VIDEOINSPECTOR: AGENTIC MULTI-VIDEO REASONING WITH COMPOSITIONAL SEGMENT TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models have made substantial progress in video understanding, yet most are trained on single-video instruction data and struggle with multi-video reasoning. Such reasoning requires models to retrieve relevant videos, localize crucial moments, and aggregate evidence across different sources. However, these intermediate capabilities are rarely explicitly supervised by existing video training data, leading models to rely on coarse representations rather than grounded cross-video evidence. We introduce **VIDEOINSPECTOR**, an agentic framework for multi-video reasoning with tool-augmented segment inspection. Given a set of videos, VIDEOINSPECTOR performs coarse-to-fine reasoning by selecting relevant candidates, inspecting fine-grained temporal segments, and composing grounded evidence for final decisions. To train this model, we construct **VIDEOINSPECTOR-INSTRUCT**, a large-scale segment-level compositional multi-video dataset. Starting from self-collected videos with dense temporal descriptions, we decompose videos into reusable semantic segments and recombine them into multi-video instances across four tasks: cross-video retrieval, cross-video temporal grounding, clip ordering, and video comparison. We further annotate tool-augmented reasoning trajectories and adopt a two-stage training pipeline: cold-start supervised fine-tuning teaches the model to imitate structured reasoning and tool use, while agentic reinforcement learning refines its policy for retrieval, inspection, and termination through exploration. Experiments on multi-video reasoning benchmarks show that VIDEOINSPECTOR improves over its non-tool variant on CrossVid, particularly in candidate analysis and temporal understanding. These results suggest that compositional segment supervision and tool-augmented reasoning provide a promising direction for building more grounded multi-video understanding systems.

1 INTRODUCTION

Multi-video understanding is becoming increasingly important in the era of multimodal AI, driven by a wide range of real-world applications such as surveillance (Sultani et al., 2018; Yuan et al., 2024), video editing (Li et al., 2025; Deng et al., 2026) and content analysis (Burgess et al., 2025). Unlike single-video understanding (Fu et al., 2025; Wu et al., 2024; Deng et al., 2025), multi-video understanding (Li et al., 2026; Zhu et al., 2025b; Deng et al., 2026) requires models to reason over a collection of videos, where the relevant evidence may be distributed across different sources and temporal locations. For example, a user may need to compare multiple camera views of the same event, determine whether two clips depict the same scene or originate from the same movie, or retrieve a relevant moment from a pool of candidate videos. These scenarios require not only visual perception within individual videos, but also cross-video retrieval, temporal localization, semantic alignment, and evidence aggregation. Therefore, multi-video reasoning cannot be solved by simply concatenating videos or increasing the input context length; it requires models to actively identify, inspect, and compose sparse evidence distributed across multiple sources.

Despite its practical importance, multi-video understanding remains relatively underexplored. Recent multimodal large language models (MLLMs) (Gemini Team, 2025; OpenAI, 2024; Qwen Team, 2025; Zhu et al., 2025a) have demonstrated strong capabilities in video understanding and reasoning. However, most existing models are primarily trained on single-video instruction data. This training

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.

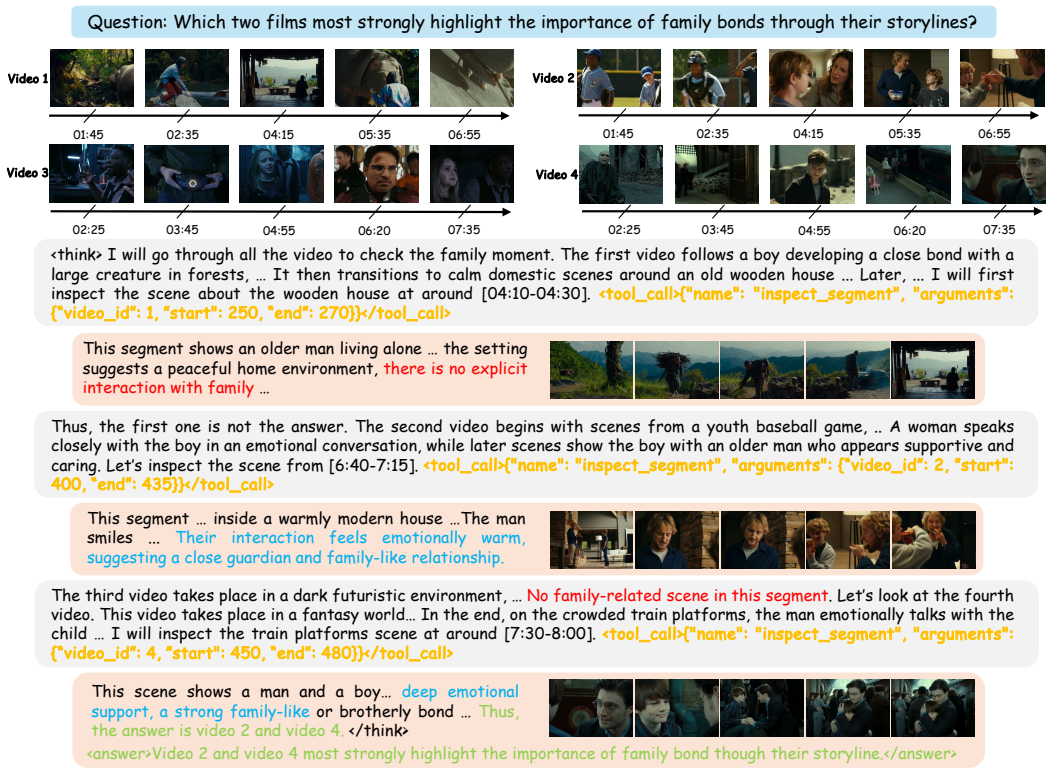


Figure 1: **Illustration of the tool-augmented Chain-of-Thought in our VIDEOINSPECTOR.** VIDEOINSPECTOR can natively invoke the *inspect_segment(video_id, start, end)* tool to perform temporally grounded perception and reasoning across multiple videos. By adaptively inspecting candidate segments during the reasoning process, the model progressively gathers fine-grained visual evidence to support cross-video understanding. Since the retrieved evidence is explicitly grounded in visual observations, the model can self-correct when its initial reasoning or localization is inaccurate, enabling more reliable multi-video reasoning and decision making.

paradigm provides limited supervision for interactions among multiple video sources. As a result, when applied to multi-video scenarios, models often rely on coarse video-level representations or superficial correlations, and struggle to identify fine-grained key visual evidence. Recent multi-video benchmarks (Zhu et al., 2025b; Li et al., 2026; Deng et al., 2026) further reveal a substantial gap between current video training practices and the requirements of multi-video reasoning.

To bridge this gap, we argue that effective multi-video understanding requires several fundamental capabilities that are not explicitly covered by existing video instruction datasets. First, the model must retrieve relevant videos from a set of candidates, since only a subset of videos may contain useful information for a given query. Second, it must localize crucial moments within each video because the visual evidence is often temporally sparse. Third, it must reason over visual evidence across multiple videos, including comparing and aggregating information from different sources. Based on this observation, we formulate multi-video learning around four representative tasks: 1) *cross-video retrieval*, which trains video-level candidate selection; 2) *cross-video temporal grounding*, which requires fine-grained localization of events within videos; 3) *clip reordering*, which emphasizes temporal reasoning across multiple segments and sources; and 4) *video comparison*, which focuses on semantic relationships such as similarity, difference, and correspondence between videos.

Beyond task design, the aforementioned capabilities inherently form a multi-step reasoning process. Training models only with final answers may encourage shortcut learning, where predictions are made from dataset biases or coarse correlations rather than grounded visual evidence. To provide structured supervision, we construct **VIDEOINSPECTOR-INSTRUCT**, a large-scale compositional segment-level dataset for multi-video reasoning. Starting from a self-collected video corpus, we use Gemini 2.5 Pro to generate dense captions for each video, and decompose videos into reusable

semantic segments and reorganize them into task-specific multi-video structures. This segment-level compositional design enables scalable data construction while preserving precise temporal alignment and controllable cross-video relationships. On top of these compositional instances, we further annotate tool-augmented reasoning trajectories, which decompose the reasoning process into a sequence of interpretable steps, while tool-based interactions allow the model to actively retrieve and inspect video segments during reasoning. In the end, we design an automatic data filtering pipeline to enhance the annotation quality. As shown in Figure 1, this design enforces explicit grounding in visual evidence, ensuring that intermediate reasoning steps are supported by actual observations rather than hallucinated content.

With the training data, we train VIDEOINSPECTOR with a two-stage pipeline. We first perform cold-start supervised fine-tuning, where the model learns to imitate structured tool-augmented reasoning trajectories. However, supervised fine-tuning is limited by its offline nature: the model only learns to reproduce fixed trajectories and does not explicitly optimize decisions. Thus, we introduce a second stage based on agentic reinforcement learning, where the model explores alternative reasoning trajectories and optimizes its decision-making policy through reward signals, enabling more effective reasoning across videos.

The contributions of this work are summarized as follows. **First**, we construct VIDEOINSPECTOR-INSTRUCT, a large-scale compositional segment-level multi-video dataset with tool-augmented supervision. **Second**, we train an agentic framework, VIDEOINSPECTOR, that performs coarse-to-fine multi-video reasoning through explicit retrieval, segment inspection, and evidence aggregation, based on a two-stage training pipeline. **Third**, we empirically demonstrate that tool-augmented VIDEOINSPECTOR improves over its non-tool variant on several benchmarks, suggesting that compositional segment supervision and grounded tool use are promising directions for multi-video reasoning.

2 RELATED WORK

Multi-Video Understanding and Benchmarks Despite advances in single-video understanding, real-world applications often require MLLMs to aggregate, compare, and align information across multiple videos. CrossVid (Li et al., 2026), one of the first large-scale cross-video reasoning benchmarks, evaluates evidence integration under a unified query through comparative analysis, temporal reasoning, and multi-view understanding. CVBench (Zhu et al., 2025b) provides a hierarchical diagnostic framework spanning object association, event association, and complex reasoning, testing entity correspondences, temporal dependencies, and causal links across videos. In realistic video editing scenarios, VEBench (Deng et al., 2026) evaluates multi-video reasoning through an operation simulation subtask that requires selecting and temporally localizing segments from multiple candidate videos to construct coherent narratives, combining perception with decision-making. Like long-video understanding, multi-video reasoning involves substantial visual-token redundancy while requiring models to identify and align sparse, critical evidence across sources. We therefore construct training data that encourages models to focus on evidence-salient video segments through tool invocation, enabling efficient cross-video evidence grounding.

Agentic Video Reasoning with Tool Use Agentic video reasoning frames understanding as active evidence seeking. Early video agents (Wang et al., 2024b; Fan et al., 2024) use LLM planners to iteratively retrieve and aggregate long-video evidence. Deep Video Discovery (Zhang et al., 2025c) supports agent-directed search over a multi-granular video database, while VideoExplorer (Yuan et al., 2025) integrates planning, temporal grounding, and scalable perception through "thinking with video." Post-training methods such as VITAL (Zhang et al., 2025b) and LongVT (Yang et al., 2025) combine visual toolboxes, cold-start supervised fine-tuning, and reinforcement learning to learn tool-augmented reasoning. For multi-video reasoning, MAGNET (Chowdhury et al., 2025) introduces AV-HaystacksQA, requiring salient-segment retrieval and linking across videos. VideoForest (Meng et al., 2025) uses person-anchored hierarchical reasoning for cross-video QA, while SAMA (Zhang et al., 2026b) combines visual tools, task-specific skills, and conflict-aware verification. These works emphasize cross-video evidence retrieval, alignment, and comparison. However, most existing systems target single videos or operate as inference-time pipelines. We instead construct compositional agentic training data for heterogeneous multi-video tasks and learn a unified tool-using policy over a compact search-and-inspect action space, enabling segment-level evidence retrieval, inspection, comparison, and aggregation across videos.

Training Data for Grounded Multi-Step Video Reasoning Most video instruction datasets provide single-video captions or final QA pairs, supporting video-language alignment without explicitly teaching evidence retrieval, temporal localization, segment comparison, or recovery from intermediate errors. Existing tool-augmented data primarily targets long-video reasoning, supervising visual tool use, clip cropping or resampling, and structured reasoning trajectories (Zhang et al., 2025b; Yang et al., 2025; Yuan et al., 2025). Multi-video reasoning additionally requires heterogeneous supervision spanning source selection, temporal localization, chronological reasoning, and cross-video comparison. We address this gap with **VIDEOINSPECTOR-INSTRUCT**, a compositional segment-level dataset with tool-augmented reasoning trajectories. By recombining timestamped semantic segments into task-specific multi-video structures, it provides controllable supervision for evidence acquisition and composition.

3 VIDEOINSPECTOR-INSTRUCT TRAINING DATA

Multi-video reasoning requires retrieving relevant sources, locating key moments, and combining evidence across videos. We construct four training tasks to develop these complementary abilities. Following recent tool-augmented video agents (Yang et al., 2025; Wang et al., 2025), our framework encourages models to use `inspect_segment` to obtain temporally localized visual evidence from selected segments. This connects high-level planning with fine-grained perception, grounding reasoning in explicit visual observations.

3.1 TASK DESIGN AND DEFINITIONS

Cross-Video Retrieval. This task requires the model to identify which candidate video contains an event described by a textual query. A typical question is: *"Which video contains the scene where X?"*, where X is derived from a segment-level event description. This task targets the most basic capability in multi-video reasoning: selecting the correct visual source from multiple candidates.

Cross-Video Temporal Grounding. Cross-video temporal grounding extends retrieval by requiring the model to temporally localize the supporting evidence. Given an event description and multiple candidate videos, the model must determine which video contains the event and the corresponding temporal span. A typical question is: *"In which video and at what time does event X occur?"* The model needs to first select a likely video and then inspect fine-grained segments to produce temporally grounded evidence.

Clip Reordering. Given a set of temporally shuffled clips sampled from one or multiple related videos, the model is required to arrange them in the correct chronological order. This task trains the model to reason over temporal progression, state changes, event continuity, and causal dependencies across disconnected segments. Unlike standard single-video temporal modeling, the clips are presented as separate candidates without their original temporal context, forcing the model to compare segment-level evidence across sources rather than relying on the original video order. A representative query is: *"Given the following shuffled clips, arrange them in chronological order."* Solving this task requires the model to inspect each clip, identify temporal cues, and infer the underlying event progression. This capability is important for multi-video understanding because evidence from different clips often needs to be aligned along an implicit temporal or causal structure.

Video Comparison. This task requires the model to compare multiple clips or videos and reason about their similarities and differences. A representative query is: *"Which two clips correspond to the same topic X, and what distinguishes them?"* This task targets higher-level cross-video reasoning, where the model must align semantically related evidence across videos, abstract common attributes, and identify fine-grained contrasts in objects, actions, scenes, or event outcomes.

3.2 DATA ANNOTATION

Constructing training data for multi-video reasoning is substantially more challenging than standard single-video instruction tuning. The annotation must not only describe visual content within each video, but also support controllable cross-video relations, temporally grounded evidence, and multi-step tool-use trajectories. To address the challenges, we design a multi-stage annotation pipeline

consisting of metadata annotation, difficulty-controlled task construction, and tool-augmented reasoning trajectory annotation.

Dense Segment Metadata Annotation. We collect a large-scale video corpus from an online video platform spanning diverse domains, including movies, sports broadcasts, edited content, etc. We first perform metadata annotation and quality filtering using Gemini-2.5-Pro (Gemini Team, 2025). For each video, Gemini generates dense temporal captions with timestamp annotations, where each caption describes a fine-grained visual event and its temporal boundary. These dense annotations convert long videos into reusable semantic segments, which serve as the basic units for subsequent composition. Beyond dense captioning, we also use Gemini to assess the overall quality and semantic richness of each video. We prioritize videos with clear narrative structure, strong temporal continuity, diverse visual events, recurring entities, and meaningful scene transitions. We filter out videos with low visual quality, weak semantic structure, excessive static content, or insufficient event diversity. Finally, we conduct manual spot checks on sampled segments to verify dense caption quality and temporal alignment. The majority of inspected segments are judged correct or acceptable, with errors mainly arising from ambiguous scene transitions or imperfect temporal boundaries.

Table 1: Statistics of VIDEOINSPECTOR-INSTRUCT.

Training Tasks	Easy	Hard	Total
Cross-video Retrieval	12K	9K	21K
Cross-video Temporal Grounding	14K	15K	29K
Clip Reordering	9K	13K	22K
Video Comparison	12K	11K	23K
Total	44K	51K	95K

Difficulty-Controlled Task Construction. Based on the segment-level metadata, we construct task-specific multi-video instances with controlled difficulty to facilitate the model to progressively learn the multi-video patterns. For cross-video retrieval, easy instances use target and distractor segments from different source videos, while hard instances sample visually or semantically similar segments from the same video or closely related videos, requiring fine-grained discrimination. For cross-video temporal grounding, hard instances require the target segment to occupy less than 25% of the video duration, making the supporting evidence sparse. For video comparison, easy instances explicitly provide the shared topic, while hard instances require the model to infer the implicit common topic before comparing fine-grained differences. For clip reordering, easy instances contain two or three clips, whereas hard instances contain four to eight clips, increasing the need for temporal cue extraction and pairwise evidence comparison. In this way, VIDEOINSPECTOR-INSTRUCT systematically controls distractor complexity, temporal sparsity, semantic similarity, and reasoning depth, exposing the model to progressively more challenging multi-video reasoning scenarios.

Tool-Augmented Reasoning Trajectory Annotation. After constructing task-specific instances, we annotate each instance with tool-augmented reasoning trajectories using GPT-4o (OpenAI, 2024). Given the dense segment descriptions and task definitions, GPT-4o produces structured trajectories that decompose the problem into intermediate steps. Each trajectory is aligned with our `inspect_segment` interface, so that the model learns not only what answer to produce, but also when to acquire visual evidence and how to use the inspected observations for reasoning. Importantly, these trajectories are not used as unconstrained free-form rationales. They are grounded in the segment-level metadata, and tool observations are tied to specific video segments and timestamps. This design reduces the risk of hallucinated reasoning and encourages the model to base intermediate decisions on explicit visual evidence. It also makes the supervision compatible with our agentic training framework, where the model must decide which video or segment to inspect before producing the final answer.

Controlled Error-Recovery Trajectories. In addition to oracle trajectories, we annotate controlled error-recovery trajectories to simulate realistic agentic reasoning. During multi-step reasoning, an agent may inspect an irrelevant candidate, gather insufficient evidence, or make an incorrect intermediate assumption. A robust model should be able to recognize such failures, revise its plan, and continue evidence acquisition. To encourage this behavior, we evenly split the data into three trajectory types: trajectories without intermediate mistakes, trajectories with one recoverable mistake followed by a rethink step, and trajectories with two recoverable mistakes followed by corresponding rethink steps. The recoverable mistakes are controlled rather than arbitrary: they correspond to plausible errors such as selecting a semantically similar distractor, inspecting an insufficient segment,

or making a premature comparison. After each mistake, the trajectory includes a rethink step that identifies why the current evidence is insufficient or inconsistent, followed by a corrected tool interaction. This design teaches the model to verify its evidence before answering, rather than blindly following the first retrieved segment. As a result, **VIDEOINSPECTOR-INSTRUCT** provides supervision for both successful tool use and self-correction under imperfect evidence acquisition.

Dataset Quality Control. We initially obtain approximately 140K samples from the aforementioned annotation process. To improve the quality of the training trajectories, we apply a three-stage automatic filtering pipeline using Qwen3-VL-30B as the judge model. In the raw annotation, we mark intentionally erroneous trajectories with auxiliary error tags. *These tagged parts are excluded from the correctness checks* in the second and third stages. First, we screen all intermediate reasoning for answer leakage and directly remove any descriptions that reveal the target answer in a way that bypasses the intended visual reasoning process. Second, we assess internal reasoning consistency, removing samples with contradictions in the untagged reasoning or a final answer that conflicts with the corrected reasoning. Explicit revisions of tagged mistakes during recovery are not treated as inconsistencies. Third, we assess evidence support, removing samples whose untagged reasoning contains hallucinated visual content or whose final answer lacks sufficient supporting evidence. Thus, intentionally introduced mistakes are preserved, while the reasoning and evidence used to recover from them must still support the final answer. After filtering, we retain approximately 95K training samples to construct VIDEOINSPECTOR-INSTRUCT. We remove all auxiliary error tags before training while preserving the underlying error-recovery trajectories.

4 TRAINING PARADIGM

4.1 COLD-START SUPERVISED FINE-TUNING

The cold-start stage aims to initialize the model with both fundamental multi-video reasoning ability and basic tool-use capability through supervised learning. This stage is entirely conducted offline and consists of two complementary components.

Without-Tool Supervision. We first train the model on multi-video data without tool usage, where each sample consists of multi-video inputs paired with direct responses. This step ensures that the model develops strong end-to-end perception and reasoning capabilities without relying on external tools, providing a stable foundation for subsequent agentic behaviors.

With-Tool Trajectory Supervision. Building upon the tool-free initialization, we conduct tool-augmented multi-turn SFT in a curriculum learning (Bengio et al., 2009) manner based on the difficulty-controlled training tasks. The supervision is performed via multi-turn instruction tuning, enabling the model to imitate structured tool-use behaviors before the upcoming RL. Despite incorporating tool interactions, this stage remains purely offline, where the model learns from fixed trajectories without exploration. As a result, the learned policy is limited to imitation and may not generalize well to unseen or uncertain scenarios.

4.2 AGENTIC REINFORCEMENT LEARNING

While cold-start SFT provides a strong initialization for structured reasoning and tool usage, it is inherently limited by fixed offline trajectories. To enable more adaptive decision-making, we further leverage an agentic RL stage based on GRPO (Shao et al., 2024), where the model learns to explore and optimize different tool-use strategies through interaction.

We construct the RL training set by filtering from the tool-augmented SFT data. Following prior work (Yang et al., 2025), we perform multiple rollouts for each hard instance using the current policy. Based on the correctness of the generated trajectories, samples are categorized into three groups: (i) always correct (too easy), (ii) always incorrect (too hard or noisy), and (iii) partially correct (uncertain). We retain only the uncertain samples for RL training, as they provide the most informative learning signals. Consequently, we obtain about 8k instances for stage 2 training.

Due to the multi-task nature of our training data, we design task-specific rewards that jointly evaluate answer correctness, reasoning quality, format compliance, and temporal grounding. For tasks with discrete answers, such as cross-video retrieval and clip reordering, we use a rule-based accuracy

reward $\mathbf{R}_{\text{acc}}$ that assigns positive rewards only to correct predictions. For open-ended tasks such as video comparison, we employ an LLM-as-a-Judge (Gu et al., 2024) reward $\mathbf{R}_{\text{judge}}$ to evaluate semantic consistency between the generated answer and the ground truth, assigning rewards for correct, partially correct, and incorrect responses. To encourage valid agentic reasoning trajectories, we include a rule-based format reward $\mathbf{R}_{\text{fmt}}$ to enforce structured rollout outputs. For temporal grounding tasks, we use a temporal IoU reward $\mathbf{R}_{\text{time}} = \frac{|\hat{\tau} \cap \tau^*|}{|\hat{\tau} \cup \tau^*|}$ between the predicted and ground-truth temporal spans, encouraging precise temporal localization rather than coarse video-level matching.

5 Experiments

5.1 Experimental Settings

Implementation Details. We use LMMs-Engine (LMMs-Lab, 2025) for cold-start SFT, and verl (Sheng et al., 2024) with vLLM (Kwon et al., 2023) for agentic RL. The learning rate is 10^{-5} for SFT and 10^{-6} for RL. The number of rollouts for GRPO is 8. We train the model with tool-free and tool-augmented data for one epoch each during the cold-start SFT stage. All experiments are conducted on 8 NVIDIA A100 GPUs.

Evaluation Settings. We evaluate VideoInspector on two general multi-video benchmarks, CrossVid (Li et al., 2026) and CVBench (Zhu et al., 2025b), as well as VEBench (Deng et al., 2026), a task-specific benchmark for video editing. In addition to the Qwen2.5-VL-7B and Qwen3-VL-8B backbone baselines, we compare VideoInspector against proprietary MLLMs, including Gemini-2.5-Pro (Gemini Team, 2025) and GPT-4o (OpenAI, 2024), as well as the open-source model VideoLLaMA3-7B (Zhang et al., 2025a). We also include specialized video models, namely LongVILA-R1-7B (Chen et al., 2025b), TimeLens-7B (Zhang et al., 2026a), and Vgent-7B (Shen et al., 2025). For agentic baselines, we include the training-free LVAgent (Chen et al., 2025a), the tool-augmented VideoThinker (Wang et al., 2025), LongVT-RFT (Yang et al., 2025), LongVideoAgent (Liu et al., 2026), and TimeSearch-R (Pan et al., 2025). Following the original LVAgent setup, we instantiate its agent pool with Qwen2-VL-7B (Wang et al., 2024a), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL-8B (Zhu et al., 2025a). We report accuracy on CrossVid, CVBench, and the footage selection subtask of OpSim (OpSim-FS), while the full OpSim task is evaluated using conditioned temporal IoU. For models evaluated on all three benchmarks, the Overall score is computed as the arithmetic mean of CrossVid Avg., CVBench Avg., and VEBench OpSim-FS.

5.2 Main Results

As shown in Table 2, our method achieves the strongest overall performance among the evaluated open-source models across the three benchmarks. Using Qwen2.5-VL-7B as the backbone, our model achieves an overall score of 38.4 without tool use, which is further improved to 40.3 after enabling tool-augmented training. The latter outperforms all evaluated specialized video models and agentic frameworks in overall performance. Compared with the Qwen2.5-VL-7B baseline, our full model obtains absolute gains of +10.4 on CrossVid average performance, +5.1 on CVBench average performance, and +3.8 on VEBench OpSim-FS, demonstrating the effectiveness of our compositional training paradigm for multi-video reasoning. On CrossVid, our approach improves comparative analysis (CA) and temporal understanding (TU) from 19.3 and 20.0 to 28.3 and 28.4, respectively. These improvements suggest that the proposed compositional segment training effectively enhances the model's ability to retrieve, localize, and reason over temporally distributed evidence across multiple videos. Our method also substantially improves multi-view reasoning (MR), increasing the score from 16.8 to 30.6, although it still trails the strongest agentic baseline on this subtask. On CVBench, our method achieves competitive performance against existing agentic frameworks. Meanwhile, on VEBench, our method also improves OpSim performance from 0.002 to 0.028, suggesting that our training and tool-augmented reasoning pipeline benefits editing-related reasoning grounded in visual evidence. The improvements also extend to the stronger Qwen3-VL-8B backbone. Our tool-free and tool-augmented variants achieve overall scores of 47.1 and 50.1, respectively, compared with 43.7 for the original backbone. The full model obtains absolute gains of +9.1 on CrossVid average performance, +6.3 on CVBench average performance, and +3.8 on VEBench OpSim-FS. These results demonstrate the effectiveness of our compositional training paradigm across both evaluated backbones.

Table 2: Performance comparison on CrossVid, CVBench, and VEBench. The purple bars indicate results from Qwen3-VL-8B backbones. **Bold** numbers indicate the best results.

Methods	CrossVid				CVBench				VEBench		Overall
	CA	TU	MR	Avg	OA	EA	CR	Avg.	OpSim-FS	OpSim	
Proprietary MLLMs											
Gemini-2.5-Pro	54.7	56.0	28.7	46.5	–	–	–	–	44.4	0.109	45.5
GPT-4o	43.1	35.5	27.4	35.3	66.9	70.8	69.1	68.9	41.9	0.004	48.7
Open-source MLLMs											
VideoLLaMA3-7B (Zhang et al., 2025a)	20.8	6.7	19.8	15.8	53.6	60.3	57.6	57.2	29.5	0.002	34.2
Qwen2.5-VL-7B (Bai et al., 2025)	19.3	20.0	16.8	18.7	53.2	54.0	51.3	52.8	30.1	0.002	33.9
Qwen3-VL-8B (Qwen Team, 2025)	42.0	28.9	31.7	34.2	61.3	57.6	63.6	60.8	36.1	0.003	43.7
Specialized Video Models											
LongVILA-R1-7B (Chen et al., 2025b)	26.2	20.1	27.4	24.6	54.4	58.4	53.4	55.5	29.1	0.001	36.4
TimeLens-7B (Zhang et al., 2026a)	**36.0**	22.6	20.6	**29.5**	56.6	**66.1**	58.6	**60.9**	26.1	0.017	38.8
Vgent-7B (Shen et al., 2025)	32.5	26.3	24.8	27.9	**56.9**	65.7	56.0	59.8	26.8	0.025	38.2
Agentic Models											
LVAgent (Chen et al., 2025a)	17.6	26.9	35.6	26.7	49.4	50.7	53.1	51.1	28.1	0.002	35.3
VideoThinker (Wang et al., 2025)	26.5	25.1	33.5	28.4	46.0	54.3	50.0	50.6	29.8	0.009	34.9
LongVT-RFT (Yang et al., 2025)	23.0	17.6	34.4	25.0	36.3	40.5	44.2	40.9	27.7	0.023	31.2
LongVideoAgent (Liu et al., 2026)	18.9	11.8	30.4	20.4	44.8	40.5	44.2	43.0	23.4	0.021	28.9
TimeSearch-R (Pan et al., 2025)	30.7	23.9	**36.8**	30.5	56.0	63.0	56.3	58.7	27.1	0.025	38.8
Qwen2.5-VL-7B + Ours (w/o tool)	26.7	27.5	25.9	26.7	55.1	56.3	57.0	56.1	32.4	0.001	38.4
Qwen2.5-VL-7B + Ours	28.3	**28.4**	30.6	29.1	56.8	57.5	**59.3**	57.9	**33.9**	**0.028**	**40.3**
Qwen3-VL-8B + Ours (w/o tool)	47.3	38.2	40.2	41.9	62.4	60.6	64.8	62.6	36.9	0.009	47.1
Qwen3-VL-8B + Ours	**48.4**	**40.1**	**41.4**	**43.3**	**63.8**	**66.5**	**71.0**	**67.1**	**39.9**	**0.031**	**50.1**

5.3 Ablation Studies

Training Data Ablation. To better understand the effectiveness and limitations of our compositional training data, we conduct comprehensive training data ablations on the CrossVid benchmark under both tool-free and tool-augmented settings. As shown in Table 3, the four proposed training tasks exhibit clear and interpretable functional specialization toward different evaluation abilities. Specifically, removing cross-video retrieval (CVR) consistently reduces comparative analysis performance and overall accuracy, suggesting that retrieval-oriented supervision is important for identifying semantically relevant candidates across videos. Removing clip reordering (CR) leads to the largest degradation in temporal understanding in the tool-free setting, particularly on procedural step sequencing (PSS), which directly corresponds to clip reordering, suggesting that alignment between the training and evaluation tasks benefits temporal reasoning. Similarly, removing cross-video temporal grounding (CVTG) substantially hurts temporal localization performance, showing that timestamp-aware supervision is crucial for fine-grained temporal perception. Finally, removing video comparison (VC) consistently degrades comparative analysis and overall reasoning performance, demonstrating the importance of explicit cross-video evidence aggregation. Importantly, removing any of the four training tasks reduces overall performance and all three category-level averages in both the tool-free and tool-augmented settings. Although tool augmentation generally improves comparative analysis and temporal understanding, the magnitude of the performance drops and the most affected subtasks vary across settings. These results suggest that the proposed training tasks remain beneficial under both settings, while their relative contributions depend on the evaluation dimension and tool-use setting. Overall, these results validate that our compositional task design provides complementary supervision signals for different aspects of multi-video reasoning.

Training Stage Ablation. Table 4 examines the contributions of SFT, RL, and their combination under both tool-free and tool-augmented settings. SFT+RL achieves the highest CrossVid and CVBench average scores and the best OpSim-FS performance in both settings, outperforming either training stage alone on these metrics. These results support the combination of cold-start SFT and subsequent RL in our training pipeline. The additional benefit of RL after SFT depends on the training setting. On CrossVid, adding RL improves average performance by 3.3 points without tools, compared with 2.1 points with tools. Similarly, the OpSim-FS gain decreases from 2.1 points in the tool-free setting to 1.1 points in the tool-augmented setting. Thus, although RL improves both

Table 3: Training data ablation evaluated on CrossVid. **Bold** numbers indicate the best results.

Methods	O.Avg	Comparative Analysis					Temporal Understanding				Multi-view Reasoning		
		BU	NC	CC	PEA	C.Avg	PI	FSA	PSS	T.Avg	MSR	MOC	M.Avg
Qwen2.5-VL-7B	18.7	19.6	19.0	23.4	15.0	19.3	58.6	1.2	0.3	20.0	11.8	21.7	16.8
Ours (w/o tool)	**26.7**	**26.9**	**33.3**	**27.5**	**19.2**	**26.7**	**69.3**	**4.1**	**9.1**	**27.5**	**23.2**	**28.6**	**25.9**
w/o CVR	24.2	25.1	31.7	25.9	18.1	25.2	65.4	3.9	8.7	26.0	18.4	24.2	21.3
w/o CR	23.4	24.4	31.0	24.7	17.8	24.5	64.7	2.6	4.6	24.0	17.6	25.8	21.7
w/o CVTG	23.4	24.2	32.2	25.1	15.8	24.4	65.2	3.4	5.8	24.8	17.3	24.9	21.1
w/o VC	22.8	24.2	32.5	26.2	16.3	24.8	65.0	2.8	6.3	24.7	14.0	23.8	18.9
Ours (w/ tool)	**29.1**	**27.3**	**34.6**	**29.8**	**21.5**	**28.3**	**70.1**	**4.1**	**11.0**	**28.4**	**28.4**	**32.7**	**30.6**
w/o CVR	27.1	26.4	32.8	27.3	18.9	26.4	67.9	3.3	9.7	27.0	24.3	31.2	27.8
w/o CR	27.3	26.9	31.7	29.3	19.9	27.0	69.2	3.8	8.2	27.1	24.6	30.9	27.8
w/o CVTG	26.2	26.2	30.9	28.8	18.5	26.1	66.3	2.7	7.8	25.6	23.9	29.9	26.9
w/o VC	25.5	25.9	31.2	27.9	17.2	25.6	67.1	3.9	6.6	25.9	19.9	30.2	25.1

Table 4: Training-stage ablation evaluated on CrossVid, CVBench, and VEBench.

Training Strategy	CrossVid				CVBench				VEBench	
	CA	TU	MR	Avg	OA	EA	CR	Avg.	OpSim-FS	OpSim
Without Tool Use										
only SFT	23.1	25.9	21.3	23.4	55.0	56.0	55.1	55.4	30.3	0.001
SFT+RL	**26.7**	**27.5**	**25.9**	**26.7**	**55.1**	**56.3**	**57.0**	**56.1**	**32.4**	0.001
only RL	23.4	24.8	19.3	22.5	54.6	56.1	55.5	55.4	29.9	**0.003**
With Tool Use										
only SFT	**28.3**	27.2	26.8	27.4	56.2	56.9	57.8	57.0	32.8	0.018
SFT+RL	**28.3**	**28.4**	**30.6**	**29.1**	**56.8**	**57.5**	**59.3**	**57.9**	**33.9**	**0.028**
only RL	26.4	24.5	25.7	25.5	55.3	56.3	55.9	55.8	31.7	0.009

variants, its marginal contribution is smaller for the tool-augmented model on these two metrics. This pattern may partly reflect the stronger performance already established during tool-augmented SFT. With SFT alone, the tool-augmented model reaches 27.0 on CrossVid Avg., 57.0 on CVBench Avg., and 32.8 on OpSim-FS, exceeding even the tool-free SFT+RL model on all three metrics. The tool-augmented variant therefore provides a stronger starting point for subsequent optimization, while RL further improves its performance and yields the best final results on these benchmark summaries.

6 Conclusion

In this work, we introduce VideoInspector-Instruct, a large-scale compositional segment training dataset for multi-video understanding, featuring temporally grounded tool use and structured multi-step reasoning trajectories. Unlike single-video instruction datasets, VideoInspector-Instruct explicitly targets the core capabilities required for multi-video reasoning. By decomposing videos into reusable semantic segments and recomposing them into diverse tasks, it provides scalable and controllable supervision for grounded evidence acquisition and aggregation. Built upon this dataset, we propose VideoInspector, an agentic multi-video reasoning framework trained through cold-start supervised fine-tuning and agentic reinforcement learning. The model progressively retrieves, inspects, and combines visual evidence across multiple videos in a coarse-to-fine manner. Experiments on three benchmarks demonstrate strong performance across diverse multi-video reasoning scenarios. These results highlight the importance of evidence-grounded agentic reasoning for multi-video understanding. We hope our dataset and training framework provide a useful foundation for future research in this direction.

AI USE STATEMENT

We used generative AI models, including Gemini 2.5 Pro and Qwen3-VL-30B, for the annotation procedures described in this paper. We also used ChatGPT for language polishing and multiple rounds of manuscript feedback to improve clarity, presentation, and overall quality. The authors take full responsibility for the accuracy, originality, and integrity of all content in this paper, including AI-assisted annotations and text, experimental results, analyses, and conclusions.

REFERENCES

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

James Burgess, Xiaohan Wang, Yuhui Zhang, Anita Rau, Alejandro Lozano, Lisa Dunlap, Trevor Darrell, and Serena Yeung-Levy. Video action differencing. *arXiv preprint arXiv:2503.07860*, 2025.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20237–20246, 2025a.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos, 2025b. URL https://arxiv.org/abs/2507.07966.

Sanjoy Chowdhury, Mohamed Elmoghany, Yohan Abeysinghe, Junjie Fei, Sayan Nag, Salman Khan, Mohamed Elhoseiny, and Dinesh Manocha. Magnet: A multi-agent framework for finding audio-visual needles by reasoning over multi-video haystacks. *arXiv preprint arXiv:2506.07016*, 2025.

Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *arXiv preprint arXiv:2510.08559*, 2025.

Andong Deng, Dawei Du, Zhenfang Chen, Wen Zhong, Fan Chen, Guang Chen, Chia-Wen Kuo, Longyin Wen, Chen Chen, and Sijie Zhu. Vebench: Benchmarking large multimodal models for real-world video editing. *arXiv preprint arXiv:2605.03276*, 2026.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

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

Gemini Team. Gemini models: Gemini 2.5 pro, 2025. URL https://developers.googleblog.com/en/gemini-2-5-video-understanding/. Accessed May 15, 2025.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Bozheng Li, Yongliang Wu, Yi Lu, Jiashuo Yu, Licheng Tang, Jiawang Cao, Wenqing Zhu, Yuyang Sun, Jay Wu, and Wenbo Zhu. Veu-bench: Towards comprehensive understanding of video editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13671–13680, 2025.

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.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. Longvideoagent: Multi-agent reasoning with long videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40404–40416, 2026.

LMMs-Lab. Lmms engine: A simple, unified multimodal framework for pretraining and finetuning., 2025. URL https://github.com/LMMs-Lab/lmms-engine.

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.

OpenAI. Gpt-4o: Openai's newest multimodal model. https://openai.com/index/gpt-4o, 2024. Accessed: 2025-05-09.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. *arXiv preprint arXiv:2511.05489*, 2025.

Qwen Team. Qwen3-vl. https://github.com/QwenLM/Qwen3-VL, 2025.

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

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding, 2025. URL https://arxiv.org/abs/2510.14032.

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

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025.

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*, pp. 58–76. Springer, 2024b.

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.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*, 2025.

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.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22052–22061, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL https://arxiv.org/abs/2501.13106.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025b.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms, 2026a. URL https://arxiv.org/abs/2512.14698.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025c.

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*, 2026b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025a. URL https://arxiv.org/abs/2504.10479.

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*, 2025b.