting	32.04	56.80	54.85	23.30	54.37
Egocentric Reasoning	60.51	69.03	69.89	55.97	69.89
NeedleQA	68.45	81.13	80.56	66.20	80.28
Order	51.74	56.37	74.90	63.71	66.80
PlotQA	70.32	74.95	76.99	71.24	80.15
Topic Reasoning	88.59	87.07	85.93	84.03	87.83
M-AVG	63.95	70.91	73.23	63.13	73.62
Per-question micro avg.	65.32	72.17	74.47	64.63	75.02

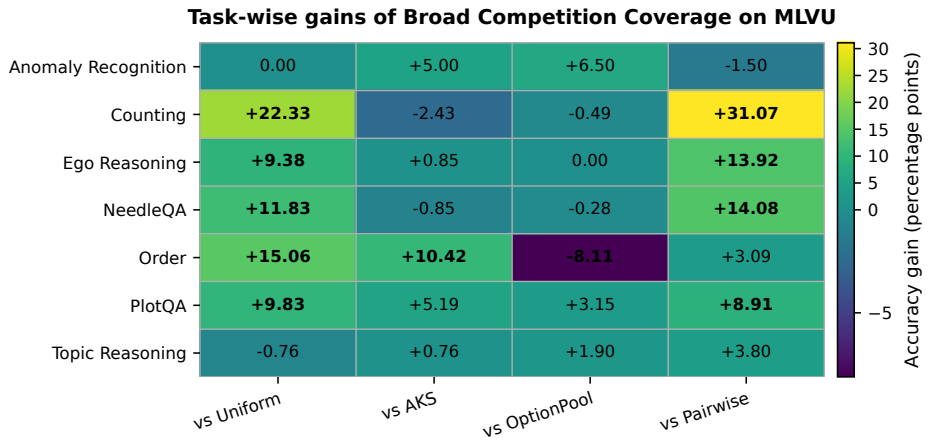


Figure H.2: **Task level gains of Broad Competition Coverage on MLVU.** Each cell shows the accuracy change of Broad Coverage (E_{32}) relative to the corresponding reference method, measured in percentage points. Positive values indicate higher accuracy for Broad Coverage, while negative values indicate higher accuracy for the reference method.

points. The overall M-AVG therefore summarizes performance across the seven tasks, while the task level results reveal how the gains are distributed among them.

Table H.2 provides the per question micro average as an auxiliary result. Method comparisons in the main paper and Appendix F use the task balanced M-AVG.

I Evidence and Prediction Changes in Qualitative Cases

Figure 4 in the main paper uses two independent examples to illustrate Broad Coverage selection and focused evidence supplementation. Frame selection, candidate answer probabilities, and supplementary group rankings in these cases are taken from saved experimental outputs. Ground truth answers are used for case selection and annotation in the figure, while evidence acquisition, answer inference, and method parameter selection follow the same procedure as the full evaluation.

Table H.2: Auxiliary micro-average accuracy on MLVU during Focused Competition Resolution. The primary metric is M-AVG.

Answer model	Broad Coverage (E_{32})	Expanded (E_{40})	Expanded (E_{48})	HDEA
Qwen3-VL-8B	75.02	75.25	75.67	75.71
InternVL3-8B	74.66	74.98	75.34	75.25
mPLUG-Owl3-7B	67.30	67.76	67.80	68.22
LLaVA-Video-7B	70.42	70.70	70.29	70.93

Table I.1: Candidate answer probabilities (%) for Uniform32 and Broad Coverage.

Configuration	A	B	C	D	Prediction
Uniform32	97.14	0.84	0.24	1.78	A
Broad Coverage (E_{32})	0.0006	0.0013	99.9971	0.0010	C

I.1 Broad Coverage Captures Evidence for a Brief Event

The VideoMME example asks **"What is the black-white cat doing in the toilet?"** The four candidate answers are A) Drinking the water, B) Taking a shower, C) Shredding paper, and D) Using the litter box. The ground truth answer is C.

Uniform32 predicts A with a probability of 97.14%. The Broad Coverage core contains the paper shredding event around 65.8 to 74.5 seconds. The three key frames shown in the figure occur at 65.8, 70.6, and 74.5 seconds. Uniform sampling misses this event, while the prediction from the Broad Coverage core changes to the correct answer C. This example illustrates how different evidence selection strategies can change the model's judgment among candidate answers.

I.2 Focused Evidence Supplementation Corrects the Core Prediction

The MLVU example asks **"Where did the cartoon cat hide in the end?"** The four candidate answers are A) Under a large cartoon dog, B) Under the piano, C) In the car, and D) On the tree. The ground truth answer is A.

The core prediction at E_{32} is strongly concentrated on the incorrect answer D (97.32%), while the correct answer A receives 2.59%. Among the six candidate pairs, A and D therefore receive the largest competition weight, $w_{AD} = 0.02524$. Focused Competition Resolution combines this priority with the discriminative scores of supplementary groups, producing the ranking in Table I.2.

Table I.2: Selected supplementary groups and their inclusion in the expanded visual contexts.

Supplementary group	Rank	v_π	Visual context
A6	1	0.001107	Added at E_{40}; retained in E_{48}
A5	2	0.001000	Added at E_{40}; retained in E_{48}
A8	3	0.000849	Added at E_{48}
A7	4	0.000834	Added at E_{48}

The first two groups, A6 and A5, add eight frames to the core and form E_{40}. A8 and A7 are then added to form E_{48} while retaining all previously selected frames. Both expansions follow the ranking determined by the same core prediction.

Supplementary evidence changes the prediction from D to the correct answer A. The probability of A reaches 98.21% at E_{40} and 99.70% at E_{48}, and the fused prediction selects A with 89.79% probability.

Table I.3: Candidate answer probabilities (%) for the core, expanded contexts, and HDEA fusion.

Output	A	B	C	D	Prediction
E_{32}	2.59	0.07	0.02	97.32	D
E_{40}	98.21	0.17	0.03	1.59	A
E_{48}	99.70	0.07	0.01	0.22	A
HDEA fusion	89.79	0.29	0.06	9.87	A