BEYOND ANSWERING: SMALL MODELS EMPOWER LARGE MODELS FOR LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in long-video reasoning enable models to actively search for visual evidence through tool use. However, these approaches commonly assign evidence acquisition and answer generation to the same model, coupling repeated visual exploration with that model's computational cost and reasoning limitations. We find that small models can more readily locate and verify relevant evidence yet struggle to integrate it into a correct answer, revealing distinct capability requirements for evidence acquisition and reasoning. Building on this observation, we propose an **evidence-driven collaboration framework** in which a small model handles repeated, multi-step evidence search and verification, while a large model integrates the acquired evidence and produces the final answer in a single invocation. Crucially, we distinguish evidence search under sparse observations from direct evidence localization: when decisive details remain unobserved, the true evidence location may not be a search target supported by the available cues. We therefore supervise search priorities grounded in the current observations, teaching the small model **where to inspect next**, while dense local verification determines each candidate's actual evidential value. This decoupled framework supports flexible test-time scaling through configurable evidence acquisition and answer routing. Experiments demonstrate that small models can strengthen large-model reasoning through evidence acquisition, extending the collaborative value of small models beyond answering. Pairing a 4B evidence search and verification model with a frozen 27B model for final reasoning achieves 74.8% on VideoMME, 79.2% on MLVU, and 58.7% on LVBench. Code, data, and models will be released.

1 INTRODUCTION

Multimodal language models have made substantial progress in video understanding (Zhang et al., 2024a; Bai et al., 2025a; Li et al., 2026), yet answering questions about long videos remains challenging. Recent methods address this challenge through active visual exploration, allowing models to search for evidence and inspect relevant video intervals on demand (Zhang et al., 2026b; Yang et al., 2026a; Zeng et al., 2026). Although the inference process now involves multiple steps of evidence acquisition and reasoning, final-answer accuracy remains the primary metric for evaluating model performance. ***But does limited standalone answering ability imply limited value in solving difficult questions?*** Our findings suggest otherwise: even when small models cannot reliably answer a question, they can still search for and verify useful visual evidence, enabling larger models to outperform standalone inference. Small models can therefore contribute beyond answering by providing better evidence for a stronger model's reasoning.

Understanding this phenomenon requires distinguishing two bottlenecks in long-video question answering. Relevant visual information is sparsely distributed throughout lengthy videos, making evidence acquisition challenging. Even when relevant content has been observed, interpreting events, connecting clues across temporal segments, and determining the answer remain distinct reasoning challenges. The top panel of Figure 1 illustrates this distinction: stronger reasoning cannot compensate for missing evidence, while successful evidence acquisition does not guarantee correct responds.

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.

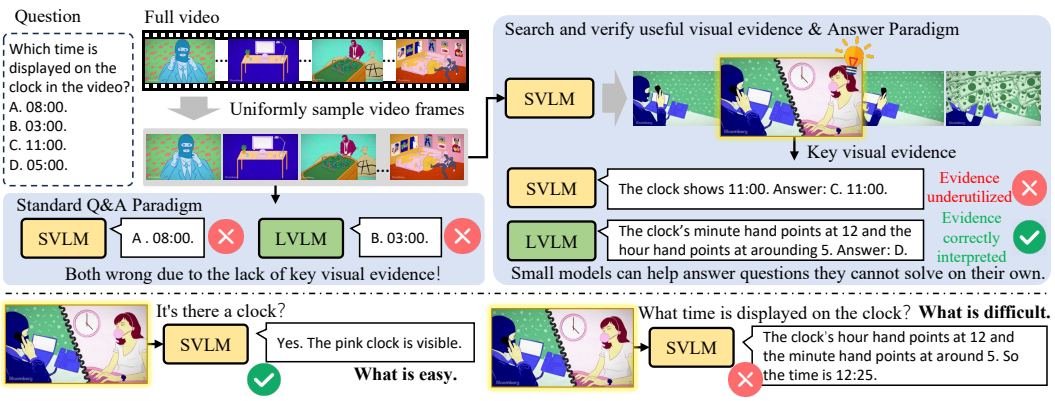


Figure 1: **Evidence acquisition and reasoning impose different capability requirements.** Top: uniform sampling can miss decisive evidence, limiting both Small Vision Language Model(SVLM) and Large Vision Language Model(LVLM). Active exploration allows a SVLM to acquire useful evidence, yet interpreting it correctly may require a LVLM. Bottom: identifying a question-relevant visual cue, such as the presence of a clock, can be easier than resolving the fine-grained information needed to answer the question.

Our central observation is that evidence acquisition and final reasoning impose different demands on model capability. Identifying a promising interval or confirming a useful visual fact does not necessarily require resolving the question, as illustrated in Figure 1 (bottom). A small model may therefore be capable of searching for and verifying useful evidence while remaining unable to derive the correct answer from it. Evaluating small models solely by their final answers can therefore overlook their ability to acquire useful evidence.

This difference in capacity requirements reveals an **exploration–reasoning dilemma** when a single model scale is used throughout the process. Larger models typically offer stronger reasoning capabilities at higher inference costs: using a small model throughout may leave useful evidence underexploited, whereas using a large model incurs substantial costs for repeated search and verification. As existing methods devote increasing computation to evidence acquisition (Yang et al., 2026a;b), this tension becomes more pronounced. The central question is therefore whether we can combine low-cost evidence acquisition by small models with the stronger reasoning capabilities of large models, enabling more extensive visual exploration without relying on a large model at every step.

Motivated by this insight, we propose an evidence-driven framework for small–large model collaboration that extends the role of small models from standalone answering to evidence acquisition for difficult questions. Specifically, a small model first attempts to answer from global video observations. For questions it cannot reliably answer, it continues to search for and verify candidate intervals, passing the resulting visual evidence to a large model for joint reasoning. Abstention thus relinquishes responsibility for the final answer without ending the small model's participation in solving the question. By combining instance-level answer routing with functional specialization within each instance, our framework allows small models both to answer some questions directly and to strengthen large-model reasoning through better evidence inputs.

To enable reliable evidence acquisition from sparse video observations, we distinguish **evidence localization** from **evidence search**. Localization asks the model to identify the temporal boundaries of the evidence (Wang et al., 2026; Zhang et al., 2026c; Zhu et al., 2026), whereas search asks where to inspect when the current observations do not yet provide sufficient evidence. Existing approaches commonly supervise search through localization targets, constructing multi-round or multi-interval trajectories around annotated or verified evidence locations (Yang et al., 2026a;b; Zeng et al., 2026). However, an interval identified from sufficiently detailed video observations may contain decisive information that is entirely absent from the sparse frames available to the search model. Requiring the model to predict that interval can therefore turn missing information into misleading training signals.This motivates a different question: ***before the evidence is visible, what makes an interval worth inspecting?*** We formulate temporal search as learning **where to inspect next**, supervising

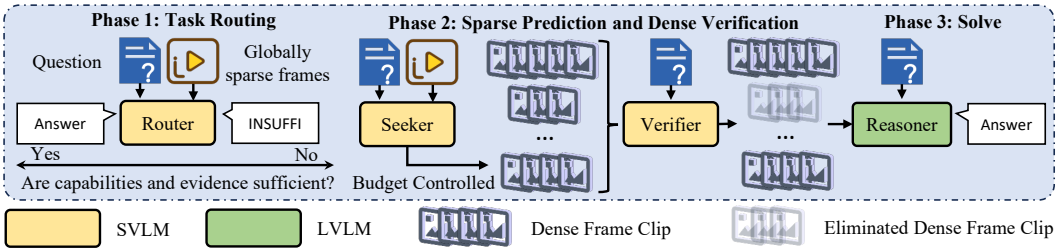


Figure 2: Overview of evidence-driven collaboration framework. Small-model evidence acquisition supports large-model reasoning through budget-controlled search and local verification.

candidate priorities according to the currently visible cues rather than requiring each prediction to recover the true evidence location.

Our main contributions are as follows:

- **We reveal different capacity requirements for evidence acquisition and evidence-based reasoning in long-video understanding.** With active visual exploration, small models with limited standalone answering ability can still effectively search for and verify key evidence, showing that a capability gap in final reasoning does not necessarily extend to evidence acquisition.
- **We propose a flexible, evidence-driven framework for small–large model collaboration.** By decoupling evidence acquisition from final reasoning, the framework supports budget-adaptive exploration and plug-and-play answering models, enabling test-time scaling through additional small-model search and verification.
- **We introduce a search method that learning inspection priorities under sparse observations.** Rather than requiring each prediction to identify a confirmed evidence location, we explicitly supervise an ordered sequence of candidates supported by visible cues, separating search priority from evidence validity established through dense local verification.

We conduct systematic evaluations across multiple long-video understanding and reasoning benchmarks. Combining a 4B model for evidence acquisition with a frozen 27B model for final reasoning, our framework achieves 74.8% on VideoMME, 79.2% on MLVU, and 58.7% on LVBench, surpassing the standalone 27B model on all three benchmarks. Performance improves consistently as inference effort increases; at a lower inference effort, the framework reduces large-model calls by 60.2% on VideoMME while still exceeding the standalone baseline. Ablation studies further examine the effects of search supervision, candidate ordering, and local evidence verification. Together, these results show that small models can strengthen larger-model reasoning through evidence acquisition, extending their utility beyond the questions they can independently answer.

2 METHOD

2.1 OVERVIEW

Figure 2 illustrates our framework, in which a small model serves as the Router, Seeker, and Verifier, alongside a frozen large-model Reasoner. Given a question and sparse global frames, the Router either answers directly or defers to evidence acquisition. The Seeker proposes temporal intervals in descending search priority up to the query budget, and the Verifier examines densely sampled local frames to retain useful evidence, including partial clues that cannot independently support a complete answer. Retained clips are arranged chronologically and uniformly sampled under a fixed frame budget. The Reasoner then integrates the selected visual evidence and produces the final answer in a single invocation.

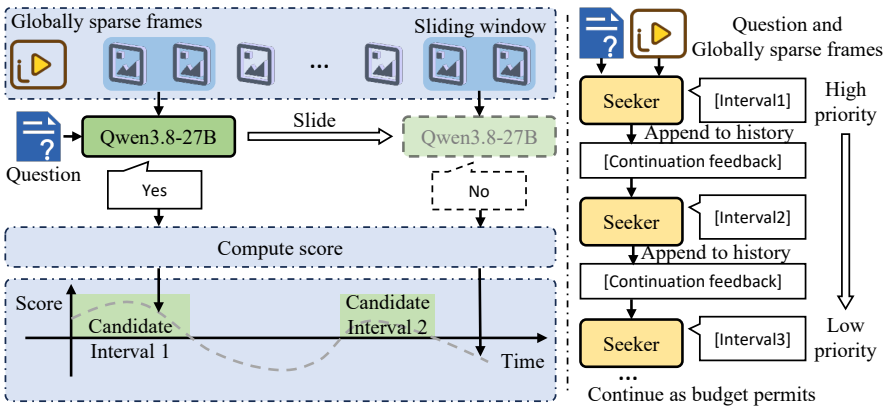


Figure 3: Learning and executing priority-ordered search. Left: Teacher scores over sliding windows of sparse frames yield a temporal priority profile for ranking candidate intervals. Right: Scripted continuation feedback prompts successive candidates in descending priority until the search budget is exhausted.

2.2 Selective Question Answering

The Router learns to answer or abstain from sparse visual observations. We train it on both full videos and local clips, using the small model's repeated-answer accuracy to construct supervision for SFT and subsequent RL. At inference time, it directly answers questions it can reliably handle from global observations and forwards the remainder to the collaborative pipeline. Detailed data construction and reward design are provided in Appendix C.2 and C.4.

2.3 Learning Search Priorities under Partial Observation

Under partial observation, the best-supported search decision need not identify the interval that actually contains the required evidence. An unsuccessful inspection therefore does not necessarily imply an unreasonable choice. Conversely, forcing the model to prioritize a ground-truth interval whose decisive details are absent from the sparse input may impose a ranking unsupported by the available observations. We therefore formulate temporal search as **priority-guided evidence search**: the Seeker learns to propose candidates according to their plausibility under the current observation, while their actual evidential value is determined through subsequent dense verification.

Constructing priority supervision. As illustrated in the left panel of Figure 3, we apply a teacher model to sliding windows of sparsely sampled video frames, conditioned on a question q. For each window O_j, the teacher judges whether the corresponding clip may contain key visual evidence, with its response restricted to `yes` or `no`. We compute

$$s_j = P_T(\texttt{yes} \mid O_j, q) - P_T(\texttt{no} \mid O_j, q), \tag{1}$$

and aggregate these window-level scores into a temporal priority profile $s(t)$. Candidate intervals are extracted from this profile and ranked by their peak scores:

$$S(\tau) = \max_{t \in \tau} s(t), \qquad S(\tau_1^*) \geq \cdots \geq S(\tau_M^*). \tag{2}$$

The resulting sequence $(\tau_1^*, \ldots, \tau_M^*)$ specifies an inspection order grounded in visible cues, rather than a chronological ordering or a set of confirmed evidence locations. Details of score aggregation and interval extraction are provided in the supplementary material.

Learning to continue the search. We organize the ordered candidates into a multi-turn interaction, as illustrated in the right panel of Figure 3. The Seeker receives the sparse video observation O and question q and predicts the first interval. Before each subsequent prediction, a feedback message states that the preceding candidate does not contain the required evidence and requests another interval. Let H_{i-1} denote the history of preceding candidates and feedback messages. We supervise

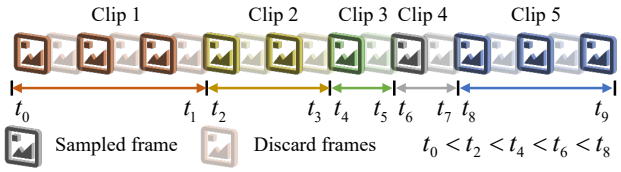


Figure 4: Visual input construction for the Reasoner. Retained clips are concatenated chronologically and uniformly sampled under a fixed frame budget.

successive responses with the ordered targets:

$$\mathcal{L}_{\text{search}} = -\sum_{i=1}^{M} \log p_\theta(\tau_i^* \mid O, q, H_{i-1}) \,, \tag{3}$$

where each interval likelihood factorizes over its response tokens. This objective teaches both the initial search preference and how to propose subsequent candidates when earlier suggestions are ruled out.

Search on demand. At inference time, continuation feedback requests the next candidate conditioned on the current observation and search history. It serves as a search-control instruction, while actual evidence validity is assessed separately by the Verifier. The resulting policy represents a conditional preference over where to inspect next, rather than a commitment to a fixed set of evidence locations. Its output is therefore an intermediate search decision that neither presupposes successful evidence acquisition nor requires a final answer. This separation allows search to remain open-ended and independent of the downstream reasoning process.

Further details on data construction, supervision, and interaction templates are provided in Appendix C.1.

2.4 Local Evidence Verification

The Verifier learns to assess evidence utility from dense local observations. We reuse the local clips from Router training, resample them densely, and obtain question-conditioned evidence labels from a teacher model for SFT. Its supervision concerns whether a clip provides useful evidence, rather than whether the small model can answer the question. At inference time, it filters Seeker-proposed intervals and passes the retained visual evidence to the Reasoner. Detailed data construction are provided in Appendix C.3.

2.5 Evidence-Based Reasoning

The Reasoner integrates the verified visual evidence to produce the final answer. As illustrated in Figure 4, we concatenate the frames from retained intervals in chronological order and uniformly sample the resulting sequence under a fixed frame budget. These frames, together with the question, are provided to the Reasoner in a single invocation. The reasoning backend is decoupled from the evidence acquisition pipeline and can be replaced by any compatible video-language model. All Reasoners used in this work are off-the-shelf models without fine-tuning.

2.6 Budget-Adaptive Inference

Our decoupled architecture allows evidence acquisition to be configured independently of final reasoning. Within this architecture, the Seeker's next-interval prediction task supports continued candidate generation and reuse across temporal scopes without requiring a complete question-answering trajectory. We exploit these properties to expand evidence acquisition through additional small-model search and verification. Figure 5 illustrates three configurations.

Low. The Router directly answers selected questions, while a single Seeker searches the full video for deferred instances using globally sparse observations.

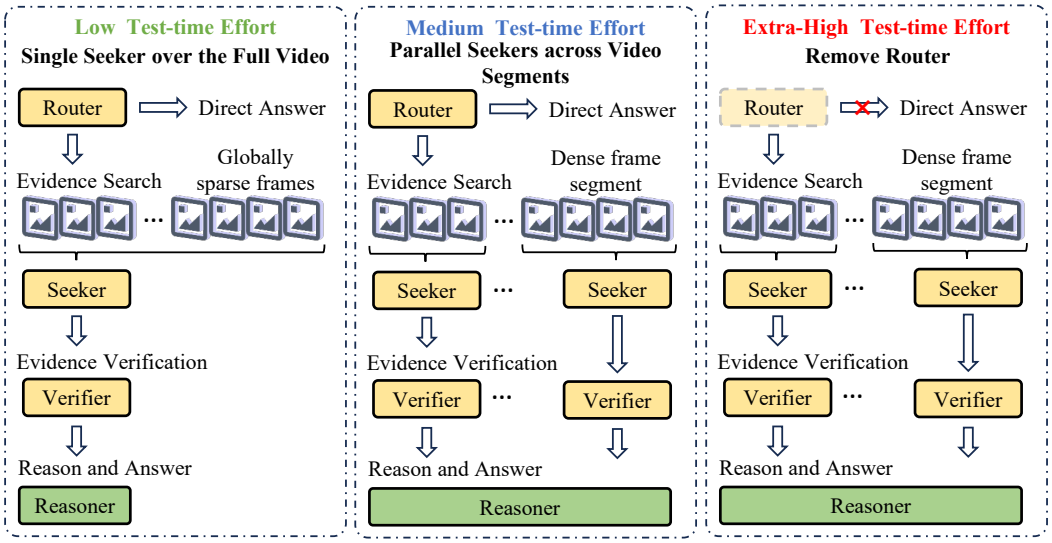


Figure 5: Three inference configurations with increasing test-time effort, differing in search scope and routing. The faded Router in Extra-High indicates that direct answering is disabled.

Medium. We retain the Router and partition the video into temporal segments searched independently by Seeker instances with shared parameters. Each instance applies the same maximum visual token budget to a shorter span, enabling denser sampling and greater overall frame coverage. Candidates are prioritized within each segment, helping preserve local cues that global sampling may miss and supporting finer localization.

Extra-High. We retain partitioned search but bypass the Router, sending every question through evidence acquisition and large-model reasoning.

For each instance entering the collaborative pipeline, the small model performs repeated search and verification, followed by a single Reasoner invocation under a fixed final-input frame budget. Independent branches can run in parallel to reduce search latency or with bounded concurrency to limit memory use. Evidence acquisition can thus scale without enlarging the per-call visual context or requiring all branches to reside in memory simultaneously.

3 Experiments

3.1 Implementation Details

Supervised Fine-Tuning. We adopt Qwen3-VL-4B-Instruct as the shared backbone for the Router, Seeker, and Verifier. The model is jointly fine-tuned on 55K examples: 20K general-purpose video understanding examples from LLaVA-Video-178K (Zhang et al., 2024b), covering multiple-choice QA, open-ended QA, and video captioning; and 35K task-specific examples constructed from CGBench (Chen et al., 2025) and LLaVA-Video-178K, comprising 10K routing examples, 12K temporal search examples, and 13K evidence verification examples. We perform multi-task supervised fine-tuning with LLaMA-Factory (Zheng et al., 2024), using a learning rate of 1×10^{-5} and a batch size of 256.

Reinforcement Learning. Starting from the jointly supervised checkpoint, we further optimize the model exclusively on the Router using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) implemented with verl (Sheng et al., 2024), on 3,200 training examples. We use a learning rate of 1×10^{-6} and a training batch size of 32, sampling 8 rollouts per input.

Evaluation. We use Qwen3-VL-4B-Instruct (Bai et al., 2025a) and Qwen3.8-27B (Qwen Team, 2026) as Reasoners, both without fine-tuning. During evaluation, all models share the same visual

Table 1: Performance on long-video understanding benchmarks. VideoMME is evaluated without subtitles; all scores are percentages. ⋆: frame limit per video input in an individual model call; †: overview/input frame limit, with additional local observations during exploration.

Model	Size	Frames	VideoMME		MLVU	LVBench
			Overall	Long	M-Avg	Overall
Existing methods						
Qwen2.5-VL (Bai et al., 2025b)	7B	768	65.1	–	70.2	45.3
Zoom-Zero (Shen et al., 2025)	7B	256	66.0	54.8	70.8	45.7
VITAL (Zhang et al., 2026b)	7B	1024†	64.1	54.0	–	–
ViTL (Wang et al., 2025a)	7B	128	–	–	62.3	47.4
Video-MTR (Xie et al., 2025)	7B	64	62.2	–	–	41.8
Conan (Ouyang et al., 2026)	7B	32	60.5	–	63.4	39.2
LongVT-RFT‡ (Yang et al., 2026a)	7B	64†	59.5	–	59.4	37.9
Video-o3 (Zeng et al., 2026)	7B	768†	66.5	–	72.1	47.6
VideoBrain (Zou et al., 2026)	8B	Adaptive	–	49.8	–	41.3
ParaVT (Yang et al., 2026b)	8B	64⋆	62.1	–	65.0	39.8
VideoSEAL (Qiu et al., 2026)	8B+7B+DS	64⋆	62.9	53.4	68.2	55.1
VideoSEAL w/o DeepSeek filter	8B+7B	64⋆	–	–	–	51.2
With a frozen 4B Reasoner						
Qwen3-VL-4B-Instruct	4B	64	61.5	49.4	62.6	33.6
Ours (Extra-High)	4B+4B	64⋆	**64.7** (+3.2)	**52.9** (+3.5)	**74.3** (+11.7)	**50.7** (+17.1)
Reasoner call reduction (%)			0.0	0.0	0.0	0.0
With a frozen 27B Reasoner						
Qwen3.8-27B (Low Think)	27B	64	68.8	58.8	65.6	43.4
Ours (Low)	4B+27B	64⋆	**67.7** (-1.1)	**56.9** (-1.9)	**70.1** (+4.5)	**54.2** (+10.8)
Reasoner call reduction (%)			60.2	53.0	42.9	26.3
Ours (Medium)	4B+27B	64⋆	**69.1** (+0.3)	**59.7** (+0.9)	**72.6** (+7.0)	**56.1** (+12.7)
Reasoner call reduction (%)			60.2	53.0	42.9	26.3
Ours (Extra-High)	4B+27B	64⋆	**74.8** (+6.0)	**66.2** (+7.4)	**79.2** (+13.6)	**58.7** (+15.3)
Reasoner call reduction (%)			0.0	0.0	0.0	0.0

input configuration: at most 64 frames, a maximum sampling rate of 2 FPS, and a maximum of $32 \times 32 \times 64$ pixels per frame. Before verification, the candidate interval budget is capped at 5 per video in *Low* mode and 8 per video in both *Medium* and *Extra-High* modes.

3.2 Main Results

Performance on video understanding benchmarks. Table 1 shows consistent gains from Low to Medium to Extra-High effort. With the 27B Reasoner, Extra-High reaches 74.8% on VideoMME (Fu et al., 2025), 79.2% on MLVU (Zhou et al., 2025), and 58.7% on LVBench (Wang et al., 2025b), improving over standalone inference by 6.0, 13.6, and 15.3 points; VideoMME Long improves by 7.4 points to 66.2%. Medium exceeds the baseline on every split while reducing Reasoner calls by 26.3–60.2%. Low retains these call savings, improving MLVU and LVBench with modest declines on VideoMME.

With a frozen 4B Reasoner, our 4B+4B configuration surpasses the 8B-based ParaVT by 2.6, 9.3, and 10.9 points on VideoMME, MLVU, and LVBench. It also outperforms default VideoSEAL on VideoMME Overall and MLVU, although it trails on VideoMME Long and LVBench. Its LVBench score of 50.7% approaches VideoSEAL's 51.2% without the DeepSeek-V3.2 filter, demonstrating competitive performance with a smaller answering model. All our Reasoners remain frozen, without any fine-tuning. So these improvements are achieved entirely through the evidence acquisition pipeline rather than adaptation of the answering model.

Performance across answering routes. Table 2 compares fixed Router-defined Direct and Deferred subsets. The Router only partitions the data: both collaborative variants use Extra-High evidence acquisition and frozen-Reasoner inference on every instance. We report relative error reduction against each corresponding standalone baseline as $(A_{\text{ours}} - A_{\text{base}})/(100 - A_{\text{base}}) \times 100\%$. MLVU scores are task-level macro averages within each subset; subset shares are sample-based.

Table 2: Performance on Router-defined subsets under Extra-High inference. Parentheses report relative error reduction (%) over the corresponding standalone model.

Method	VideoMME Overall			VideoMME Long			MLVU		
	Direct	Deferred	All	Direct	Deferred	All	Direct	Deferred	All
Subset share (%)	60.2	39.8	100	53.0	47.0	100	42.9	57.1	100
4B standalone	73.0	44.0	61.5	61.5	35.7	49.4	72.1	51.6	62.6
Ours (4B+4B)	74.2 (+4.4)	50.4 (+11.4)	64.7 (+8.3)	61.9 (+1.0)	42.8 (+11.0)	52.9 (+6.9)	80.7 (+30.8)	65.4 (+28.5)	74.3 (+31.3)
27B standalone	79.6	52.5	68.8	67.5	49.1	58.8	74.5	54.2	65.6
Ours (4B+27B)	82.0 (+11.8)	63.9 (+24.0)	74.8 (+19.2)	71.5 (+12.3)	60.2 (+21.8)	66.2 (+18.0)	83.7 (+36.1)	72.6 (+40.2)	79.2 (+39.5)

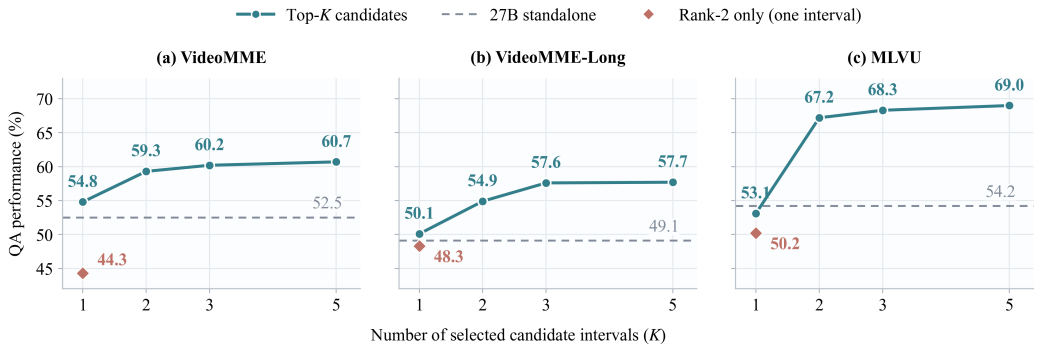


Figure 6: Candidate-budget scaling on Deferred subsets. Dashed lines denote standalone 27B performance; diamonds denote using only the second-ranked candidate.

Despite its higher baseline accuracy, the 27B Reasoner achieves greater relative error reduction than the 4B Reasoner on every subset, indicating that **stronger answering models benefit more from small-model evidence acquisition**. Improvements on both Direct and Deferred instances also show that this contribution extends beyond questions rejected by the Router.

MLVU and VideoMME Long further illustrate this complementarity. On MLVU, the 4B Reasoner gains 11.7 points, approaching the 27B model's 13.6-point gain: better evidence resolves many failures even with limited answering capacity. On VideoMME Long, the 27B model gains 7.4 points versus 3.5 for the 4B model, with relative error reduction increasing from 6.9% to 18.0%. This contrast is consistent with a stronger reasoning bottleneck on VideoMME Long, where useful evidence remains difficult for the small model to exploit. Small–large collaboration addresses both limitations through small-model evidence acquisition and stronger final reasoning.

3.3 Ablation Studies

All ablations are evaluated on fixed Router-deferred subsets, allowing us to examine evidence acquisition on questions that the small model does not answer directly. We use the same frozen 27B Reasoner and final visual input budget throughout, with experiment-specific settings detailed below.

Candidate search budget. We vary the number of candidate intervals under Low inference. Figure 6 shows that performance improves with continued search, with gains largely saturating after three candidates. Although the second-ranked candidate alone underperforms the first, combining them substantially improves accuracy, suggesting that lower-priority candidates provide complementary evidence and justify exploration beyond the initial interval.

Effectiveness of search priorities. Under Low inference, we compare the first two (ranks 1–2) and last two (ranks 4–5) candidates from the same five-interval sequence, with verification unchanged. Higher-priority candidates outperform lower-priority ones by 13.7, 7.8, and 23.6 percentage points on VideoMME, VideoMME Long, and MLVU, respectively (Table 3). They also consistently ex-

Table 3: Effectiveness of search priorities (%).

Configuration	VideoMME		MLVU
	Overall	Long	M-Avg
27B standalone	52.5	49.1	54.2
Top-2	**59.3**	**54.9**	**67.2**
Bottom-2	45.6	47.1	43.6

Table 4: Effect of evidence verification (%).

Configuration	VideoMME		MLVU
	Overall	Long	M-Avg
w/o Verifier	59.6	55.6	66.8
w/ Verifier	**60.7**	**57.7**	**69.0**
Gain (pp)	+1.1	+2.1	+2.2

ceed the standalone baseline, whereas lower-priority candidates fall below it, demonstrating that the learned priorities help allocate limited inspection budgets to more useful intervals.

Local evidence verification. Using the same top-five candidates under Low inference, we compare retaining all candidates with filtering them through the Verifier. Verification improves performance by 1.1–2.2 percentage points across the three evaluation sets (Table 4), supporting the separation of search priority from evidence validity: an interval worth inspecting need not provide useful evidence for the final answer.

Evidence Localization vs. Evidence Search. We compare learning to locate the true evidence with learning where to inspect when current observations are insufficient. We partition the Seeker dataset by evidence sufficiency under a fixed frame budget and construct training mixtures containing 0% or 50% evidence-missing instances. For each mixture, we independently retrain a Evidence Localization (EL) model and a Seeker using identical source QA examples, initialization, and training settings. Both predict one interval at evaluation ($K = 1$ for Seeker). We expand each EL interval to match its Seeker counterpart's duration before passing it to the shared answering pipeline, controlling for TG's typically narrower predictions.

As shown in Figure 7, the two paradigms perform similarly when training evidence is visible. Introducing 50% evidence-missing instances reduces EL model performance by 1.86 and 3.11 percentage points on VideoMME and MLVU, respectively, while improving Seeker performance by 0.94 and 1.93 points. This divergence highlights the importance of distinguishing evidence search from localization under partial observation. Requiring the model to locate evidence it cannot observe can introduce misleading supervision, consistent with EL's degradation. Seeker instead learns inspection priorities from available cues, turning these same instances into useful training for a central challenge of long-video understanding: searching for evidence before it has been fully observed.

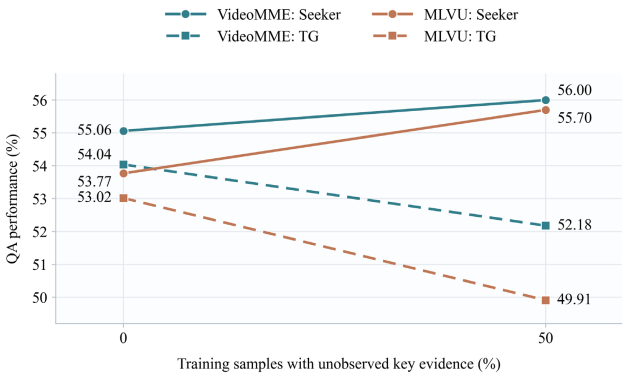


Figure 7: Evidence localization vs. evidence search.

4 Conclusion

We show that small models can contribute to long-video understanding beyond standalone answering by acquiring useful evidence for larger models. Our framework separates repeated evidence search and verification from final reasoning, enabling flexible allocation of test-time computation. We further distinguish evidence search under partial observation from direct evidence localization, training the Seeker to prioritize where to inspect based on visible cues. Together, these findings support organizing small–large collaboration around the different capability requirements of evidence acquisition and reasoning.

AI USE STATEMENT

We used Qwen3.8-27B to generate teacher annotations for search-priority and evidence-verification supervision, as detailed in Appendix C. We also used generative AI tools to assist with translation, manuscript editing, literature discovery, experimental design discussions, interpretation of results, figure preparation, and code review. The authors reviewed the AI-assisted content and take responsibility for the final manuscript, experimental results, and claims.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Chengsong Huang, Tong Zheng, Langlin Huang, Jinyuan Li, Haolin Liu, and Jiaxin Huang. Relayllm: Efficient reasoning via collaborative decoding. *arXiv preprint arXiv:2601.05167*, 2026.

Byeongchan Lee, Jonghoon Lee, Dongyoung Kim, Jaehyung Kim, Kyungjoon Park, Dongjun Lee, and Jinwoo Shin. Efficient llm collaboration via planning. *arXiv preprint arXiv:2506.11578*, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026.

Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, et al. Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35549–35561, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 34433–34448, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5503a7c69d48a2f86fc00b3dc09de686-Paper-Conference.pdf.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41089–41099, 2026.

Rui Pan, Yinwei Dai, Zhihao Zhang, Gabriele Oliaro, Zhihao Jia, and Ravi Netravali. Specreason: Fast and accurate inference-time compute via speculative reasoning. *Advances in Neural Information Processing Systems*, 38:12730–12749, 2026.

Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. Videoseal: Mitigating evidence misalignment in agentic long video understanding by decoupling answer authority. *arXiv preprint arXiv:2605.12571*, 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-zero: Reinforced coarse-to-fine video understanding via temporal zoom-in. *arXiv preprint arXiv:2512.14273*, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Chendong Wang, Donglin Bai, Yifan Yang, Xiao Jin, Anlan Zhang, Rui Wang, Shiqi Jiang, Yuqing Yang, Hao Wu, Qi Dai, et al. Video-in-the-loop: Span-grounded long video qa with interleaved reasoning. *arXiv preprint arXiv:2510.04022*, 2025a.

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

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478*, 2025.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026a.

Zuhao Yang, Kaichen Zhang, Sudong Wang, Keming Wu, Zhongyu Yang, Bo Li, Xiaojuan Qi, Shijian Lu, Xingxuan Li, and Lidong Bing. Paravt: Taming the tool prior paradox for parallel tool use in agentic video reinforcement learning. *arXiv preprint arXiv:2605.20342*, 2026b.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026.

Chuang Zhang, Zizhen Zhu, Yihao Wei, Bing Tian, Junyi Liu, Henan Wang, Wang Xavier, and Yaxiao Liu. Confidence-calibrated small-large language model collaboration for cost-efficient reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4480–4501, 2026a.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026b.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026c.

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

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024b. URL https://arxiv.org/abs/2410.02713.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

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

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, et al. Timelens2: Generalist video temporal grounding with multimodal llms. *arXiv preprint arXiv:2607.17423*, 2026.

Junbo Zou, Ziheng Huang, Shengjie Zhang, Liwen Zhang, and Weining Shen. Videobrain: Learning adaptive frame sampling for video understanding. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

A Appendix

This appendix provides additional details and analyses:

- **Related work** (Section B): small–large model collaboration and agentic video understanding.
- **Data construction and supervision** (Section C): data sources, annotation procedures, and supervision for the Router, Seeker, and Verifier.
- **Implementation details** (Section D): inference configurations and computational cost accounting.
- **Additional experimental results** (Section E): efficiency analysis and experimental details for comparing evidence localization with evidence search.
- **Qualitative analysis** (Section F): continued small-model contributions after deferral, complementary evidence acquisition, and differences in final reasoning capability.
- **Limitations** (Section G): limitations of task allocation, search supervision, and local evidence verification.

B Related Work

Small–Large Model Collaboration. Small–large model collaboration exploits differences in model capability and computational cost to improve inference quality and efficiency. One line of work allocates computation at the instance level. RouteLLM (Ong et al., 2025) learns query-to-model routing policies from preference data, while COREA (Zhang et al., 2026a) calibrates small-model answer confidence and delegates low-confidence instances to a larger model. These approaches optimize model selection and escalation to allocate computational resources according to instance requirements. Another line of work introduces collaboration within individual problem-solving processes. COPE (Lee et al., 2025) uses plans as intermediate representations to coordinate planning and execution across model sizes; RelayLLM (Huang et al., 2026) allows a small model to request that a larger model temporarily take over token generation; and SpecReason (Pan et al., 2026) uses a small model to generate intermediate reasoning steps, which a larger model evaluates and revises when necessary. Together, these studies suggest that standalone problem-solving ability does not fully characterize a model's collaborative value. Our work further identifies different capability requirements for evidence acquisition and evidence reasoning in long-video understanding: under active visual exploration, small models with limited standalone QA performance can still effectively search for and verify relevant evidence. A capability gap in final reasoning therefore need not extend to evidence acquisition.

Agentic Video Understanding. The visual token costs of long videos constrain observation under limited context and GPU memory budgets, motivating reasoning paradigms that acquire visual information on demand. Thinking With Videos (VITAL) (Zhang et al., 2026b) uses video cropping to inspect local details; LongVT (Yang et al., 2026a) introduces self-correction through global browsing and local inspection; and Video-o3 (Zeng et al., 2026) interleaves clue seeking with multi-hop reasoning. ParaVT and Video-MTR improve exploration through parallel interval inspection and turn-level retrieval rewards, respectively (Yang et al., 2026b; Xie et al., 2025), while VideoSEAL combines indexing, retrieval, and visual inspection to separate search planning from answer authority (Qiu et al., 2026). These approaches extend evidence acquisition to multiple rounds and intervals, increasing the cost of repeated visual processing before the final answer. However, search and verification need not require the same model capacity as final reasoning. We accordingly decouple evidence search, local verification, and final answering, assigning repeated exploration to a small model and evidence integration to a large model. Independently configurable search scopes and candidate budgets enable flexible test-time scaling: additional small-model computation broadens evidence coverage and refines observations while keeping the visual token limit per call fixed.

Furthermore, multi-round exploration and multi-interval prediction do not necessarily align search supervision with the information available under sparse observations. LongVT (Yang et al., 2026a) uses video length to determine the probability of constructing multi-round training trajectories; Video-o3 (Zeng et al., 2026) synthesizes multi-hop trajectories around previously localized and verified evidence; and ParaVT (Yang et al., 2026b) combines distilled trajectories with multi-segment grounding examples from MUSEG (Luo et al., 2026). These designs expand interaction depth or localization targets without explicitly supervising the inspection order justified by the current observations. Our method instead learns observation-conditioned search priorities, separating candidate proposal from evidence verification through subsequent dense inspection.

C Data Construction and Supervision

We use Qwen3.8-27B as the teacher model for all teacher-assisted data construction.

C.1 Seeker Data Construction

We construct ordered search supervision from teacher assessments of sparse visual observations. The resulting candidate intervals also provide the local clips used for Router and Verifier data construction.

Observation-aligned teacher scoring. Each video is sampled using the same configuration as training, with a maximum of 64 frames and a sampling rate capped at 2 FPS. We apply sliding windows of four sampled frames with a stride of one frame, ensuring that the teacher evaluates only visual details available in the student's input. For each window, the teacher judges whether the continuous interval covered by its sparse frames is likely to contain key visual evidence for the question. The judgment concerns the potential relevance of the covered interval, rather than whether the displayed frames alone suffice to answer. We score each window by the probability difference between the `yes` and `no` tokens, then average the scores of overlapping windows at each sampled timestamp to obtain a temporal priority profile. The teacher prompt is provided below.

Teacher prompt for search-priority supervision

```
You are given a sparse set of frames sampled from a continuous video
    interval.

Question:
{question}

Task:
Decide whether the continuous interval covered by these sparse
    frames is likely to contain
key visual evidence for answering the question.
```

```
Important:

- The shown frames are only sparse observations of a longer
    continuous interval.
- Some frames inside the interval are missing and are not shown.
- Do NOT judge only whether the displayed frames themselves already
    contain the final answer.
- Judge whether the full interval covered by these sparse
    observations is likely to contain relevant evidence.

Answer "yes" only when there is reasonably strong and discriminative
    visual evidence that this interval is relevant, such as:

- the shown frames already contain the crucial visual evidence
- the shown frames contain a question-specific action, event, state
    change, interaction, or object attribute
- the shown frames distinguish between possible answers
- the shown frames strongly suggest that a crucial visual moment is
    unfolding within this interval
- the interval contains more than just common objects or scenes
    mentioned in the question

Answer "no" when:

- the shown content only contains frequently appearing objects,
    people, scenes, or interfaces
- the shown content is only weakly, vaguely, or generically related
    to the question
- the relevance depends mostly on speculation rather than visible
    cues
- the scene is too generic without question-specific evidence
- the interval may be related only in a broad semantic sense, but
    there is no clear visual grounding
- the interval shows the general topic of the question but not the
    visual clue needed to answer it

Decision rule:

- Judge the whole covered interval, not only the displayed frames.
- Be confident.
- If the relevance is ambiguous, weak, generic, or speculative,
    answer "no".
- Only answer "yes" when the visual match to the question is fairly
    specific and convincing.

Output exactly one word:
yes or no
```

We compute each window's score as the difference between the probabilities of the `yes` and `no` output tokens. Each sampled timestamp receives the average score of the windows covering it, yielding a temporal priority profile over the sparse observations.

Interval extraction and ranking. We compute the mean of the temporal priority profile and set scores below this mean to -1. Contiguous regions surviving this threshold form candidate intervals. Each candidate is assigned its peak score, and up to five intervals are retained in descending order. These ordered intervals provide successive targets for the multi-turn search interaction described in Section 2.3.

Figure 8 illustrates how visible cues inform the ordering. The presence of the queried person alone provides limited support, while a subtitle matching an answer option offers a more specific connection. The strongest cues combine the person holding a "NEVER" card with the subtitle corresponding to option A. The highest-ranked candidate, $[114, 134]$ seconds, has a peak score of 0.90 and covers the annotated evidence interval $[125, 127]$ seconds. Its broader boundaries reflect the pur-



Moderate relevance: The person matches the question, and the subtitle matches option C, but no "NEVER" card is visible.
Low relevance: The person matches the question, but no question-specific action or text is visible.
High relevance: The person holds a "NEVER" card while the subtitle matches option A.

Figure 8: Example of Seeker supervision construction. (a) The thresholded temporal priority profile derived from teacher judgments under sparse observations. (b) Five candidate intervals ranked by peak score, with the annotated evidence interval shown for reference. Representative frames illustrate different levels of question-specific visual relevance.

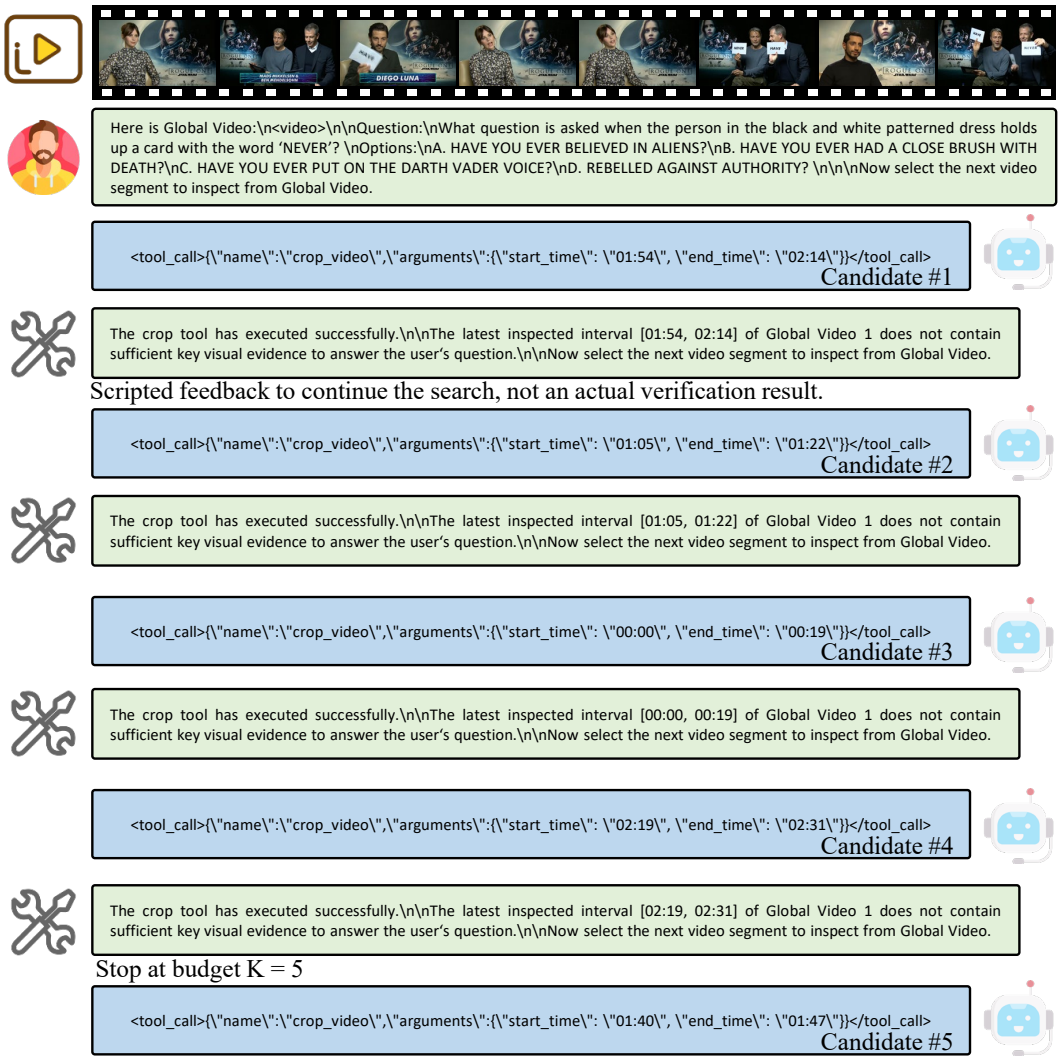


Figure 9: Multi-turn supervision for ordered interval prediction. Given sparse video observations and a question, the Seeker is trained to generate candidate intervals in descending search priority. Scripted continuation messages elicit successive candidates; they serve as search-control instructions rather than actual verification results. At inference time, the interaction can stop after the requested number of candidates, allowing the search depth to adapt to the available budget.

pose of the supervision: identifying a promising region for further inspection, rather than precisely reproducing the evidence boundaries.

Annotation requirements and applicability. The procedure requires video-question pairs but no manually annotated temporal boundaries or predefined question-type labels. It applies the same question-conditioned relevance criterion across tasks, producing search priorities grounded in the available sparse observations rather than treating every selected interval as confirmed evidence.

Constructing multi-turn search dialogues. We convert the ranked intervals into multi-turn training dialogues, as illustrated in Figure 9. The initial input contains the sparsely sampled video and question, and each assistant response specifies the next interval through a `crop_video` call. Between successive responses, we insert a scripted message that rules out the preceding candidate and requests another interval. This message advances the search sequence rather than reporting an actual verification result, so it is applied even when a candidate overlaps the ground-truth evidence. The resulting dialogue supervises the model to move through candidates in descending priority while

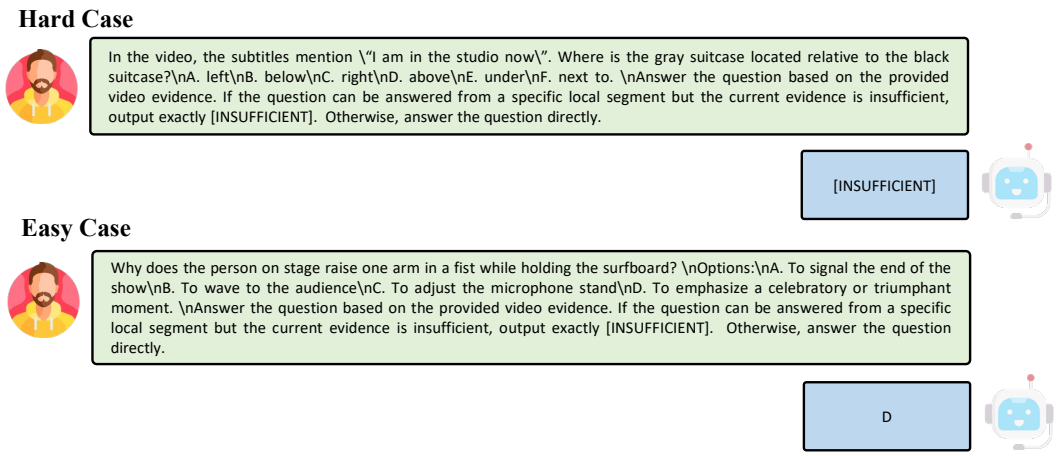


Figure 10: Router training examples with abstention (top) and direct-answer (bottom) targets, assigned according to the small model's correctness over eight sampled responses.

conditioning on its previous proposals. Although each training dialogue contains up to five candidates, the interaction does not require a predefined output count: at inference time, the system requests successive candidates until the search budget is reached.

C.2 Router Data Construction

We construct Router training data from a mixture of full videos and local clips, with the latter drawn from the candidate intervals extracted during Seeker data construction. Both are sparsely sampled to reflect the Router's observation conditions. For each video–question pair, we sample eight responses from the base small model and compute their correctness rate against the reference answer. We then apply predefined thresholds to assign direct-answer or abstention labels. Instances with at most three correct responses out of eight are labeled for abstention; those with four or more correct responses are labeled for direct answering.

For supervised fine-tuning, direct-answer instances use the reference answer as the target, whereas abstention instances use `[INSUFFICIENT]`, as illustrated in Figure 10. The same labeling procedure also supplies the instance categories used in Router reinforcement learning. These labels reflect the small model's ability to answer under the supplied observations: abstention may arise from missing evidence or an inability to reason reliably from the evidence available.

C.3 Verifier Data Construction

We construct Verifier data using local intervals extracted during Seeker data construction. Each interval is sampled densely and paired with its question for teacher annotation. The teacher checks which question-relevant facts are visually supported and assigns either `[USEFUL_EVIDENCE]` or `[NO_USEFUL_EVIDENCE]`. Useful evidence need not support a complete answer: confirmed partial information or evidence that eliminates a candidate answer also qualifies, whereas generic relevance and unverified interpretations do not. The resulting labels supervise the small model to assess local evidence utility, as illustrated in Figure 11. Unlike Router labels, which reflect the small model's empirical answering performance, Verifier labels reflect the teacher's assessment of the observed visual content. The Verifier is trained through supervised fine-tuning only. An English translation of the teacher prompt is provided below.

Teacher Prompt for Evidence Verification

```
Verify the key facts actually visible in the provided video interval
and assess whether they offer useful visual evidence for the
following question.
```

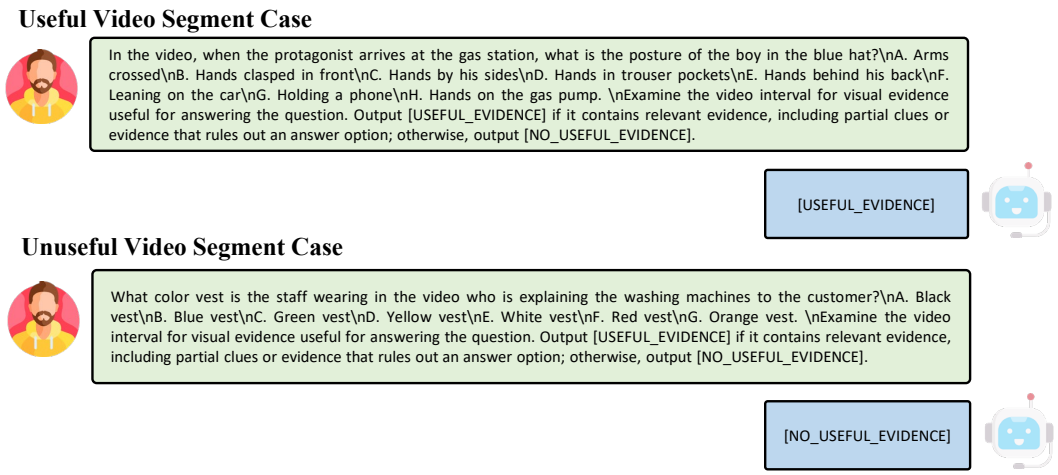


Figure 11: Verifier training examples with useful-evidence (top) and no-useful-evidence (bottom) labels, assigned by the teacher through dense local observation.

```
Question: {question}

Organize your analysis as follows: first, briefly identify the key
facts that need verification, without assuming an answer or that
these facts occur in the interval; then, concisely describe the
relevant visible content, checking whether the entities, actions,
relationships, or events correspond to those facts; finally, review
what has and has not been confirmed and assess the usefulness of the
evidence.

Follow these rules:

    1. The video may lack some or all of the necessary information.
       Collect the evidence that is actually visible and reason from
       it without adding unseen details.

    2. The question and answer options specify what needs
       verification; they are not sources of facts about the video.
       Do not guess an answer and then reinterpret similar scenes to
       support it. Do not fill gaps in visual evidence using common
       sense, external knowledge, or assumptions about the narrative.

    3. Identify people, objects, and actions using visible
       distinguishing features. Similar appearance, color,
       position, or setting alone does not establish that an
       observed entity is the one referenced in the question. If
       a crucial detail is unclear, identity cannot be established,
       or multiple interpretations remain possible, acknowledge
       the uncertainty rather than presenting a resemblance as a
       confirmed observation.

    4. The presence of people or objects does not establish that the
       required event or relationship occurred. Verify the actions,
       participant roles, and relevant relationships. For example,
       two people appearing together, approaching each other, or
       holding an object does not by itself establish a handover.
       Do not reinterpret similar actions as the event described
       in the question or invent unseen action sequences or causal
       relationships.

    5. The interval need not independently support a complete answer.
       Reliable complete evidence, useful partial evidence, and
       evidence that rules out even one answer option all count as
```

```
        useful. Uncertainty about some details does not invalidate
        other confirmed and useful facts. Assign a negative label
        only when no such fact can be confirmed. Failure to observe
        an event in this interval does not establish that it never
        occurred in the full video.
    6. In your final assessment, consider whether the key facts are
        actually visible, whether they address the question, and
        whether their interpretation still depends on unverified
        assumptions. Prior knowledge cannot replace visual
        verification and must not be presented as an observation. Do
        not provide the final answer to the question or mechanically
        repeat the preceding analysis.
    7. Keep the analysis concise, typically reaching a conclusion
        in one or two sentences. Omit irrelevant scenes and avoid
        frame-by-frame descriptions, full plot summaries, exhaustive
        possibilities, or section headings.
Return a JSON object with the analysis in English and the response
restricted to one of the following labels:

[USEFUL_EVIDENCE]: At least one confirmed visual fact materially
helps answer the question, including complete, partial, or
eliminative evidence.

[NO_USEFUL_EVIDENCE]: No such visual fact is confirmed. Similar
entities, vague cues, or imagined events are insufficient.

{"analysis": "ANALYSIS_CONTENT", "response": "RESPONSE_TOKEN"}
```

C.4 Training Procedure and Hyperparameters

Joint supervised fine-tuning. We initialize the small model from Qwen3-VL-4B-Instruct and jointly train its Router, Seeker, and Verifier roles on 54,728 examples: 20,000 general video understanding examples from LLaVA-Video-178K, covering multiple-choice QA, open-ended QA, and video captioning; and 34,728 task-specific examples constructed from CGBench and LLaVA-Video-178K, comprising 9,773 routing examples, 12,084 temporal search examples, and 12,871 evidence verification examples. All three roles share the same model parameters and are distinguished by their prompts. We perform supervised fine-tuning using LLaMA-Factory with a learning rate of 10^{-5} and a global batch size of 256.

Router reinforcement learning. Following SFT, we perform reinforcement learning exclusively on the Router task using 3,200 examples, with a batch size of 32, eight rollouts per input, and a learning rate of 10^{-6}. Rewards follow the instance labels established during data construction: for direct-answer instances, only a correct answer receives a positive reward; for abstention instances, only `[INSUFFICIENT]` receives a positive reward, even if a generated answer happens to be correct. All other responses receive zero reward. Seeker and Verifier tasks receive no additional RL supervision, and all Reasoner models remain frozen throughout training. Table 5 summarizes the configurations.

D Implementation Details

D.1 Inference Configurations

We implement three inference configurations, Low, Medium, and Extra-High, which differ in routing and search scope. All configurations use the same trained models. Table 6 summarizes their settings.

Low effort. The Router first attempts to answer the question from a sparse global observation. Accepted answers are returned directly, while abstained instances proceed to evidence acquisition. A single Seeker operates over the full video and generates up to five candidate intervals through suc-

Table 5: Training configurations for joint supervised fine-tuning and Router reinforcement learning.

Parameter	SFT	Router RL
Training data		
General video understanding examples	20,000	–
Router examples	9,773	3,200
Seeker examples	12,084	–
Verifier examples	12,871	–
Total examples	54,728	3,200
Optimization		
Learning rate	10^{-5}	10^{-6}
Global batch size	256	32
Training epochs	1	1
Optimizer	AdamW	AdamW
Learning rate scheduler	Cosine	Constant
Warmup ratio	0.03	0
Weight decay	0	0.01
RL and rollout settings		
Policy optimization algorithm	–	GRPO
Rollouts per input	–	8
Sampling temperature	–	0.7
KL coefficient	–	0.01
Training infrastructure		
Training framework	LLaMA-Factory	verl
GPU type and count	4× NVIDIA H200	4× NVIDIA H200

cessive continuation requests. The Verifier examines densely sampled frames from each candidate, and the retained visual evidence is passed to the Reasoner for a single final answering call.

Medium effort. The Router follows the same procedure as in Low effort. For abstained instances, the number of temporal partitions depends on the video duration: one partition for videos shorter than 6 minutes, two for videos between 6 and 12 minutes, and four for videos longer than 12 minutes. Each partition is searched independently under the same per-call visual budget, enabling finer temporal sampling for longer videos. The total candidate budget is capped at eight per video across all partitions, before verification. Verified evidence is aggregated for a single Reasoner call.

Extra-High effort. This configuration bypasses the Router and applies the partitioned search and verification procedure to every question. It uses the same eight-candidate budget as Medium effort, but allocates evidence acquisition and final Reasoner computation to all instances, including those that the Router would otherwise answer directly.

E Additional Experimental Results

E.1 Efficiency Analysis

We examine the quality–cost trade-offs of the 4B+27B framework using a parameter-weighted video-input proxy. For each stage, we count the number of video-input sessions and weight it by the corresponding model size:

$$C = \frac{4(V_{\text{Router}} + V_{\text{Seeker}} + V_{\text{Verifier}}) + 27V_{\text{Reasoner}}}{27N}, \tag{4}$$

where N is the number of evaluated questions and V denotes the video-input count for each module. Standalone 27B inference therefore has a normalized cost of 1. A Seeker session may generate multiple candidate intervals from the same video input; these continuation turns are not counted as additional video inputs. This proxy summarizes model-size-weighted visual-input workload,

Table 6: Inference configurations. Candidate budgets are upper limits per video, summed across all searches and measured before verification. Visual input limits apply to each model call.

Setting	Low	Medium	Extra-High
Router	Enabled	Enabled	Bypassed
Search scope	Full video	Segments	Segments
Number of video partitions	1	Duration-dependent	Duration-dependent
Candidate budget per video	5	8	8
Evidence verification	Enabled	Enabled	Enabled
Maximum frames per call	64	64	64
Maximum sampling rate (FPS)	2	2	2
Maximum pixels per frame	65,536	65,536	65,536
Reasoner calls per searched instance	1	1	1

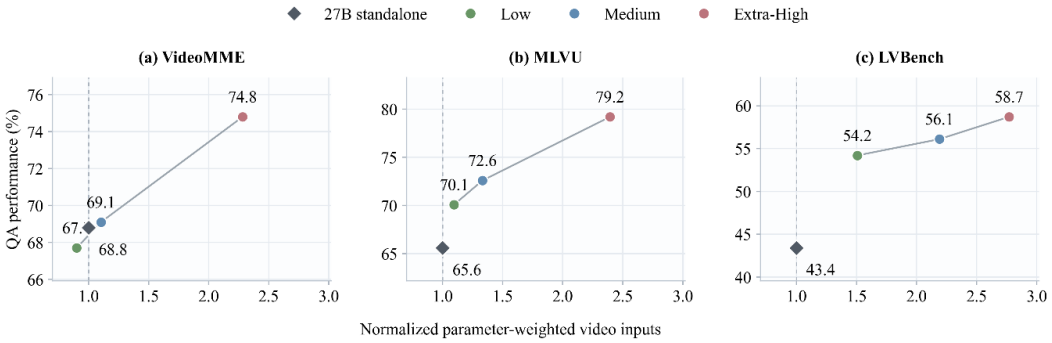


Figure 12: Quality–cost trade-offs of the 4B+27B framework. Cost is measured by parameter-weighted video-input sessions, normalized to standalone 27B inference on each benchmark. All configurations are evaluated on the full evaluation sets.

rather than total computation or latency. It does not account for differences in actual input lengths, continuation and output tokens, caching, or video preprocessing. Costs are averaged per question, while MLVU performance follows the task-level M-Avg metric.

The Qwen3.8-27B Reasoner operates in Low Think mode with a maximum generation budget of 8,192 tokens, whereas the fine-tuned small-model modules directly emit task outputs without explicit chains of thought. Our proxy excludes autoregressive generation costs, including the Reasoner's additional reasoning tokens. The stage-wise values thus represent visual-input workload allocations rather than total computational cost; the generation budget is an upper bound, not the actual output length.

Quality–cost trade-offs. Figure 12 shows that increasing inference effort consistently improves performance across all three benchmarks, with different resource requirements. On VideoMME, Low operates at $0.90\times$ the baseline proxy cost with a 1.1-point accuracy decrease, whereas Medium surpasses the baseline at $1.10\times$ cost. Medium also improves MLVU and LVBench by 7.0 and 12.7 points at $1.33\times$ and $2.19\times$ cost, respectively. Extra-High achieves the strongest results, improving over standalone 27B inference by 6.0, 13.6, and 15.3 points at normalized costs of 2.28, 2.39, and 2.77. The configurations thus offer different operating points, with higher accuracy obtained through additional evidence acquisition and, when routing is bypassed, more frequent large-model reasoning.

Stage-wise resource allocation. Figure 13 separates the proxy cost into routing, search, verification, and final reasoning. Low and Medium invoke the Reasoner on the same questions; their cost difference comes entirely from additional small-model search and verification. The corresponding accuracy gains therefore demonstrate the value of increasing evidence acquisition while keeping the number of large-model invocations fixed. Verification accounts for the largest share of small-model

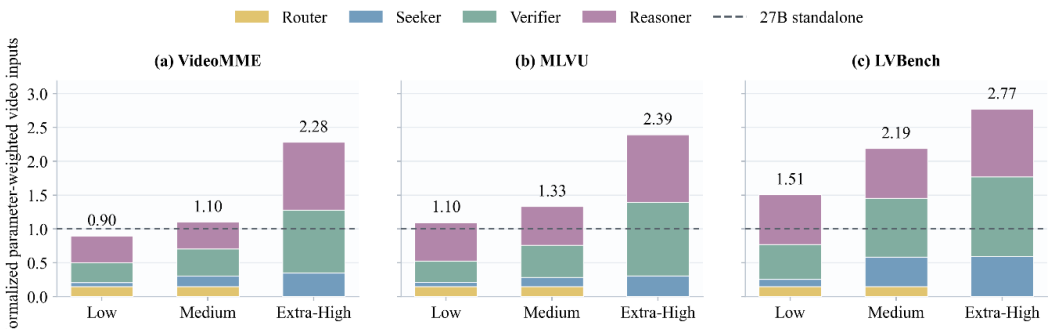


Figure 13: Stage-wise decomposition of the normalized video-input cost proxy. Stacked bars show contributions from the four modules; the dashed line denotes standalone 27B inference. Values above the bars indicate the total proxy cost.

visual-input workload in every configuration, reflecting repeated dense inspection of candidate intervals. Extra-High removes routing and extends evidence acquisition and final reasoning to all questions, so its additional cost includes both small-model exploration and large-model answering. Across configurations, each question entering the collaborative pipeline requires only one final Reasoner invocation.

E.2 Experimental Details for Evidence Localization vs. Evidence Search

We partition the training data using the temporal scores produced during Seeker data construction. Specifically, we label an instance as evidence-missing if the smallest interval covering its ground-truth evidence has a score of -1, and as evidence-visible otherwise. These labels reflect evidence availability under the fixed sparse observation budget: for every instance, we verify that the model can answer correctly when the ground-truth interval is sampled sufficiently densely, ensuring that the question is answerable given adequate visual evidence. We construct training mixtures with 0% or 50% evidence-missing instances, keeping each training set fixed at 3,200 examples. For each mixture, the temporal grounding model and Seeker are independently trained from the same base-model initialization on the same source QA examples, with a batch size of 64. Seeker retains its multi-turn priority supervision during training, while evaluation restricts both models to a single predicted interval.

F Qualitative Analysis

We present two examples illustrating how small-model-guided evidence acquisition supports questions that the small model cannot reliably answer, and how search priority differs from the evidential value established through local inspection.

Continued contribution after deferral. Figure 14 illustrates a question asking which action a magician does not perform. Although relevant visual cues are already present in the sparse overview, the Router defers the question. The small model nevertheless continues to contribute: the Seeker proposes candidate intervals, and the Verifier retains useful observations while discarding an irrelevant segment. The retained evidence is then consolidated for the large Reasoner, which selects option B. This example illustrates that deferral does not necessarily indicate an absence of relevant visual information, nor does it imply that the small model has exhausted its utility. Its contribution continues through evidence acquisition even when it declines to provide the final answer.

Search priority and complementary evidence. Figure 15 asks whether a judge is to the performer's left or right from the performer's perspective. Under sparse observation, the first candidate is more strongly supported by visible cues because both the performer and the judge are identifiable. Dense inspection confirms the presence of the judge, ruling out option D, but does not resolve the requested spatial relation. The Verifier therefore retains this partial evidence. Continuing to the sec-

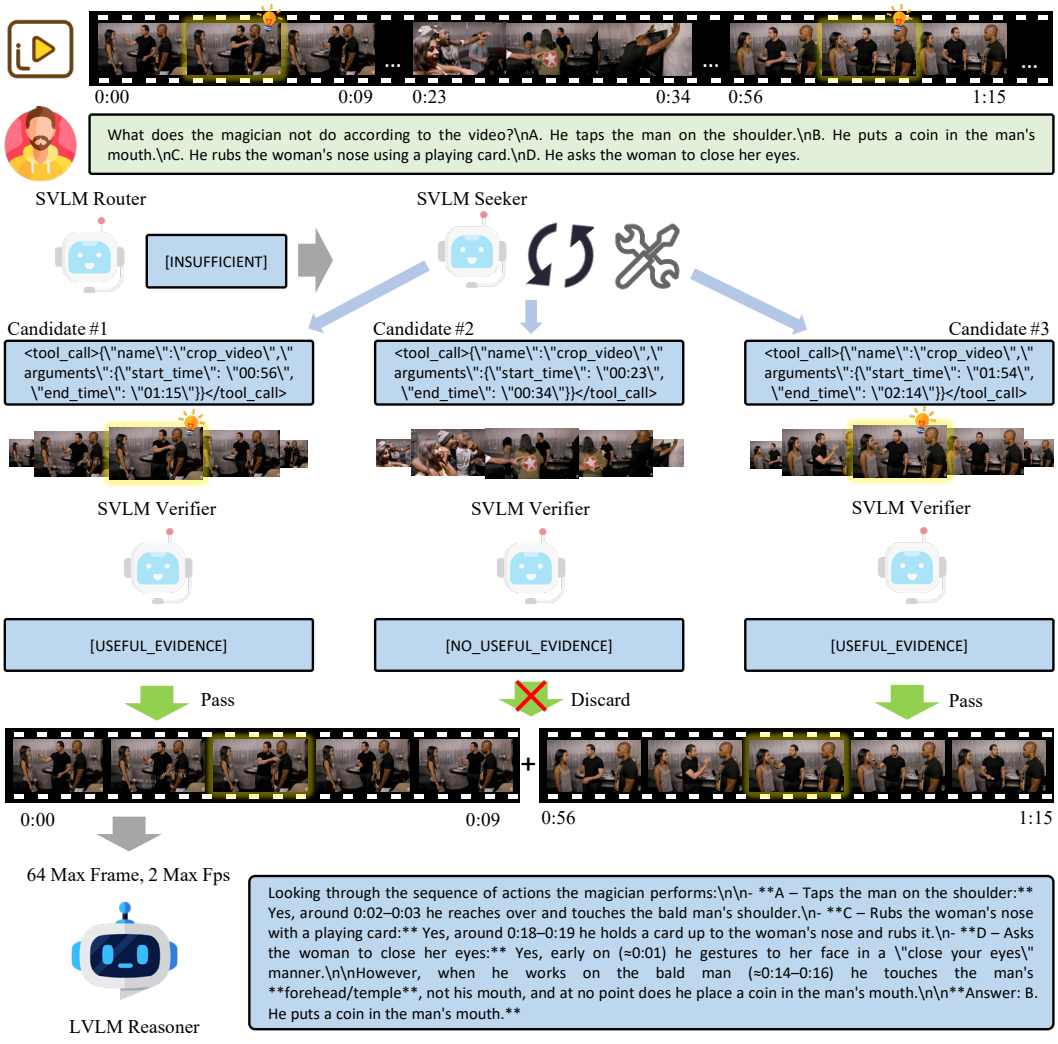


Figure 14: Continued small-model contribution after Router deferral. Despite declining to answer from the sparse overview, the small model searches for and verifies useful evidence, enabling the large model to produce the correct answer.

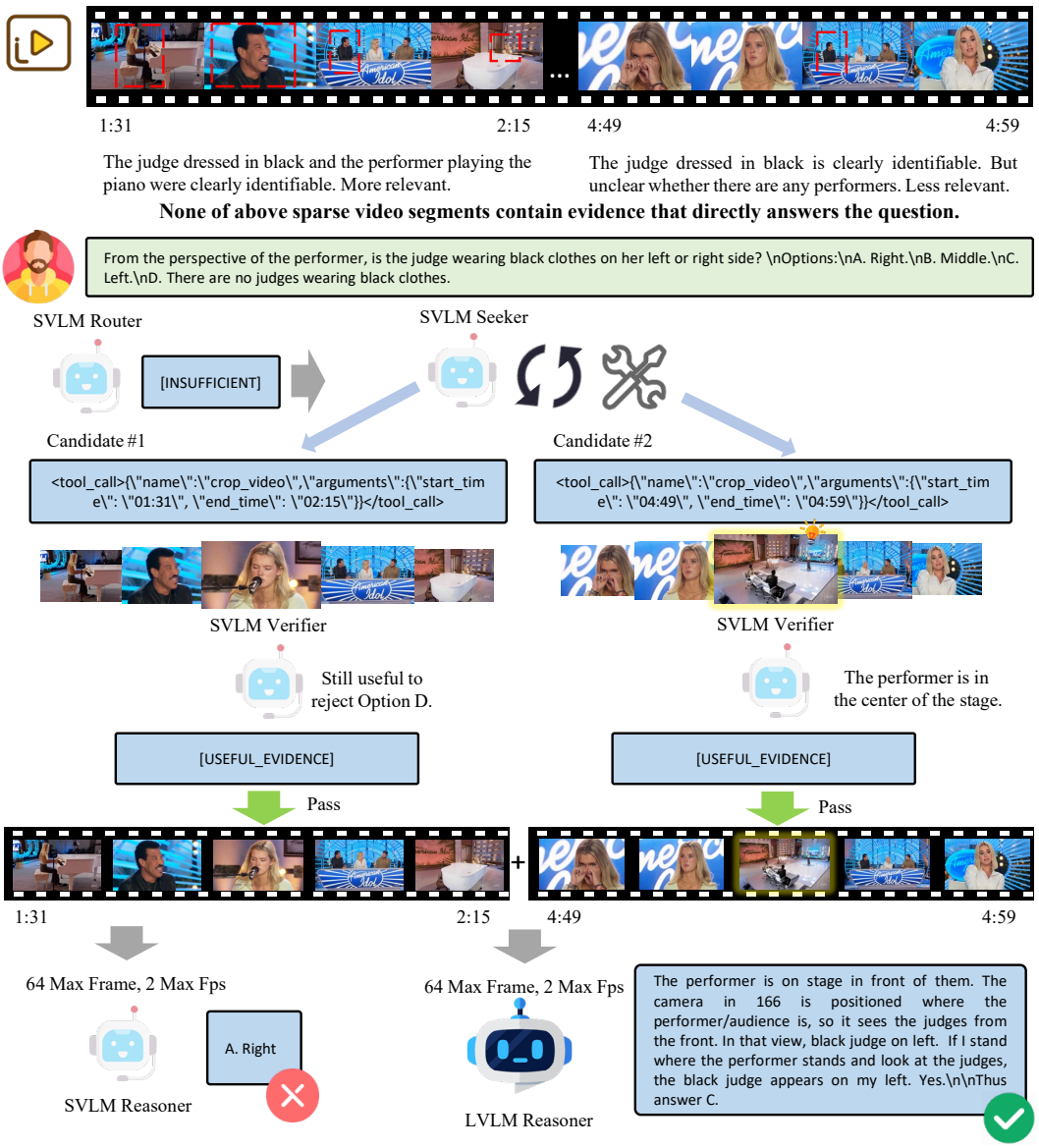


Figure 15: Complementary evidence from successive search candidates. The first interval provides useful partial evidence, while the second reveals additional spatial context. Given the same aggregated evidence, the small model answers incorrectly, whereas the large model resolves the spatial relation correctly.

ond, initially less salient candidate reveals a wider view that helps establish the spatial arrangement. This case illustrates why search should extend beyond the most promising initial candidate: priority reflects the available observations, whereas decisive details may emerge only through subsequent inspection. It also demonstrates that evidence verification should preserve useful partial observations without requiring each interval to independently support a complete answer.

Evidence acquisition beyond answering capability. The same example further illustrates the distinction between acquiring evidence and reasoning over it. Given the same aggregated evidence under the same frame budget, the small Reasoner incorrectly selects "Right," whereas the large Reasoner selects the correct answer, "Left." The small model can thus identify and retain useful visual evidence while still failing to resolve the perspective-dependent spatial relation. This contrast illustrates the motivation for our collaboration framework: the small model improves the evidence

available for reasoning, while the large model supplies the capability needed to interpret that evidence and complete the answer.

G LIMITATIONS

Our routing policy does not fully separate evidence availability from reasoning difficulty. Ideally, allocation would distinguish questions that a small model can or cannot answer, both before and after acquiring additional evidence. The current Router primarily assesses the initial observation, without explicitly predicting whether further exploration would enable small-model answering. This may cause unnecessary search when reasoning is the main bottleneck, or unnecessary large-model calls when additional evidence would suffice.

Search supervision and adaptation also remain limited. Candidate priorities are derived from a 27B teacher's relevance judgments, which may overlook subtle cues or introduce systematic ranking errors. Moreover, scripted continuation feedback advances the candidate sequence without conveying newly verified visual facts. Although this supports budget-controlled exploration, it does not fully exploit local discoveries to revise subsequent search priorities.

Finally, evidence usefulness can depend on relationships across intervals. A locally assessed clip may appear uninformative yet become valuable when combined with another observation. Teacher-supervised verification may therefore discard complementary evidence or retain redundant content. Improving the individual modules remains important, but further gains may also require routing, search, and verification to account more explicitly for the evidence already acquired and the capabilities of the final answering model.