VIDEOSEEKER: EMPOWERING CROSS-VIDEO UNDERSTANDING VIA MULTI-TURN EVIDENCE-SEEKING AGENTIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-video understanding (CVU) requires models to locate key evidence across videos, establish temporal correspondences, and integrate intermediate findings through explicit comparison. Evidence needs become clearer as reasoning unfolds, yet end-to-end methods that answer directly from fixed frame samples lack mechanisms to supplement and verify evidence based on intermediate judgments. Under limited context budgets, sparse sampling may miss brief yet critical events, while dense encoding of entire videos scales poorly to multiple long videos. To this end, we formulate CVU as multi-turn agentic reasoning driven by active evidence seeking, integrating targeted segment replay, temporally interleaved comparison, and persistent evidence memory into a closed loop of observation, comparison, and reasoning. To learn reliable evidence-seeking policies, we first construct 4,000 multi-turn trajectories through verifiable trajectory distillation for supervised fine-tuning. During reinforcement learning, we assign fine-grained credit to evidence notes using information-gain rewards estimated from changes in the log-likelihood of the GT under the current policy. We further introduce a visual grounding gate that restricts positive credit to notes citing previously observed frames, mitigating reward hacking. Extensive experiments demonstrate the effectiveness of our multi-turn agentic reasoning framework for CVU. We will release our models and dataset to the community.

1 INTRODUCTION

Video understanding Fu et al. (2025); Ning et al. (2025); Tong et al. (2025) is central to dynamic scene perception and temporal reasoning in vision-language models (VLMs) Bai et al. (2025); Wang et al. (2025b). As applications expand from single-video question answering to multi-source video analysis, cross-video understanding (CVU) Peng et al. (2026); Ghazanfari et al. (2026) requires models not only to understand individual videos but also to establish object associations, event correspondences, and temporal relationships across them. For example, multi-camera association requires cross-view entity matching, while clip ordering requires inferring temporal relations among events across clips. CVU thus goes beyond a collection of independent single-video questions, requiring models to locate key evidence across videos, establish temporal correspondences, and integrate intermediate findings through explicit comparison Bai et al. (2026). Existing benchmarks such as CVBench Zhu et al. (2025) and CrossVid Li et al. (2026) systematically evaluate these capabilities, revealing substantial limitations in current models' cross-video association, temporal alignment, and compositional reasoning.

Existing methods Meng et al. (2025); Yan et al. (2026a) typically answer CVU questions directly from sampled frames, confining reasoning to a fixed visual context. Under limited context budgets, sparse sampling may miss brief yet critical events, while denser sampling increases encoding and context-processing costs, limiting scalability to multiple long videos Shen et al. (2024); Hu et al. (2025a). More importantly, evidence needs become clearer as reasoning unfolds, yet this paradigm lacks mechanisms to supplement and verify evidence based on intermediate judgments. For example, identifying an action in one video may reveal the need to examine the corresponding stage in another. CVU therefore requires adaptive evidence seeking and comparison guided by information gaps and cross-video differences, rather than simply more input frames. Fig. 1 illustrates

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.

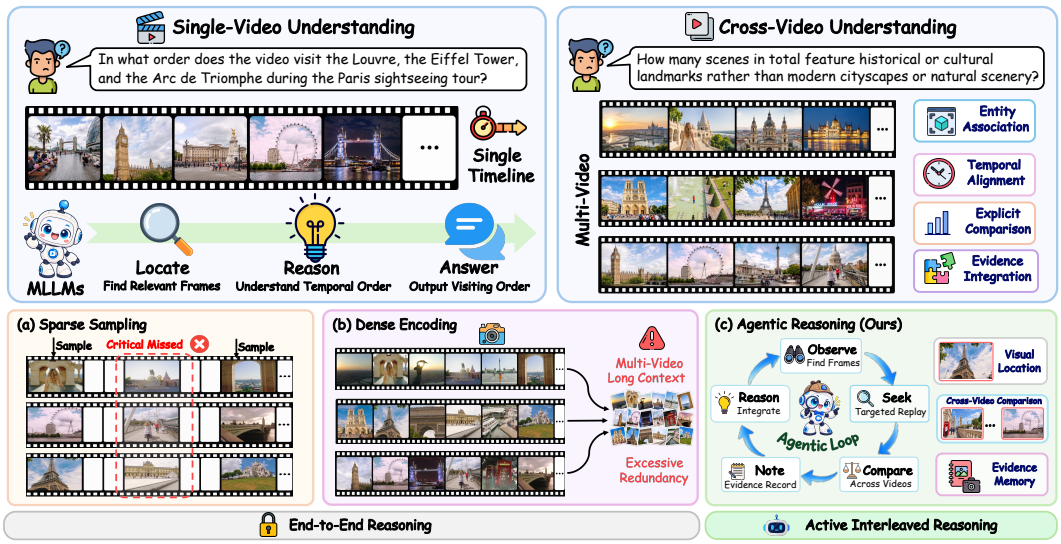


Figure 1: **Adaptive evidence seeking for cross-video understanding.** Single-video understanding operates over a single temporal stream, whereas CVU requires locating distributed evidence, establishing cross-video correspondences, explicitly comparing observations, and integrating findings across videos. (a) Sparse sampling may miss critical evidence, while (b) dense encoding introduces substantial redundant context. (c) Our agentic framework alternates observation, targeted replay, cross-video comparison, evidence recording, and reasoning to progressively acquire and integrate task-relevant evidence.

this contrast and motivates an agentic alternative that adaptively acquires and integrates evidence as reasoning unfolds.

Tool-augmented multi-turn agents (Hong et al., 2025; Chng et al., 2025) offer a path beyond fixed frame sampling. FrameThinker He et al. (2025), Video-DeepResearch Fang et al. (2026) and VideoSearcher Gao et al. (2026) enable models to progressively acquire and integrate evidence through video retrieval, memory, or external search. However, these methods primarily focus on single-video evidence gathering or external knowledge retrieval, whereas CVU requires temporal alignment, explicit comparison, and evidence accumulation across videos and interaction turns Peng et al. (2026). Evidence notes preserve visual findings for cross-turn reasoning, yet outcome-only rewards offer limited insight into individual notes' contributions and lack explicit grounding constraints. Reliable cross-video evidence seeking thus requires multi-turn supervision with verifiable answers and visual evidence, alongside fine-grained note-level credit assignment. Positive credit should be grounded in observed frames to avoid rewarding unsupported answer guesses.

To address these challenges, we introduce VIDEOSEEKER, an active evidence-seeking agentic framework that performs CVU through a multi-turn loop of observation, comparison, and reasoning. Starting from a sparse overview, the agent uses three purpose-built tools. `choose_frames` provides dense local replay to inspect potentially missed details. `compare` temporally interleaves frames from selected video segments to support explicit cross-video comparison. `note` records intermediate findings with video and frame references in persistent memory, allowing later reasoning to build on traceable evidence. We train the agent through cold-start supervised fine-tuning (SFT) followed by reinforcement learning (RL). To teach coordinated tool use and evidence recording, we develop a verifiable trajectory distillation pipeline spanning nine task types with independently sourced ground truth. Following multiple answer-leakage checks, a teacher model explores these tasks in a real tool environment without access to the answers. Joint validation of answer correctness, evidence-seeking behavior and evidence recording yields 4,000 multi-turn trajectories for SFT.

RL then refines evidence recording through note-level credit assignment using information gain. We use teacher forcing to compute the change in answer log-likelihood under the current policy before and after each note, treating this change as a proxy for the note's information contribution. However, unsupported answer assertions can also increase this likelihood. We therefore introduce a

visual grounding gate that restricts positive credit to notes citing previously observed frames, coupling the information-gain signal with a visual provenance constraint to mitigate reward hacking. Extensive experiments demonstrate consistent improvements across CVU benchmarks, with particularly strong gains on temporal alignment tasks. Ablation studies and note interventions further validate the effectiveness of the proposed framework and its use of persistent evidence memory.

In summary, our contributions are as follows:

- We propose an active evidence-seeking agentic framework for CVU that integrates targeted segment replay, temporally interleaved cross-video comparison, and persistent evidence memory into a multi-turn reasoning loop.
- We develop a verifiable trajectory distillation pipeline that jointly validates answers and evidence-seeking processes, yielding 4,000 cross-video reasoning trajectories across nine task types with independently sourced ground truth.
- We assign note-level credit using information gain estimated from the answer log-likelihood changes. A visual grounding gate limits positive credit to notes citing observed frames, mitigating potential reward hacking.
- Extensive experiments across multiple benchmarks validate the effectiveness of our overall framework and its key design components.

2 Related work

2.1 Long- and Multi-Video Understanding

Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), LVBench (Wang et al., 2025a), and LongVideoBench (Wu et al., 2024) extend video understanding evaluation in terms of video duration, task coverage, and fine-grained temporal reasoning. CVBench (Zhu et al., 2025), CrossVid (Li et al., 2026), and MVX-Bench (Zhang et al., 2026) further address joint multi-video understanding, covering entity association, event comparison, step alignment, and multi-view reasoning, and reveal limitations in models' ability to associate and integrate cross-video evidence. We focus on active evidence seeking in this setting, supporting cross-video relationship modeling and compositional reasoning through multi-turn observation, explicit comparison, and evidence accumulation.

2.2 Tool-Augmented Agentic VLMs

OpenThinkIMG (Su et al., 2025), DeepEyes (Zheng et al., 2025), and DeepEyesV2 (Hong et al., 2025) enhance multimodal reasoning through interactions with visual tools, while MMSearch-R1 (Wu et al., 2025a) and SenseNova-MARS (Chng et al., 2025) further integrate visual and textual search. In the video domain, VideoAgent (Wang et al., 2024) and VideoExplorer (Yuan et al., 2025) explore memory augmentation and iterative temporal localization, while VideoDR (Liu et al., 2026) and VideoSearcher (Gao et al., 2026) combine video evidence gathering with external knowledge retrieval. These video agents primarily focus on single-video understanding, whereas cross-video settings additionally require temporal correspondence across videos, explicit comparison, and the accumulation and reuse of evidence over multiple turns. We integrate local evidence replay, temporally interleaved cross-video comparison, and persistent evidence memory into an active evidence-seeking agent, and learn reliable cross-video evidence-seeking policies through verifiable trajectory distillation and note-level credit assignment.

2.3 Memory and Evidence-Note Mechanisms

MemAgent (Yu et al., 2026) and Memory-R1 (Yan et al., 2026b) optimize memory updating and management through reinforcement learning, while ReSum (Wu et al., 2025b) combines context summarization with reinforcement learning to support long-horizon search. More directly related to evidence recording, EviNote-RAG (Dai et al., 2025) trains models to distill evidence notes from retrieval results and uses an entailment-based reward to assess how well those notes support the answer. Unlike these text-memory and evidence-summarization methods, we introduce video–frame-level provenance citations into notes and combine their informational contribution with visual-source constraints, enabling traceable preservation and reuse of cross-video evidence.

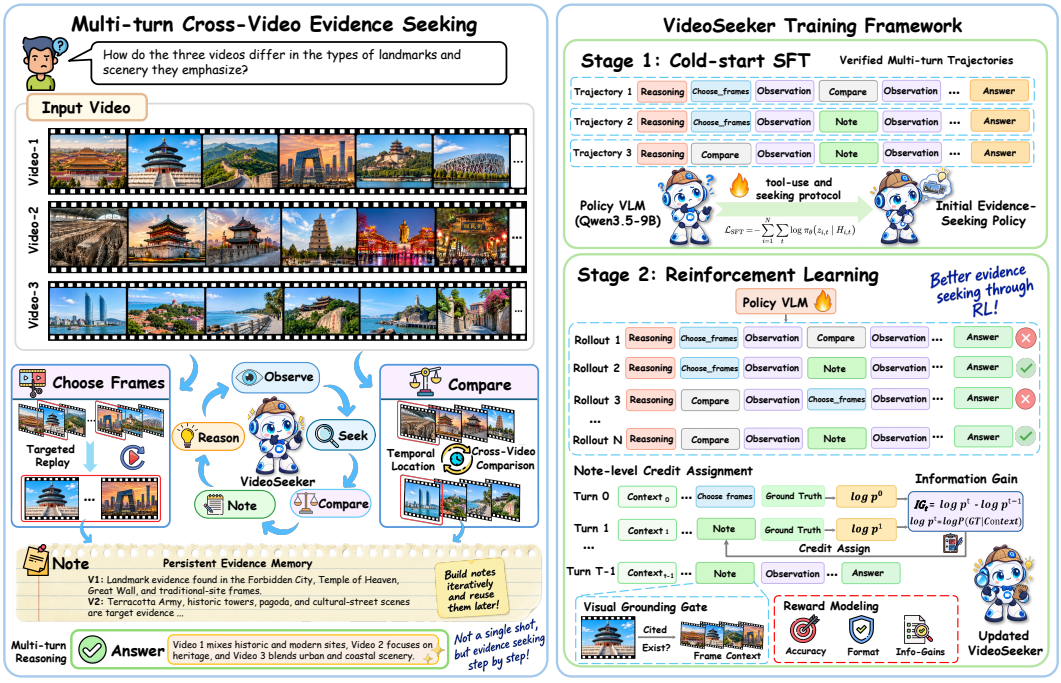


Figure 2: Overview of VideoSeeker for Cross-Video Understanding, integrating multi-turn evidence seeking with a two-stage training framework.

3 METHOD

3.1 PROBLEM FORMULATION

Task Definition. Given a collection of videos $\mathcal{V} = \{V^{(i)}\}_{i=1}^{N}$ with $N \geq 2$ and a question q, CVU requires producing an answer y by associating and integrating evidence across videos. We formulate this process as multi-turn interaction between an agent π_θ and a video environment. Starting from a sparse overview o_0, the agent constructs an interaction trajectory $\tau = (a_1, o_1, \ldots, a_T, o_T)$ before answering, where a_t denotes the action at turn t and o_t the corresponding observation returned by the environment. The agent adaptively localizes relevant moments, compares video segments, and records intermediate findings to support subsequent reasoning.

Observation Space. At turn t, the agent conditions on the interaction history H_t and a persistent notebook M_t, initialized with $H_1 = (q, o_0)$ and an empty notebook. The observation space is $\mathcal{O} = \mathcal{O}^v \cup \mathcal{O}^m$, where $\mathcal{O}^v$ contains retrieved video frames and $\mathcal{O}^m$ contains notebook feedback. Each visual observation preserves video and frame identifiers (i, j), allowing evidence from different videos to be distinguished and referenced. After executing a tool action, the history is updated as $H_{t+1} = (H_t, a_t, o_t)$. The `note` action appends an entry $n_t = (c_t, R_t)$ to the notebook, where c_t is a textual finding and R_t contains its video–frame references. The notebook is included in subsequent contexts, preserving intermediate findings across interaction turns.

Action Space. At each turn, the agent generates intermediate reasoning and either invokes a tool or produces the final answer. The available actions are:

- `choose_frames`: Select a video and a temporal interval for dense local replay.
- `compare`: Select intervals from multiple videos and temporally interleave their frames to facilitate explicit cross-video comparison.
- `note`: Record an intermediate finding in persistent evidence memory.
- `answer`: Produce the final prediction and terminate the interaction.

Tool arguments must refer to valid videos and intervals. The environment tracks observed frame identifiers, allowing note references to be checked against actual visual observations for subsequent

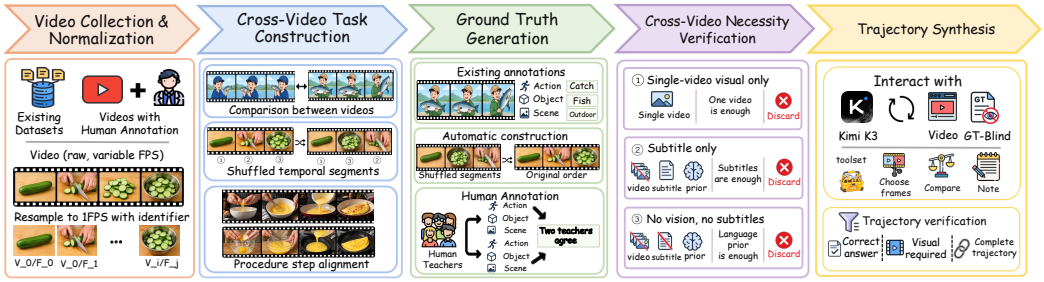


Figure 3: Overview of the verifiable cross-video data construction and trajectory synthesis pipeline.

credit assignment. Unlike replay and comparison, `note` updates the agent's memory without acquiring new visual evidence.

3.2 Cross-Video Data Pipeline

To train reliable evidence-seeking behaviors for CVU, we develop a verifiable trajectory distillation pipeline that constructs cross-video tasks with independently sourced ground truth and distills multi-turn tool-interactive trajectories under strict process verification. The overall data construction pipeline is illustrated in Fig. 3, with further implementation details provided in Appendix B.

Task-Oriented Video Curation. We collect source videos from topic-based YouTube crawls, Kinetics-600 Carreira et al. (2018), LSMDC Rohrbach et al. (2017), and RoomTour3D Han et al. (2025), with no overlap with the evaluation benchmarks. Each video is resampled to 1 FPS and overlaid with a "Video i / Frame j" identifier. Videos are then organized according to task requirements, including related clips for comparison, shuffled segments from the same video for temporal ordering, and videos of the same procedure for step alignment.

Task Construction with Independent Ground Truth. We instantiate nine task types with ground truth independent of the answering agent, as detailed in Appendix B.2. The GT is derived from three sources: *constructed* GT induced directly by task construction, such as the original order of shuffled segments; *label-anchored* GT inherited from existing action labels; and *teacher-authored* GT generated from dense frame grids and retained only when two independent teacher samples agree. Questions refer to moments through visual events rather than timestamps and are required to rely on evidence from multiple videos.

Cross-Video Necessity Verification. We verify that each task genuinely requires cross-video visual evidence under three input settings: video frames, subtitles only, and no visual input. Each setting is evaluated with two independent teacher samples. We discard tasks that cannot be solved from visual frames or can be answered from subtitles or language priors alone.

GT-Blind Exploration and Trajectory Verification. For each verified task, the teacher explores the tool environment without access to the ground-truth answer, ensuring that trajectories reflect evidence seeking rather than post-hoc justification. The environment enforces only procedural constraints, rejecting premature answers and trajectories with an empty notebook. We retain trajectories only when the final answer is correct, visual evidence is acquired before answering, notes are recorded incrementally as findings emerge, and image placeholders are correctly aligned with their referenced frames. This process yields 4,000 high-quality trajectories averaging 3.37 turns and 66.9 observed frames, while the remaining verified tasks are reserved for RL.

3.3 Model Training

We train VideoSeeker in two stages. Cold-start SFT establishes the multi-turn evidence-seeking protocol and coordinated tool-use behaviors, while RL further refines evidence recording through fine-grained note-level credit assignment. The overall training framework is illustrated in Fig. 2.

3.3.1 COLD-START SFT

Given the curated dataset $\mathcal{D}_{\mathrm{SFT}} = \{(x_k, \tau_k^\star)\}_{k=1}^{K}$, where $x_k = (\mathcal{V}_k, q_k)$ and $\tau_k^\star$ is a verified multi-turn trajectory, we optimize

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{k=1}^{K} \sum_{t} \log \pi_\theta \left(z_{k,t} \mid H_{k,t} \right), \tag{1}$$

where $z_{k,t}$ denotes the assistant-generated sequence at turn t. Tool observations are provided as conditioning context but excluded from the loss. This stage teaches the agent to coordinate replay, comparison, and evidence recording before RL.

3.3.2 RL WITH NOTE-LEVEL CREDIT ASSIGNMENT

After SFT, we further optimize VIDEOSEEKER through online RL in the same tool environment. In CVU, `note` is essential for preserving intermediate evidence across videos and turns, but different notes contribute unequally to subsequent reasoning. Outcome-only rewards cannot distinguish useful evidence from redundant or misleading notes. We therefore introduce note-level credit assignment based on information gain, with visual grounding to constrain positive credit. For each task x, the old policy $\pi_{\theta_{\mathrm{old}}}$ samples a group of G trajectories $\{\tau_g\}_{g=1}^{G}$.

Outcome Reward. We retain the task objective through an outcome reward gated by format and interaction validity,

$$r_g^{\mathrm{acc}} = \mathbb{I}[\mathrm{fmt}_g] \cdot \mathbb{I}[\mathrm{gate}_g] \cdot \mathrm{Corr}(y_g, y^\star), \tag{2}$$

where fmt_g checks answer formatting, gate_g requires at least one successful visual action before answering, and $\mathrm{Corr} \in [0, 1]$ evaluates task-specific answer correctness.

Note-level Information Gain. To quantify the contribution of each evidence note, we measure how it changes the policy's support for the ground-truth answer. For $y^\star = (y_1^\star, \ldots, y_L^\star)$, we define

$$V(H) = \frac{1}{L} \sum_{l=1}^{L} \log \pi_\theta \left(y_l^\star \mid H, \texttt{<answer>}, y_{<l}^\star \right), \qquad \mathrm{IG}_t = V(H_t^+) - V(H_t^-), \tag{3}$$

where H_t^- and H_t^+ denote the histories immediately before and after note n_t, respectively. $V(H)$ is computed by teacher forcing under the current policy with gradients stopped. A positive IG_t indicates that the new note increases the likelihood of the ground-truth answer, providing an endogenous credit signal without an external process scorer.

Visual Grounding Gate. Unlike retrieved documents, a note is written by the policy itself. A note asserting the answer raises $V(H_t^+)$ without any visual basis, so rewarding IG_t directly invites reward hacking. Let $\mathcal{F}_t$ denote the set of frame identifiers (i, j) observed in H_t, which the environment tracks, and let R_t be the references of note n_t. Positive credit is granted only when the cited frames were actually observed,

$$g_t = \mathbb{I}[|R_t \cap \mathcal{F}_t| > \rho \, |R_t|], \qquad \widetilde{\mathrm{IG}}_t = g_t \, \mathrm{IG}_t + (1 - g_t) \min(\mathrm{IG}_t, 0). \tag{4}$$

with $\rho = 0.8$. Thus, an ungrounded note cannot receive positive credit, while negative information gain is preserved regardless of grounding.

Credit Assignment and Policy Objective. We separately normalize the outcome reward and grounded information gain within each rollout group to account for their different scales,

$$\tilde{r}_g^{\mathrm{acc}} = \frac{r_g^{\mathrm{acc}} - \mu^{\mathrm{acc}}}{\sigma^{\mathrm{acc}} + \epsilon}, \qquad \tilde{r}_{g,t}^{\mathrm{ig}} = \frac{\widetilde{\mathrm{IG}}_{g,t} - \mu^{\mathrm{ig}}}{\sigma^{\mathrm{ig}} + \epsilon}, \tag{5}$$

where $(\mu^{\mathrm{acc}}, \sigma^{\mathrm{acc}})$ are computed over the G outcome rewards, and $(\mu^{\mathrm{ig}}, \sigma^{\mathrm{ig}})$ over all grounded information gains within the group. We assign $\tilde{r}_{g,t}^{\mathrm{ig}}$ to note turns, $\tilde{r}_g^{\mathrm{acc}}$ to the final answer turn, and zero reward otherwise. To propagate note-level credit to preceding evidence-seeking actions, we define the turn-level advantage as the cumulative future reward:

$$\hat{A}_{g,t} = \sum_{t' \geq t} \tilde{r}_{g,t'}. \tag{6}$$

Table 1: Performance on cross-video understanding benchmarks under *Without Tool*, *Agentic Model (Zero-shot)*, and *Agentic Model* settings. For CrossVID, CA, TU, MVR, and FF denote Comparative Analysis, Temporal Understanding, Multi-view Reasoning, and Free-form QA, respectively. For CVBench, OA, EA, and CR denote Object Association, Event Association, and Complex Reasoning. For MVU-Eval, P and R denote Perception and Reasoning. C.Avg, CV.Avg, and M.Avg denote the corresponding benchmark-level averages, while O.Avg denotes the arithmetic mean of C.Avg, CV.Avg, and M.Avg. The best and second-best results are shown in **bold** and underlined, respectively.

Models	CrossVID				C.Avg	CVBench			CV.Avg	MVU-Eval		M.Avg	O.Avg
	CA	TU	MVR	FF		OA	EA	CR		P	R		
Without Tool													
Open-source													
Qwen3.5-4B	28.70	19.13	21.96	29.80	24.59	59.51	57.61	56.99	57.85	59.88	48.76	52.83	45.09
Qwen3.5-9B	32.69	34.38	32.96	26.17	32.60	60.32	62.23	64.38	62.58	62.63	51.86	55.80	50.33
Qwen3.5-27B	54.02	38.87	32.46	31.00	42.86	48.58	60.05	58.31	56.54	**75.80**	59.81	65.66	55.02
Qwen3.5-35B-A3B	46.42	38.53	32.85	30.09	39.71	63.16	64.67	67.28	65.29	66.16	56.27	59.89	54.96
Qwen3.5-122B-A10B	43.51	40.86	31.21	35.07	39.41	66.40	64.67	70.71	67.40	73.20	64.13	67.45	58.09
Qwen3.6-27B	50.87	44.42	29.81	29.59	42.60	58.30	63.59	64.64	62.68	68.30	60.16	63.14	56.14
Open-source large-scale model													
MiniMax-M3	42.15	39.97	26.46	23.14	36.46	61.94	64.95	60.16	62.37	66.77	57.16	60.67	53.17
GLM-5.3-Flash	**58.40**	47.41	42.50	**37.26**	49.81	53.44	64.13	58.05	59.15	69.98	64.75	66.67	58.54
Qwen3.5-397B-A17B	49.49	48.83	34.70	31.24	44.51	56.28	62.23	64.38	61.57	71.52	63.60	66.50	57.53
Agentic Model (Zero-shot)													
Open-source													
Qwen3.5-4B	31.46	19.36	12.51	30.09	23.90	51.82	55.71	51.72	53.22	45.94	33.75	38.21	38.44
Qwen3.5-9B	37.61	29.35	28.79	25.64	32.17	59.51	63.04	63.32	62.27	62.79	56.80	58.99	51.14
Qwen3.5-27B	31.33	44.05	18.32	14.09	30.82	57.09	65.49	65.17	63.28	73.05	64.75	67.79	53.96
Qwen3.5-35B-A3B	43.01	36.74	36.05	30.45	38.48	61.94	65.22	69.13	65.90	63.55	57.77	59.89	54.76
Qwen3.5-122B-A10B	46.65	41.93	34.29	31.42	41.24	69.23	66.85	67.55	67.71	70.75	64.40	66.72	58.56
Qwen3.6-27B	48.25	45.50	33.08	27.66	42.33	56.28	63.32	67.81	63.28	71.21	59.28	63.64	56.42
Open-source large-scale model													
MiniMax-M3	41.60	39.47	18.97	26.79	34.95	57.09	64.95	62.53	62.07	64.47	57.69	60.17	52.40
GLM-5.3-Flash	56.34	53.03	38.74	34.63	49.65	69.64	70.38	70.45	70.22	68.45	62.72	64.82	61.56
Qwen3.5-397B-A17B	46.01	50.42	32.13	29.71	42.93	62.75	68.48	68.07	66.90	75.19	**67.23**	**70.14**	59.99
Agentic Model													
VideoSeeker-9B	50.63	59.53	43.96	33.64	50.27	68.02	**76.63**	75.46	74.04	71.98	63.87	66.83	63.71
Δ v.s. Qwen3.5-9B	**+17.94**	**+25.15**	**+11.00**	**+7.47**	**+17.67**	**+7.70**	**+14.40**	**+11.08**	**+11.46**	**+9.35**	**+12.01**	**+11.03**	**+13.38**
VideoSeeker-35B-A3B	51.67	**63.27**	**44.57**	33.01	**51.86**	**73.68**	75.82	**75.99**	**75.35**	74.43	64.66	68.24	**65.15**
Δ v.s. Qwen3.5-35B-A3B	**+5.25**	**+24.74**	**+11.72**	**+2.92**	**+12.15**	**+10.52**	**+11.15**	**+8.71**	**+10.06**	**+8.27**	**+8.39**	**+8.35**	**+10.19**

The same advantage is shared by all generated tokens within turn t, allowing earlier actions to receive credit for useful evidence they enable while preserving final-answer correctness as the task objective. Without note-level information gain, Eq. 6 reduces to a single outcome advantage shared across the trajectory. We optimize the policy with a clipped surrogate objective:

$$\mathcal{L}_{\text{RL}} = -\frac{1}{\sum_g |\tau_g|} \sum_g \sum_t \sum_l \min\left(w_{g,t,l}\hat{A}_{g,t},\ \text{clip}(w_{g,t,l}, 1-\epsilon_c, 1+\epsilon_c)\hat{A}_{g,t}\right) + \beta\, \mathbb{D}_{\text{KL}}\left[\pi_\theta \,\|\, \pi_{\text{sft}}\right], \tag{7}$$

where $w_{g,t,l}$ denotes the token-level importance ratio between π_θ and $\pi_{\theta_{\text{old}}}$, and observation tokens are excluded from optimization. This objective preserves trajectory-level answer correctness while providing fine-grained, visually grounded credit for evidence recording and the preceding actions that support it.

4 Experiments

4.1 Implementation Details

Training Details. VideoSeeker is initialized from Qwen3.5-9B and Qwen3.5-35B-A3B Qwen Team (2026a). Both variants follow the same two-stage training pipeline, using ms-swift Zhao et al. (2024) for cold-start SFT and veRL Sheng et al. (2025) for online RL. During SFT, we update all model parameters, including the vision encoder, for one epoch on 4,000 verified trajectories with a learning rate of 1×10^{-5} and a global batch size of 8. The RL stage starts from the SFT checkpoint and is trained for one epoch on a pool of 2,936 tasks, using a prompt batch size of 32, 8 rollouts per prompt,

Table 2: Performance on general video understanding benchmarks. VIDEOSEEKER is evaluated with the same tool environment as on cross-video benchmarks; baselines follow the official protocol of each benchmark.

Models	MMVU	TempCompass	Video-MMMU	VideoMathQA	Avg.
Qwen3.5-4B	55.40	71.58	62.56	33.71	55.81
Qwen3.5-9B	63.30	78.20	69.67	38.05	62.31
Qwen3.5-35B-A3B	**69.70**	**81.78**	<u>76.56</u>	<u>52.05</u>	<u>70.02</u>
Video-R1-7B	51.00	63.46	50.11	48.71	53.32
VideoChat-R1.5-7B	51.50	71.37	30.22	46.24	49.83
VideoRFT-7B	50.50	55.85	37.56	43.48	46.85
VideoCoM	<u>65.40</u>	71.30	50.20	27.80	53.68
VideoSeeker-9B	63.60	78.54	70.78	49.33	65.56
VideoSeeker-35B-A3B	**69.70**	<u>80.70</u>	**77.89**	**58.43**	**71.68**

Table 3: Algorithmic ablation on fine-grained credit assignment. We vary which interaction turns receive credit and whether note-level credit is filtered by the visual grounding gate, while keeping the tool set fixed across all variants.

Setting	CrossVID	CVBench	MVUEval
SFT	44.15	72.74	60.22
GRPO	49.76	73.04	66.33
CA (Choose Frame)	49.67	73.54	65.27
CA (Compare)	49.26	73.94	64.82
CA (Note) w/o Gate	48.97	73.84	65.71
Ours	**50.27**	**74.04**	**66.83**

and an actor learning rate of 1×10^{-6}. Additional training and rollout hyperparameters are provided in the Appendix.

Benchmarks. We evaluate on two categories of benchmarks. For cross-video understanding, we use CVBench Zhu et al. (2025), CrossVid Li et al. (2026), and MVU-Eval Peng et al. (2026), which cover object and event association, temporal alignment, multi-view matching, and open-ended comparison over groups of 2 to more than 10 videos. For general video understanding, we further test on MMVU Zhao et al. (2025), TempCompass Liu et al. (2024), Video-MMMU Hu et al. (2025b), and VideoMathQA Rasheed et al. (2025a) to verify that evidence-seeking training does not compromise single-video capability. Training videos have no overlap with any benchmark. It should be noted that we correct erroneous annotations and remove unresolved samples from the original benchmark test sets and re-evaluate all baselines on the resulting cleaned subsets. The details are provided in the Appendix D.3.

Baselines. We compare VideoSeeker against a broad set of strong open-weight MLLMs. These include the Qwen3.5 series Qwen Team (2026a), Qwen3.6 Qwen Team (2026b), which provide controlled comparisons across model scales and generations, as well as recent frontier models GLM-5.3-Flash Z.ai (2026), and MiniMax M3 MiniMax (2026). We additionally compare with Video-R1 (Feng et al., 2025), VideoChat-R1.5 (Yan et al., 2026c), VideoRFT (Wang et al., 2026), and Video-CoM (Rasheed et al., 2025b). We evaluate models under three settings. (1) *Without Tool*, where models follow the official protocol of each benchmark, uniformly sample frames from all videos, and answer in a single pass. (2) *Agentic Model (Zero-shot)*, where pretrained models are equipped with the same tool interface and interaction protocol as VideoSeeker. (3) *Agentic Model*, where VideoSeeker is evaluated after our two-stage SFT and RL training. This setup separates the gains from tool-enabled inference from those introduced by agentic training.

4.2 MAIN RESULTS

Cross Video Understanding. As shown in Tab. 1, VideoSeeker achieves strong performance across cross-video understanding benchmarks. On CrossVid, CVBench, and MVU-Eval, VideoSeeker consistently improves over the corresponding Qwen3.5 base models, demonstrating the effectiveness of active evidence seeking for associating, comparing, and integrating information distributed across

Table 4: Tool ablation under the same training objective. We evaluate individual and pairwise combinations of *Choose Frame*, *Compare*, and *Note* to isolate their respective contributions and complementarity in cross-video evidence seeking.

Choose Frame	Compare	Note	CrossVID	CVBench	MVUEval
✓			44.82	70.02	59.78
	✓		43.66	70.52	61.18
		✓	44.94	70.02	60.84
✓	✓		49.25	73.64	65.04
✓		✓	49.39	73.24	65.38
	✓	✓	49.69	73.24	65.21
✓	✓	✓	**50.27**	**74.04**	**66.83**

multiple videos. The gains are particularly pronounced on tasks requiring temporal understanding and cross-video reasoning. Notably, simply equipping pretrained models with the same tools does not consistently improve performance, whereas VideoSeeker substantially benefits from dedicated agentic training, indicating that effective cross-video reasoning requires learning how to actively acquire and organize evidence rather than tool access alone.

General Video Understanding. Tab. 2 further evaluates VideoSeeker on general video understanding benchmarks. Despite being specifically trained for multi-turn cross-video evidence seeking, VideoSeeker preserves and generally improves the underlying video understanding capability of the base models. In particular, consistent gains on reasoning-intensive benchmarks such as VideoMathQA suggest that the learned evidence-seeking behavior transfers beyond the cross-video setting. These results show that our agentic training improves cross-video reasoning while preserving transferable video understanding.

4.3 Ablation Study

Algorithmic Ablation. Tab. 3 studies where fine-grained credit should be assigned and whether it should be constrained by visual grounding. Outcome-only GRPO already improves over SFT, while assigning additional credit to `Choose Frame` or `Compare` turns does not yield consistent gains. In contrast, our grounded note-level credit achieves the best performance across all three benchmarks. Removing the visual grounding gate consistently degrades performance, supporting the importance of grounding note-level rewards in actually observed visual evidence to avoid spurious credit. Notably, note-level credit without the gate still remains competitive with other credit-assignment variants, suggesting that `Note` provides an effective granularity for attributing intermediate reasoning progress.

Tool Ablation. Tab. 4 evaluates the complementary roles of `Choose Frame`, `Compare`, and `Note` under the same training objective. Individual tools perform substantially worse than their combinations, while different pairwise variants exhibit complementary strengths across benchmarks. The complete system consistently achieves the best performance, indicating that targeted evidence acquisition, explicit cross-video comparison, and persistent evidence recording provide distinct yet complementary capabilities for cross-video reasoning. This further suggests that effective evidence seeking relies on coordinating these capabilities rather than optimizing any single tool in isolation.

5 Conclusion

In this work, we revisit cross-video understanding as a multi-turn evidence-seeking problem rather than fixed-context end-to-end prediction. We introduce VideoSeeker, an agentic framework that actively localizes relevant moments, compares evidence across videos, and preserves intermediate findings through persistent evidence memory. To learn reliable evidence-seeking behavior, we develop a verifiable trajectory distillation pipeline for cold-start supervision and further refine evidence recording through note-level credit assignment based on information gain, with visual grounding to mitigate reward hacking. Extensive experiments across cross-video understanding benchmarks demonstrate consistent improvements over strong base models, supporting the effectiveness of active evidence seeking for CVU.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing and to improve the clarity and readability of the manuscript. They were not used to generate research ideas, design the methodology or experiments, produce experimental results, or draw scientific conclusions. All AI-assisted revisions were carefully reviewed and verified by the authors, who take full responsibility for the content of the paper. Generative AI models that constitute part of the proposed methodology, data construction, or evaluation pipeline are separately described in the main text and appendix.

ETHICS STATEMENT

All videos and related materials used in our data construction are obtained from publicly accessible online sources and existing open-source datasets. We collect and process these resources in accordance with the corresponding platform policies and dataset licenses, and use them solely for academic research and evaluation. The collected videos are used exclusively to construct video-grounded QA instances and multi-turn tool-interaction trajectories.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility, we will release all code, training data, cleaned evaluation sets, and the complete data construction pipeline used in this work. The release will cover the multi-turn tool environment, task construction and filtering procedures, trajectory generation and verification pipeline, supervised fine-tuning and reinforcement learning implementations, benchmark preprocessing and cleaning protocols, and all configurations required to reproduce the reported experiments. We will also provide the processed training trajectories and cleaned test subsets used for evaluation, enabling end-to-end reproduction of the data construction, training, and evaluation process.

REFERENCES

Purui Bai, Tao Wu, Jiayang Sun, Xinyue Liu, Huaibo Huang, and Ran He. Mvpbench: A multi-video perception evaluation benchmark for multi-modal video understanding. *arXiv preprint arXiv:2603.22756*, 2026.

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*, 2025.

João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about Kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

Yuqin Dai, Guoqing Wang, Yuan Wang, Kairan Dou, Kaichen Zhou, Zhanwei Zhang, Shuo Yang, Fei Tang, Jun Yin, Pengyu Zeng, et al. Evinote-rag: Enhancing rag models via answer-supportive evidence notes. *arXiv preprint arXiv:2509.00877*, 2025.

Zhen Fang, Yu Zeng, Wenxuan Huang, Yiming Zhao, Shiting Huang, Tianfei Ren, Qi Lu, Qingnan Ren, Qisheng Su, Lionel Z Wang, et al. Video-deepresearch: Towards the next-generation multimodal deepresearch agent. *arXiv preprint arXiv:2608.03979*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, et al. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025.

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

Zhenkun Gao, Yicheng Bao, Jinlong Peng, Xueheng Li, Theo Huang, Bangwei Liu, Kunquan Li, Zhenye Gan, Tao Hu, Chengjun Xie, et al. Videosearcher: Empowering video deep research with multi-tool agentic reasoning via reinforcement learning. *arXiv preprint arXiv:2607.02927*, 2026.

Sara Ghazanfari, Siddharth Garg, Prashanth Krishnamurthy, and Farshad Khorrami. Syncr: A cross-video reasoning benchmark with synthetic grounding. *arXiv preprint arXiv:2605.08412*, 2026.

Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. Roomtour3d: Geometry-aware video-instruction tuning for embodied navigation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27586–27596. IEEE, 2025.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*, 2025.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

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

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025b.

Jingyao Li, Jingyun Wang, Molin Tan, Haochen Wang, Cilin Yan, Likun Shi, Jiayin Cai, Xiaolong Jiang, and Yao Hu. CrossVid: A comprehensive benchmark for evaluating cross-video reasoning in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6244–6252, 2026. arXiv:2511.12263.

Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, Kunyi Wang, Rui Xu, Sen Hu, Jianheng Hou, et al. Watching, reasoning, and searching: A video deep research benchmark on open web for agentic video reasoning. *arXiv preprint arXiv:2601.06943*, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Yiran Meng, Junhong Ye, Wei Zhou, Guanghui Yue, Xudong Mao, Ruomei Wang, and Baoquan Zhao. Videoforest: Person-anchored hierarchical reasoning for cross-video question answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 836–845, 2025.

MiniMax. Minimax m3: Frontier coding, 1m context, native multimodality—all in one model. *URL https://www.minimax.io/blog/minimax-m3*, 2026.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *Computational Visual Media*, 2025.

Tianhao Peng, Haochen Wang, Yuanxing Zhang, Noah Wang, Zili Wang, Ge Zhang, Jian Yang, Shihao Li, Yanghai Wang, Xintao Wang, et al. Mvu-eval: Towards multi-video understanding evaluation for multimodal llms. *Advances in Neural Information Processing Systems*, 38, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026a.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025a.

Hanoona Rasheed, Mohammed Zumri, Muhammad Maaz, Ming-Hsuan Yang, Fahad Shahbaz Khan, and Salman Khan. Video-com: Interactive video reasoning via chain of manipulations. *arXiv preprint arXiv:2511.23477*, 2025b.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

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. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in neural information processing systems*, 38:4350–4376, 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a.

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

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, 2024.

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

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*, 2025a.

Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, et al. Resum: Unlocking long-horizon search intelligence via context summarization. *arXiv preprint arXiv:2509.13313*, 2025b.

Pengyu Yan, Akhil VSS Gorugantu, Mahesh Bhosale, Abdul Wasi, Vishvesh Trivedi, and David Doermann. Trace: Evidence grounding-guided multi-video event understanding and claim generation. In *Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2026)*, pp. 120–129, 2026a.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026b.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026c.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, et al. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. In *International Conference on Learning Representations*, volume 2026, pp. 39458–39486, 2026.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. Videoexplorer: Think with videos for agentic long-video understanding. *arXiv preprint arXiv:2506.10821*, 2025.

Z.ai. Glm-5.3-flash: Frontier intelligence, flash cost. *URL https://z.ai/blog/glm-5.3-flash*, 2026.

Yue Zhang, Liqiang Jing, Jia Li, Yapeng Tian, Xinya Du, Yunhui Guo, and Vibhav Gogate. A skill-augmented agentic framework and benchmark for multi-video understanding. *arXiv preprint arXiv:2603.14733*, 2026.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8475–8489, 2025.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL `https://arxiv.org/abs/2408.05517`.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. CVBench: Benchmarking cross-video synergies for complex multimodal reasoning. *arXiv preprint arXiv:2508.19542*, 2025.

APPENDIX

A IMPLEMENTATION DETAILS

Training Details. VideoSeeker is initialized from Qwen3.5-9B and Qwen3.5-35B-A3B Qwen Team (2026a), and both variants are trained with the same two-stage pipeline. For cold-start SFT, we use ms-swift Zhao et al. (2024) to optimize all model parameters, including the vision encoder, for one epoch over the 4,000 verified multi-turn trajectories constructed by our data pipeline. We use a global batch size of 8 and a learning rate of 1×10^{-5}. Online RL is performed with veRL Sheng et al. (2025), initialized from the SFT checkpoint. We train for one epoch over a pool of 2,936 verified cross-video tasks. Each prompt is sampled with 8 rollouts, with a prompt batch size of 32 and an actor learning rate of 1×10^{-6}. Rollouts are generated with temperature 1.0 and top-p 1.0 to maintain sufficient exploration during policy optimization.

Each video frame is encoded with at most 64 visual tokens. We set the maximum context length to 28,672 tokens and the maximum generated response length to 16,384 tokens. Each trajectory contains at most 18 assistant turns, and each of the three tools—`choose_frames`, `compare`, and `note`—can be invoked at most five times. These limits balance long-horizon evidence seeking with tractable training cost. For the RL objective, we set the clipping coefficient to $\epsilon_c = 0.2$ and the KL regularization coefficient to $\beta = 0.001$. The visual grounding threshold in the note-level credit assignment is set to $\rho = 0.8$. The native thinking mode of Qwen3.5 is disabled, while explicit `<think>` blocks are retained as part of the multi-turn interaction format. For open-ended tasks, Qwen3.6-27B is used as the training-time judge to evaluate answer correctness. Incremental note taking is encouraged by a 0.05 bonus added to the correctness term rather than by a hard gate. For the `note`-only variant of Table 4 the gate action is `note` itself, and when `note` is disabled no turn receives credit and the objective reduces to outcome-only GRPO.

B DATA PIPELINE

B.1 TRAINING DATA CONSTRUCTION

We provide additional details of the cross-video data construction pipeline used to obtain supervision for VideoSeeker. Our goal is to construct not only question–answer pairs that genuinely require cross-video visual reasoning, but also verifiable multi-turn trajectories that expose how relevant evidence is localized, compared, recorded, and integrated. To this end, the pipeline separates task construction and ground-truth generation from trajectory exploration, and applies multiple stages of verification before a trajectory is admitted into the training set.

We provide additional details of the cross-video data construction pipeline used to obtain supervision for VideoSeeker. Our goal is to construct not only question–answer pairs that require cross-video visual reasoning, but also verifiable multi-turn trajectories that explicitly capture how relevant evidence is localized, compared, recorded, and integrated. To reduce answer leakage and post-hoc rationalization, task construction and ground-truth generation are separated from trajectory exploration, followed by multiple stages of task- and trajectory-level verification.

Source Video Collection and Organization. We construct the source video pool from topic-based YouTube crawls and existing video datasets, including Kinetics-600 Carreira et al. (2018), LSMDC Rohrbach et al. (2017), and RoomTour3D Han et al. (2025). All training videos are disjoint from the evaluation benchmarks. Each video is resampled at 1 FPS and every frame is overlaid with an explicit "Video i / Frame j" identifier. Frame indices are maintained independently for each video, making every visual observation uniquely addressable and allowing subsequent evidence notes to preserve explicit visual provenance. Videos are then organized according to the reasoning structure required by different tasks. Semantically related clips are grouped for cross-video comparison, temporally continuous segments are shuffled to construct ordering tasks, and videos depicting the same or related procedures are grouped for step alignment. Through these constructions, task-relevant evidence is distributed across multiple visual streams rather than concentrated in a single video.

Task Construction with Independent Ground Truth. We instantiate nine cross-video task types covering complementary forms of association, comparison, temporal reasoning, and evidence in-

tegration. A key design principle is that the ground-truth answer is obtained independently of the agent that later generates the reasoning trajectory. We derive supervision from three sources.

First, *constructed ground truth* is induced directly by the task construction process. For example, when temporally continuous video segments are shuffled, their original chronological order naturally defines the target ordering. Second, *label-anchored ground truth* is inherited from existing semantic or action annotations associated with the source data. Third, for tasks whose answers cannot be directly derived from construction metadata or existing labels, we use *teacher-authored ground truth*: a teacher model observes dense frame grids and independently generates the reference answer twice, and only instances for which the two samples agree are retained. Importantly, the resulting ground-truth answer is not exposed during subsequent trajectory exploration. Questions refer to relevant moments through visual events or semantic descriptions rather than explicit timestamps or frame indices, preventing the task formulation itself from revealing where the decisive evidence is located.

Cross-Video Necessity Verification. Before trajectory generation, we verify that each candidate task genuinely requires visual evidence rather than being solvable through textual shortcuts or language priors. Each task is evaluated under three information conditions: video frames, subtitles only, and no visual input, with two independent teacher samples under each condition. We discard tasks that cannot be reliably solved when visual evidence is available, as such instances may be ambiguous or incorrectly constructed. Conversely, tasks that can be answered from subtitles alone or without visual input are also removed. Combined with the task-oriented video construction above, this filtering stage reduces shortcut solutions and retains tasks that provide meaningful supervision for cross-video visual reasoning.

GT-Blind Multi-Turn Exploration. For each verified task, a teacher model interacts with the same multi-turn video environment used by VIDEOSEEKER, but without access to the ground-truth answer. This GT-blind setting forces the trajectory to emerge from actual evidence acquisition rather than from constructing a plausible explanation around a known answer.

The teacher initially receives only a sparse overview formed by uniformly sampled frames from all input videos. It can then iteratively invoke three tools. `choose_frames` selects an interval from a single video and returns up to eight sampled frames for targeted local replay. `compare` selects intervals from two videos and returns their frames in temporally interleaved order, facilitating explicit comparison of objects, actions, scenes, and temporal correspondences. `note` records question-relevant intermediate findings together with their video–frame references in persistent memory. Each tool can be called at most five times, and only one tool action is permitted at each interaction turn. The teacher is not allowed to answer solely from the initial sparse overview and must perform at least one successful `choose_frames` or `compare` operation before producing the final answer. This encourages the trajectory to actively verify the decisive visual evidence rather than relying on coarse observations or parametric knowledge.

Incremental Evidence Recording. To obtain trajectories with explicit and reusable intermediate evidence, we enforce a *record-on-discovery* protocol. Whenever a tool observation reveals question-relevant evidence—for example, identifying an entity, locating a temporal boundary, supporting or eliminating a candidate, or establishing a cross-video correspondence—the subsequent action must record that finding with `note` before further inspection.

Each note is expected to remain concise and factual, specifying the relevant video–frame references, the observed content, and its implication for the question. Observations that do not contribute useful evidence are not recorded. This protocol encourages the notebook to accumulate evidence progressively as reasoning unfolds, rather than degenerating into a retrospective summary generated immediately before the final answer.

Trajectory Verification and Filtering. Teacher-generated trajectories are retained only after both outcome-level and process-level verification. First, the final answer must agree with the independently obtained ground truth. For structured tasks, correctness is checked according to the corresponding answer format. For open-ended questions, candidate answers are evaluated against predefined scoring points, with coverage and factual correctness assessed separately. We further require that visual evidence is actively acquired before answering and reject trajectories with an empty notebook. Relevant findings must be recorded incrementally as they emerge, and all image placeholders and video–frame references are checked for consistency with the observations actually returned by

the environment. These constraints ensure that intermediate findings are traceable to observed visual evidence rather than unsupported assertions.

The environment itself enforces only procedural constraints and never reveals the ground-truth answer to the teacher. Premature answers are rejected, only one action is accepted per turn, and trajectories that exhaust the tool budget must answer based on the evidence accumulated so far. After all verification stages, we retain 4,000 high-quality multi-turn trajectories for cold-start SFT, with an average of 3.37 interaction turns and 66.9 observed frames per trajectory. These trajectories provide supervision for coordinated evidence localization, explicit cross-video comparison, persistent evidence recording, and final-answer reasoning. The remaining verified tasks are reserved for online RL, where the policy further improves its evidence-seeking behavior through on-policy interaction with the same environment.

B.2 Training Data Composition

Our cold-start dataset contains 4,000 verified multi-turn trajectories spanning nine cross-video task types, as summarized in Tab. 5. The task mixture is designed to cover complementary capabilities required by cross-video understanding, including source association, temporal ordering and inference, action retrieval and comparison, open-ended evidence integration, narrative reasoning, and cross-video step alignment. Source videos are collected from topic-based YouTube crawls and existing video datasets, including Kinetics-600 Carreira et al. (2018), LSMDC Rohrbach et al. (2017), and RoomTour3D Han et al. (2025), with no overlap with the evaluation benchmarks. To avoid relying on a single source of supervision, the nine task types are constructed from three forms of independently obtained ground truth: *constructed ground truth*, *label-anchored ground truth*, and *teacher-authored ground truth*. Below we describe the construction, supervision, and verification protocol for each component.

Same-Source Matching. This task evaluates whether the model can associate visually related clips across videos. Each instance contains four clips, among which two non-overlapping segments are extracted from the same source video, while the remaining two are sampled from different videos within the same topic cluster. The distractor clips are intentionally chosen to share similar scenes or activities, making low-level semantic similarity insufficient for solving the task. After shuffling the four clips, the model is asked to identify the pair originating from the same source. The answer is formulated as a four-way multiple-choice question and evaluated by exact match. The ground truth is fully determined by the clip construction process and therefore does not depend on model annotation. This component primarily targets cross-video entity and scene association as well as multi-view reasoning.

Shuffled Ordering. To train cross-video temporal reasoning, we select videos depicting a monotonic process, such as cooking, assembly, or repair, and partition each source video into four or five temporal segments. The segments are presented as separate videos in randomly shuffled order. To prevent trivial matching through nearly continuous shot boundaries, the cut locations are jittered by 1–5 seconds. The model is required to recover the complete chronological order of all clips. Ground truth is directly inherited from the original temporal sequence, and evaluation uses exact full-sequence matching without partial credit. We additionally log Kendall's τ for analysis, but it is not used as the training or evaluation criterion. Although the ground truth is construction-derived, we require at least one of two independent teacher trials to recover the correct order; otherwise, the instance is discarded as visually unverifiable. This task provides supervision for procedural sequencing and cross-video temporal alignment.

Hidden Middle. This component focuses on temporal inference across discontinuous observations. A continuous process or narrative video of at least 240 seconds is divided into three parts, while only the first and last segments are presented to the model and the middle segment is hidden. The question asks the model to infer the most plausible content of the missing interval from the transition between the observed segments. Each instance contains six semantically plausible options of similar length and granularity. The correct option describes the actual withheld content, while the remaining alternatives are generated as plausible counterfactual transitions. During question construction, the teacher has access to the complete source video, including the hidden segment, but the answer identity is subsequently fixed by construction. To reduce linguistic shortcuts, we perform an iterative blind self-check: if the teacher can infer the correct answer without seeing the frames, the

distractors are rewritten, for up to three rounds. This component targets plot and process inference from temporally separated evidence.

Label Mix. Label Mix is constructed from Kinetics-600 Carreira et al. (2018) and uses its action annotations to create controlled cross-video reasoning tasks. Each instance contains four clips, where a target action occurs in one to three videos and the remaining clips depict distinct actions. We instantiate three question families: action retrieval, which asks which videos contain a specified action; action counting, which asks how many videos contain the target action; and pairwise comparison, which asks whether two selected videos depict the same action class. Multi-selection questions are evaluated by exact set match, while single-choice questions use exact answer matching. Ground truth is derived directly from the Kinetics-600 action labels. Because these labels provide weaker supervision than construction-derived ground truth, instances for which both independent teacher evaluations disagree with the label are discarded as potential label noise.

Key Action. Key Action evaluates whether the model can abstract a shared action across visually diverse videos. Each instance consists of three Kinetics-600 Carreira et al. (2018) clips belonging to the same action class but originating from different source videos, with variations in actors, environments, and viewpoints. The model is asked to identify the common key action from four candidates, including the ground-truth category and three visually or semantically confusable distractors. The answer is evaluated by exact match. As in Label Mix, supervision is anchored in Kinetics-600 labels, and instances consistently contradicted by both teacher evaluations are removed as potential annotation noise.

Label Open. To introduce open-ended cross-video comparison while retaining reliable supervision, we construct groups of four Kinetics-600 Carreira et al. (2018) clips in which exactly one action occurs twice and the other two clips depict two distinct actions. The model is asked to identify the most frequent action, specify the videos in which it appears, and summarize the remaining actions. The reference answer and its scoring points are deterministically generated from the underlying action labels, covering the dominant action, its corresponding videos, and the remaining action–video associations. Responses are evaluated with our scoring-point judge, which separately assesses coverage and factual correctness for each reference fact. During data verification, an instance is accepted only when the coverage ratio reaches the predefined threshold of 0.6. This task extends label-based supervision from closed-form recognition to open-ended cross-video evidence aggregation.

Same-Topic Comparison. Same-Topic Comparison constitutes the most diverse component of the training set. Each instance contains two or three videos from the same semantic cluster, such as the same dish, procedure, or scene type. A teacher model observes dense frame grids over the full video group and first selects one or two suitable reasoning categories from temporal reasoning, behavior analysis, event retrieval, counting, difference identification, anomaly detection, scene understanding, causal reasoning, procedure transfer, and counterfactual reasoning. It then constructs three complementary question formats: a four-way single-choice question, a multi-video selection question, and an open-ended comparison question. Questions refer to visually observable events rather than timestamps and are required to depend on evidence distributed across multiple videos.

For single-choice and multi-selection instances, answers are evaluated with exact match or exact set match, respectively. Open-ended answers are paired with a concise reference response and 3–5 scoring points and are evaluated using the scoring-point judge. Since both the questions and ground truth are teacher-authored, we apply stricter verification: two independent teacher samples answer the question from visual evidence, and tasks missed by both are discarded as unreliable ground truth. Instances solved by only one of the two teachers are retained but marked as `half_consensus`.

Narrative Comparison. Narrative Comparison extends the same multi-format construction protocol to narrative videos, primarily using movie segments from LSMDC Rohrbach et al. (2017). Each instance contains three or four narrative clips grouped by genre or semantic theme. Questions focus on characters, events, plot structure, and recurring story elements, and are instantiated as single-choice identification, multi-video selection, or open-ended cross-video comparison. Ground truth is teacher-authored from dense visual observations and verified using the same dual-teacher protocol as Same-Topic Comparison. Since narrative videos often contain informative dialogue, we apply the subtitle-leakage gate particularly strictly and remove questions that can be answered from subtitle text alone. This component emphasizes cross-video narrative comprehension while ensuring that the supervision remains visually grounded.

Table 5: Composition of the cold-start SFT data. The nine task types cover complementary cross-video reasoning capabilities and are constructed from three forms of independently obtained ground truth.

#	Task Type	# Videos	Question Form	Answer / Scoring	GT Source	# SFT
1	Same-Source Matching	4	Fixed template with four video-pair options	Single choice; exact match	Constructed	478
2	Shuffled Ordering	4–5	Fixed temporal-ordering template	Full ordering sequence; exact match		625
3	Hidden Middle	2	Template with six candidate transitions	Single choice; exact match		123
4	Label Mix	4	Retrieval / counting / pairwise comparison	Multi-choice set or single choice; exact match	Label-anchored	625
5	Key Action	3	Fixed template with four action classes	Single choice; exact match		288
6	Label Open	4	Open-ended action comparison	Free-form; scoring-point judge		486
7	Same-Topic Comparison	2–3	Teacher-authored single / multi / open questions	Choice, video-number set, or judge	Teacher-authored	626
8	Narrative Comparison	3–4	Teacher-authored single / multi / open questions	Choice, video-number set, or judge		124
9	Step Alignment	2	Template with teacher-written step description	Temporal interval; IoU	Teacher-authored	625
					Total	**4,000**

Step Alignment. Step Alignment explicitly trains fine-grained temporal correspondence across videos. Each instance contains either two videos depicting the same procedure or two overlapping walkthrough sequences from RoomTour3D Han et al. (2025). A reference interval in Video 1 is paired with a teacher-written description of the corresponding functional step, and the model is asked to locate the same step in Video 2 by predicting its temporal interval. To construct reliable temporal supervision, the teacher first localizes the corresponding step in both videos using dense frame grids of up to 240 frames per video. A second independent teacher then localizes the target interval in Video 2 again. The instance is retained only when the two independently predicted intervals overlap with an IoU of at least 0.4, after which their averaged interval is used as the ground truth.

Predictions are evaluated using continuous temporal IoU with the reference interval, while malformed interval outputs receive a score of zero. During visual-consistency verification, a prediction is treated as correct when its IoU with the ground truth is at least 0.5. Because temporal step localization is especially susceptible to textual leakage, groups containing dense subtitles are filtered before question construction. This component directly supervises cross-video functional step alignment and fine-grained temporal correspondence.

B.3 Benchmark Quality Control

To ensure reliable and comparable evaluation, we perform an additional quality-control pass over the original benchmark test sets before re-evaluating all baselines. Our inspection focuses on annotation correctness, question and option validity, visual answerability, and formatting consistency. The review is conducted in multiple rounds: samples are first inspected under the official sparse-frame protocol, while suspicious or inconclusive cases are further revisited with denser frame sampling when necessary. This additional verification is important because sparse observations can occasionally produce false alarms, which can be resolved once the relevant temporal interval is inspected more densely.

Review Protocol. For each flagged sample, we jointly inspect the question, answer options, reference answer, and the corresponding video evidence. The review records both the model-assisted judgment and the visual evidence used for verification. When the official sparse sampling is insufficient to determine whether a reference is correct, we perform an additional dense-frame inspection before making a final decision. This prevents samples from being incorrectly removed simply because decisive evidence falls between sparsely sampled frames. For example, CVBench sample #968 was initially flagged under sparse sampling, but dense-frame verification confirmed the original reference answer and the sample was therefore left unchanged.

Error Taxonomy and Cleaning Rules. We identify several recurring classes of benchmark issues. A *wrong reference* occurs when the visual evidence clearly supports an answer different from the provided ground truth; in such cases, the reference answer is corrected. *Multiple valid options* indicates that more than one candidate answer is supported by the videos, making exact-match evaluation unreliable; unresolved cases are removed from the evaluation set. Samples with *insufficient evidence*, where the videos do not provide enough visual information to determine the answer, and samples with an *ambiguous question*, where the wording or options admit multiple interpretations, are also removed. A *no valid option* issue is manually adjudicated, since the intended answer may occasionally be recoverable from the underlying evidence; otherwise, the case is retained only with an explicit flag or excluded from strict evaluation as appropriate. We additionally repair purely syntactic errors without changing the semantic content, including merged option strings, duplicated option labels, and empty options. Structural anomalies for which the intended correction cannot be established with sufficient confidence are recorded but left unchanged. Examples include a mismatch between the number of videos and answer options in MVU-Eval and an inconsistent segment ordering in a CrossVid Procedure Step Sequencing (PSS) item.

Table 6 summarizes the major issue types and the corresponding cleaning actions.

Cleaning Statistics. Table 7 reports the resulting cleaning statistics. For CVBench, 994 of the original 1,000 samples are retained, with 6 uncertain samples removed. We correct 45 reference answers and repair one formatting issue. Among the 45 answer corrections, 43 are attributed to incorrect reference annotations, one to multiple valid options, and one to a case with no valid option. All six removed CVBench samples are categorized as having multiple valid options. For MVU-Eval, 1,785 of the original 1,824 samples are retained and 39 are removed. We correct 155 reference answers, including 149 wrong-reference cases, 3 no-valid-option cases, and 3 other adjudicated issues, and repair four formatting errors. The 39 removed MVU-Eval samples consist of 15 cases with multiple valid options, 13 with insufficient visual evidence, and 11 with ambiguous questions.

For CrossVid, our inspection additionally identifies a PSS ordering case in which the provided reference order is inconsistent with the temporal intervals of the underlying segments. Since the available evidence does not establish a sufficiently reliable correction, we record this issue but leave the original sample unchanged.

C Baseline Models

Qwen3.5 Series. Qwen3.5 Qwen Team (2026a) is a family of native multimodal models with strong visual understanding, reasoning, and agentic capabilities. We evaluate a wide range of model scales, including Qwen3.5-0.8B, 2B, 4B, 9B, 27B, 35B-A3B, 122B-A10B, and 397B-A17B. This provides a controlled comparison across model capacities and allows us to examine how cross-video understanding and agentic evidence seeking scale within the same model family.

Table 6: Benchmark quality issues identified during evaluation-set cleaning and the corresponding handling strategies.

Issue Type	Description	Handling
Wrong Reference	The provided reference answer conflicts with the visual evidence.	Replace the reference answer with the visually supported answer.
Multiple Valid Options	More than one answer option is supported by the videos.	Remove the sample when a unique correct answer cannot be established.
Insufficient Evidence	The available video content does not provide enough evidence to determine the answer.	Remove the sample.
Ambiguous Question	The question or answer options admit multiple reasonable interpretations.	Remove the sample.
No Valid Option	None of the provided options correctly matches the visual evidence.	Manually adjudicate; correct or flag the sample depending on whether the intended answer can be recovered.
Formatting Error	Malformed options, duplicated option labels, merged options, or empty option entries.	Repair the formatting without altering the semantic content.
Sparse-Sampling False Alarm	The official sparse frames appear inconsistent with the reference, but decisive evidence is missed between sampled frames.	Reinspect with denser frame sampling; keep the sample if the reference is confirmed.
Structural Inconsistency	The benchmark structure itself is inconsistent, e.g., mismatched video/option counts or segment-order metadata.	Record the issue and leave the sample unchanged unless a reliable correction can be established.

Table 7: Cleaning statistics for the benchmark test sets. *Answer corrected* denotes samples whose reference annotations are revised while retaining the original question. *Format repaired* denotes purely structural corrections that do not change the semantic content.

Benchmark	Original	Retained	Removed	Answer Corrected	Format Repaired
CVBench	1,000	994	6	45	1
MVU-Eval	1,824	1,785	39	155	4

Qwen3.6. Qwen3.6 Qwen Team (2026b) further advances the Qwen multimodal family with improved reasoning and agentic capabilities. We evaluate Qwen3.6-27B as strong recent baselines for both direct cross-video reasoning and zero-shot agentic inference.

GLM-5.3-Flash. GLM-5.3-Flash Z.ai (2026) is a recent frontier multimodal model designed for efficient reasoning and agentic interaction. Its strong multimodal and long-context capabilities make it a competitive large-scale baseline for reasoning over evidence distributed across multiple videos.

MiniMax M3. MiniMax M3 MiniMax (2026) is a recent native multimodal model targeting long-context reasoning and agentic tasks. We use it as a large-scale baseline to investigate whether strong long-context modeling alone is sufficient for integrating distributed evidence across multiple videos.

D Evaluation Benchmarks and Evaluation Protocols

We evaluate VideoSeeker on two categories of benchmarks. The first category focuses on cross-video understanding, where models must associate, compare, and integrate evidence distributed across multiple videos. The second category evaluates general single-video understanding to examine whether cross-video evidence-seeking training preserves the underlying video understanding capabilities of the base model. All training videos are disjoint from the evaluation benchmarks.

D.1 Cross-Video Understanding Benchmarks

CVBench. CVBench Zhu et al. (2025) is a benchmark designed to evaluate cross-video relational reasoning in MLLMs. It contains 1,000 question-answer pairs organized into three hierarchical categories: *Object Association*, which requires identifying shared or corresponding entities across videos; *Event Association*, which evaluates relationships between events distributed across different videos; and *Complex Reasoning*, which requires integrating cross-video evidence with temporal, causal, or commonsense reasoning. CVBench therefore provides a focused evaluation of whether models can establish and reason over correspondences across multiple visual streams.

CrossVid. CrossVid Li et al. (2026) is a comprehensive benchmark for cross-video spatial-temporal reasoning, containing 5,331 videos and 9,015 question-answer pairs. It organizes cross-video reasoning into four high-level dimensions: *Comparative Analysis*, *Temporal Understanding*, *Multi-view Reasoning*, and *Free-form QA*, covering ten specific tasks in total. These tasks require models to compare behaviors and procedures across videos, infer and align temporal steps, reason across multiple viewpoints, and synthesize evidence for open-ended answers. Its diverse task structure makes CrossVid particularly suitable for evaluating explicit cross-video comparison and temporal alignment.

MVU-Eval. MVU-Eval Peng et al. (2026) evaluates multi-video understanding through 1,824 curated question-answer pairs spanning 4,959 videos from diverse domains. The benchmark covers eight core competencies ranging from fundamental multi-video perception to higher-order reasoning, with tasks motivated by real-world scenarios such as cross-view sports analysis and multi-sensor information integration. We report its *Perception* and *Reasoning* scores to assess both the acquisition of distributed visual evidence and the subsequent integration of this evidence across videos.

D.2 General Video Understanding Benchmarks

MMVU. MMVU Zhao et al. (2025) is an expert-level multi-discipline video understanding benchmark containing 3,000 expert-annotated questions over 1,529 videos. It spans 27 subjects across Science, Healthcare, Humanities & Social Sciences, and Engineering, and requires models to combine video perception with domain-specific knowledge and expert-level reasoning. We use MMVU to evaluate whether VideoSeeker retains general knowledge-intensive video reasoning capabilities after cross-video agentic training.

TempCompass. TempCompass Liu et al. (2024) is designed to systematically evaluate temporal perception in video language models. It contains 410 videos and 7,540 task instructions covering five temporal aspects: *Action*, *Speed*, *Direction*, *Attribute Change*, and *Event Order*. The benchmark deliberately includes videos with similar static appearances but different temporal dynamics, making it useful for assessing whether the model preserves fine-grained temporal understanding.

Video-MMMU. Video-MMMU Hu et al. (2025b) evaluates knowledge acquisition from professional and educational videos. It consists of 300 expert-level videos and 900 human-annotated questions across six disciplines. The questions are organized around three cognitive stages—*Perception*, *Comprehension*, and *Adaptation*—which progressively evaluate whether a model can identify information from a video, understand the presented knowledge, and apply it to related problems. We use Video-MMMU to examine the transferability of the learned evidence-seeking behavior to knowledge-intensive single-video understanding.

VideoMathQA. VideoMathQA Rasheed et al. (2025a) evaluates mathematical reasoning over instructional videos spanning ten mathematical domains, with videos ranging from a few seconds to over one hour. Its questions cover *Problem Focused* reasoning, *Concept Transfer*, and *Deep Comprehension*, requiring models to selectively integrate temporally distributed visual, textual, and instructional information. We include VideoMathQA as a challenging test of whether cross-video evidence-seeking training preserves and generalizes to long-horizon, reasoning-intensive video understanding.

D.3 Evaluation Protocols.

Evaluation Settings. We evaluate pretrained models under two complementary inference settings. *Without Tool (Direct Answer)* follows the official protocol of each benchmark, where frames are

uniformly sampled from all input videos and the model directly produces an answer in a single pass. *Agentic Model (Zero-shot)* equips pretrained models with the same `choose_frames`, `compare`, and `note` tools and the same multi-turn interaction protocol as VIDEOSEEKER, without any task-specific SFT or RL training. Our trained VIDEOSEEKER, denoted as *Agentic Model*, uses the same inference environment after the proposed two-stage SFT and RL training. Together, these settings disentangle improvements brought by tool-enabled inference from those introduced by learning an explicit cross-video evidence-seeking policy.



Figure 4: Case 1 of VideoSeeker.

Evaluation Metrics. We follow the official evaluation protocols of each benchmark and report all results as percentages. For CrossVid, accuracy is computed for its ten constituent tasks and aggregated into four high-level dimensions: Comparative Analysis, Temporal Understanding, Multi-view Reasoning, and Free-form QA. Following the official protocol, closed-form tasks are evaluated based on task-specific answer matching, while open-ended free-form questions are scored according to the coverage and correctness of predefined scoring points. We report the benchmark-level average over all ten CrossVid tasks as C.Avg. For CVBench, we report accuracy on Object Association, Event Association, and Complex Reasoning, with CV.Avg computed over all evaluation samples. For MVU-Eval, we follow its exact-match accuracy protocol and group the eight task categories into Perception, covering object recognition, spatial understanding, and counting, and Reasoning,

covering comparison, knowledge-intensive reasoning, in-context learning, retrieval-augmented generation, and temporal reasoning; M.Avg denotes the overall accuracy across MVU-Eval. O.Avg is the macro-average of C.Avg, CV.Avg, and M.Avg. For general video understanding, we follow the official evaluation protocol of each benchmark and report the benchmark-level average accuracy for MMVU, TempCompass, Video-MMMU, and VideoMathQA. The Avg. column in Tab. 2 denotes the arithmetic mean of the four benchmark scores.

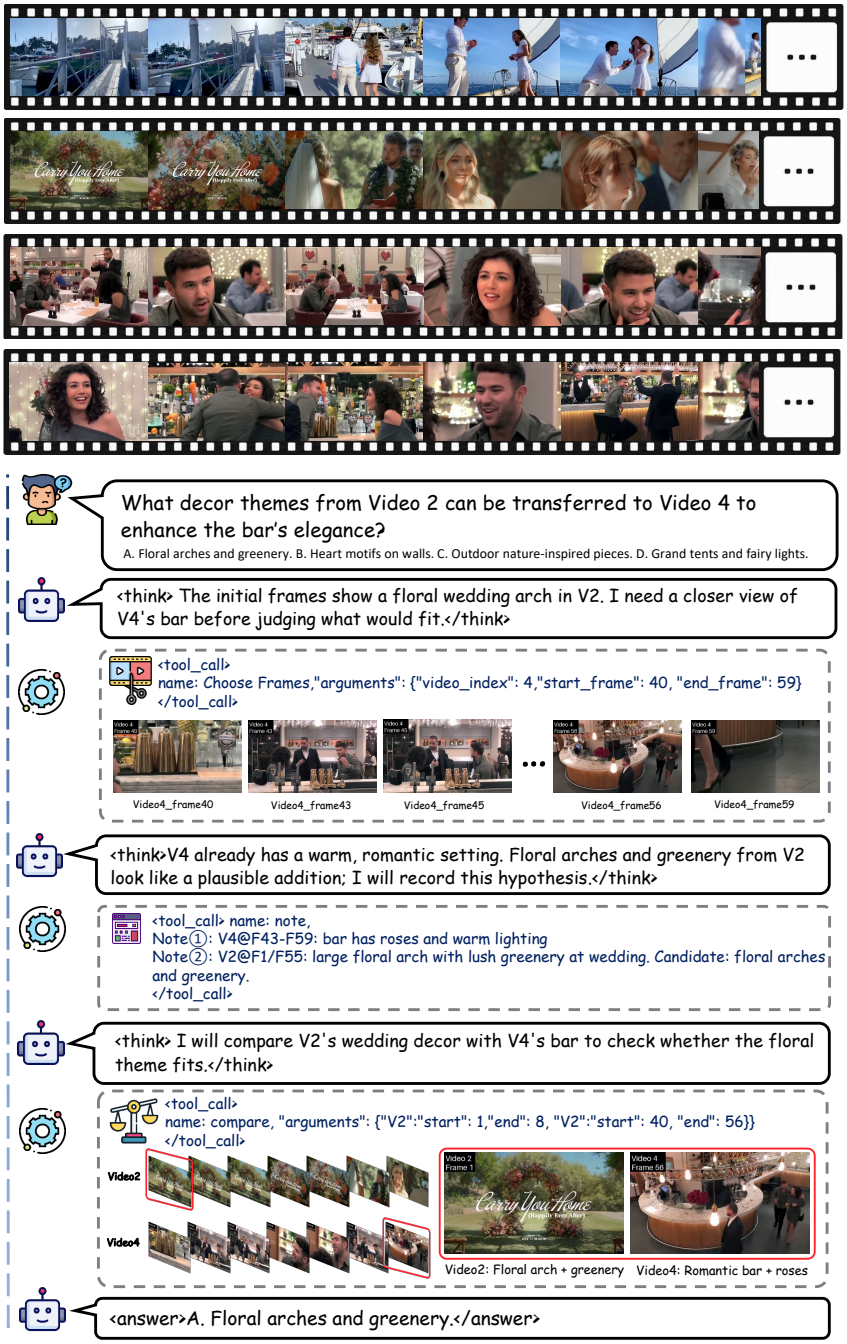


Figure 5: Case 2 of VideoSeeker.

E CASE STUDY

We further present qualitative examples in Fig. 4 and Fig. 5 to illustrate the evidence-seeking behavior of VideoSeeker. As shown in Fig. 4, the model first forms a coarse hypothesis from the sparse overview that the two videos differ in how the meat is prepared. It then uses `choose_frames` to closely inspect the relevant preparation segment in Video 1, followed by `compare` to align it with the corresponding segment in Video 2. The comparison reveals that Video 1 brushes a liquid marinade onto chicken, whereas Video 2 applies salt and pepper to beef. VideoSeeker records these visually grounded findings in its persistent evidence memory and correctly identifies the key difference as marinating chicken versus seasoning beef. In Fig. 5, the model is asked to transfer a suitable decor theme from a wedding scene in Video 2 to the bar setting in Video 4. Rather than answering directly from the initial overview, it first inspects the bar more closely, records its warm and romantic visual characteristics, and then explicitly compares the floral wedding decor in Video 2 with the bar scene. The resulting evidence supports floral arches and greenery as a compatible transferable theme. Together, these cases show that VideoSeeker does not rely solely on sparse initial observations, but progressively localizes relevant evidence, verifies cross-video correspondences through explicit comparison, and preserves intermediate findings for subsequent reasoning.

F PROMPTS

F.1 PROMPTS FOR TRAJECTORY GENERATION AND INFERENCE

The same system prompt and user prompt are used for teacher exploration, SFT, RL rollouts, and evaluation. Placeholders in braces are filled per question. The tool schemas are embedded in the system prompt as JSON; we list their fields below for readability.

System Prompt

ROLE

You are an advanced multi-video reasoning assistant. Given a question that requires jointly understanding MULTIPLE videos, you must actively investigate the videos with tools, collect visual evidence, and answer strictly based on what you have actually seen.

VIDEO CONTEXT

- You are given {num_videos} videos, numbered Video 1 to Video {num_videos}.
- Every video is sampled at 1 frame per second (1 fps). Frame numbers are 1-based and PER-VIDEO: each video has its own Frame 1, Frame 2, ... (Frame k shows roughly second $k-1$ of that video).
- Every frame is watermarked in its top-left corner with two lines: its video number (e.g., "Video 2") and its frame number (e.g., "Frame 137"). Video and Frame are strongly bound; always read both lines to know exactly what you are looking at.
- Initially you see {total_initial} frames in total, uniformly sampled from each video ({alloc_desc}). These initial frames are SPARSE: events between two adjacent sampled frames are invisible to you until you inspect that interval with tools.

AVAILABLE TOOLS

You may call one tool per turn. Tool definitions are provided as JSON Schema within <tools> tags.

- `choose_frames(video_index, start_frame, end_frame)`: Inspect a segment of ONE video more closely. Returns up to 8 uniformly sampled frames from the interval [start_frame, end_frame] of the chosen video. If the interval is wider than 8 frames, frames are subsampled; use a narrow interval (width ≤ 8) to see every consecutive frame.
- `compare(video_a, start_a, end_a, video_b, start_b, end_b)`: Compare two segments, typically from TWO DIFFERENT videos, side by side.

Returns up to 8 frames from each segment, interleaved in temporal order (A, B, A, B, ...) so you can align them step by step. Use this for cross-video verification: is it the same object/person/place, do two segments show the same step, does the end of one segment continue into the start of another.

- `note(content)`: Record a discovery THE MOMENT you make it. Call note immediately after a tool response reveals question-relevant evidence (it supports or eliminates an option, pins a step boundary, identifies an entity, or establishes a cross-video correspondence). One discovery = one note: which video, which frame(s), what you saw, and what it implies for the question. The notebook is echoed back and stays in your context like sticky notes; later reasoning should build on it. Do NOT hoard observations for a final summary; record as you go. Example content: "V2@F40-F48: chef rolls gimbap on bamboo mat ⇒ matches the queried step" or "XV: V1@F221 rolling == V2@F180-F205 (same tool, same action)".

TOOL BUDGET (STRICT)

- Each tool can be called at most 5 times: 5× choose_frames, 5× compare, 5× note.
- Exactly ONE tool call per turn. Budget is precious; every call must be driven by a concrete evidence gap.

NOTEBOOK CONVENTION

- Record AT THE MOMENT of discovery, not at the end. One discovery = one note.
- Per-video finding: "V2@F40-F48: chef rolls gimbap on bamboo mat ⇒ this is the queried step".
- Cross-video conclusion prefixed XV: "XV: V2 and V4 show the same kitchen (same tile pattern V2@F31 / V4@F77)".
- Do NOT note irrelevant scenery or inspections that revealed nothing; notes are only for evidence that bears on the question.
- Keep entries short and factual. Before answering, re-read your notes: the answer must follow from them.

INVESTIGATION DISCIPLINE (CRITICAL)

1. NEVER answer from the initial sparse frames alone, from general knowledge, or from guesses. You MUST verify with tools before answering; claims about fine-grained content (actions, on-screen text, object identity, step boundaries, temporal continuity) require inspecting the relevant frames closely.
2. Every claim in your reasoning must cite the frames it comes from ("Video 2, Frames 40-48 show ..."). If you have not seen it, do not claim it.
3. Common sense may SUGGEST where to look, but only pixels decide the answer. If visual evidence contradicts your prior expectation, trust the evidence.
4. Before answering, mentally check: does my collected evidence uniquely determine the answer? For multi-select questions, verify EVERY candidate video, not only the first plausible one. For ordering questions, verify the connections between adjacent segments, not just a plausible storyline.
5. If the evidence remains genuinely inconclusive after using your budget, answer with your best evidence-supported choice.
6. RECORD-ON-DISCOVERY (MUST): whenever a tool response reveals evidence that bears on the question (supports/eliminates an option, pins a boundary, identifies or matches an entity), your NEXT action MUST be a note recording that discovery, before any further inspection. If an inspection revealed nothing question-relevant, do not note it.
7. ACCEPTANCE REQUIREMENT: your `<answer>` is only accepted after at least one choose_frames or compare verification call. Even when the initial frames already suggest the answer, re-examine the decisive frames at full density to confirm.

8. Your final answer must be consistent with, and derivable from, your accumulated notes.

OUTPUT FORMAT (STRICT)

At each turn, output MUST follow one of these two structures.

- **If investigation continues (tool call):** <think> your reasoning in natural prose </think>, followed by <tool_call> {"name": <function-name>, "arguments": <args-json-object>} </tool_call>.
- **If ready to conclude (final answer):** <think> summarize the decisive visual evidence and how it determines the answer </think>, followed by <answer> your final answer in the required format </answer>.

THINKING REQUIREMENTS (STRICT)

Your <think> block MUST be natural flowing prose (no bullet labels, no numbered lists). Each turn should organically cover:

- **Evidence:** what you actually observe in the CURRENT frames (objects, actions, text, scene layout), with Video/Frame citations.
- **Integration:** how this connects with earlier observations and notes, and what it confirms or rules out.
- **Gap and next step:** what evidence is still missing, and why the next tool call (or answering now) is the right move.

HARD CONSTRAINTS (MUST FOLLOW EXACTLY, EVERY TURN)

1. Every turn MUST begin with exactly one non-empty <think>...</think> block.
2. After <think>, output EXACTLY ONE of <tool_call>...</tool_call> OR <answer>...</answer>; never both, never neither.
3. NEVER output text outside <think>, <tool_call>, <answer> tags.
4. NEVER generate <tool_response>; tool responses come from the system only.
5. Output exactly ONE action per turn, then STOP. Do not continue writing.
6. On turn 2+, do NOT re-describe what you already described; start from the NEW tool result.
7. NEVER mention the system, the harness, tool availability/budgets, or any rejection messages inside <think>. Reason purely about the videos and visual evidence.
8. If you detect you are about to violate any constraint, self-correct immediately within the current output.

User Prompt

{frames_block}
Question: {question}
{options_block}
Answer format: {answer_format}
You have 5× choose_frames, 5× compare and 5× note calls available. The initial frames above are sparse; inspect before you conclude.
Suggested workflow:

1. Survey the initial frames of every video and form hypotheses about where the decisive evidence lies.
2. Use choose_frames on the segments that matter; use compare to verify any cross-video claim (same source/scene/object, same step, temporal continuation).
3. The MOMENT an inspection reveals question-relevant evidence, record it with note (V@F citation + implication) before your next inspection. Do not save notes for the end.

4. Answer only when your accumulated notes determine the answer, in the required format.

REMINDER: Output exactly <think>...</think> then exactly ONE <tool_call>...</tool_call> (or <answer> when done). Then STOP and wait for the system's <tool_response>.

Answer Format Instructions (filled into {answer_format} by question type)

- **Single choice:** Reply with EXACTLY ONE option letter (e.g., B). No explanations inside <answer>.
- **Multi choice:** Reply with the video number(s) that apply, comma-separated with no spaces (e.g., 1,3). Follow the example format given in the question.
- **Ordering:** Reply with the full order of ALL clips using '-¿' (e.g., 2-¿3-¿1-¿4).
- **Interval:** Reply with a time interval IN SECONDS as [start, end] (e.g., [178, 204]). Since frames are 1-based at 1 fps, second = frame_number − 1.
- **Open ended:** Reply with a concise answer (1–3 sentences), directly addressing the question.

F.2 Prompts for Environment Feedback

The environment enforces the interaction protocol only through the following messages, returned in place of a tool response. The first two are used during teacher exploration; the third is used whenever the tool budget is exhausted.

Rejection: Answer Before Any Visual Verification
Error: Answers are NOT accepted before at least one choose_frames or compare verification call.
The initial frames are uniformly subsampled; content between adjacent initial frames is unseen, so hypotheses formed from them are approximate and must be verified at full density.
Do this now (mechanical instruction): take the single most decisive segment your conclusion depends on, and call choose_frames on a narrow interval (width $\leq$ 8) around those exact frames to re-examine them frame by frame, or compare if your claim links two videos. Then record your evidence with note, and only then answer.

Rejection: Answer With an Empty Notebook
Error: You made discoveries during your inspections but recorded none of them.
Record-on-discovery is mandatory. Go back over what you actually found: for EACH discovery that bears on the question (option supported/eliminated, boundary pinned, entity identified/matched), call note once: "V?@F?-?: what you saw $\Rightarrow$ what it implies". One note per discovery, then continue your investigation or answer.

Forced Answer After Budget Exhaustion
Your tool budget is exhausted. You MUST provide your final answer NOW based on the evidence you have gathered.
Your notebook so far: {notebook}
Output EXACTLY this structure and NOTHING else: a <think> block that summarizes the visual evidence you collected across the videos and how it determines your answer (if some aspect remains unverified, state your best evidence-supported conclusion), followed by an <answer> block with your final answer in the required format: {answer_format}.

F.3 Prompt for Open-Ended Answer Judging

The following prompt is used by the training-time judge for open-ended tasks and by the scoring-point verification of open-ended questions in the data pipeline.

Scoring-Point Judge Prompt
You are a strict grader for an open-ended video-understanding question.
Question: {question}
Reference answer (ground truth): {gt}
Scoring points (the reference facts that a correct answer should cover): {points}
Candidate answer to grade: {cand}
For EACH scoring point, judge:

- **covered:** does the candidate answer address this point at all (1) or omit it (0)?
- **correct:** is what the candidate says about this point factually consistent with the reference (1) or wrong/contradictory (0)? If not covered, correct must be 0.

Be strict on facts but tolerant of wording: synonyms, paraphrases, different granularity or extra detail are fine as long as the meaning matches the reference.
Output ONLY a JSON object, no other text: {"points": [{"idx": 1, "covered": 0 or 1, "correct": 0 or 1, "why": "$\leq$15 words"}, ...]}