ANCHORGROUP: OBJECT-ANCHORED VIDEO UNDERSTANDING BENCHMARK WITH AN AGENTIC ANNOTATION LOOP

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Question Answering (VideoQA) has seen advances in both Multimodal Large Language Models (MLLMs) and evaluation benchmarks. Existing benchmarks typically evaluate independent QA pairs and report performance for each task. Questions from different tasks may concern different objects, even within the same video. Separate task scores therefore provide limited insight into whether models succeed across tasks involving the same object. To address this gap, we introduce **AnchorGroup**, a video understanding benchmark organized into object-anchored QA groups. Each QA group contains perception, reasoning, and summarization questions about the same object. The shared anchor provides a common reference for comparing task performance and examining associations between task outcomes. Existing annotation pipelines primarily focus on individual QA pairs rather than shared-anchor constraints across tasks. We therefore introduce **A^{3}Loop**, an anchor-guided agentic annotation loop that constructs QA groups with direct access to video clips. The loop generates QA pairs from summarization to reasoning and then perception. During construction, verifier feedback guides generators in revising QA pairs to meet task-specific and group-level requirements. Revised QA pairs are then verified again. Evaluations on AnchorGroup show that perception has the lowest average accuracy among the three tasks. Correct perception answers are generally associated with higher reasoning and summarization accuracy within the same QA groups. We release the dataset and annotation framework to support reproducible evaluation of video understanding across tasks within shared object contexts.

1 INTRODUCTION

Video Question Answering (VideoQA) aims to answer questions grounded in video content. Recent Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Feng et al., 2026b; Lin et al., 2024; Sun et al., 2026b; Zhang et al., 2024; 2026) have achieved strong performance on diverse video understanding benchmarks (Zheng et al., 2024; Fu et al., 2025a; Wang et al., 2025a; Wu et al., 2024; Chen et al., 2025a; Cheng et al., 2025). Alongside model development, recent benchmark research has explored advances in evaluation design and annotation methods. Multi-level evaluation supports finer-grained capability analysis (Fu et al., 2026; Meng et al., 2026) and annotation pipelines aim to improve scalability and sample quality (Alam et al., 2026; Wu et al., 2026).

However, existing benchmarks commonly use individual QA pairs as the basic unit of evaluation (Hu et al., 2026a; Zhou et al., 2025; Ma et al., 2025). Questions from different task categories do not necessarily share an object context (Fu et al., 2026; Meng et al., 2026; Shaar et al., 2026). As illustrated in Fig. 1(a) and Fig. 1(b), questions targeting different capabilities may refer to different objects, even within the same video. Such QA organization supports task-specific evaluation. Yet, differences in task performance also reflect differences in the objects being evaluated. Object differences make associations between task outcomes harder to interpret. Meanwhile, existing VideoQA annotation pipelines primarily focus on generating and verifying individual QA pairs. Extending these pipelines to QA groups requires joint checks of shared object references and evidence relations, placing additional demands on the annotation process.

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.

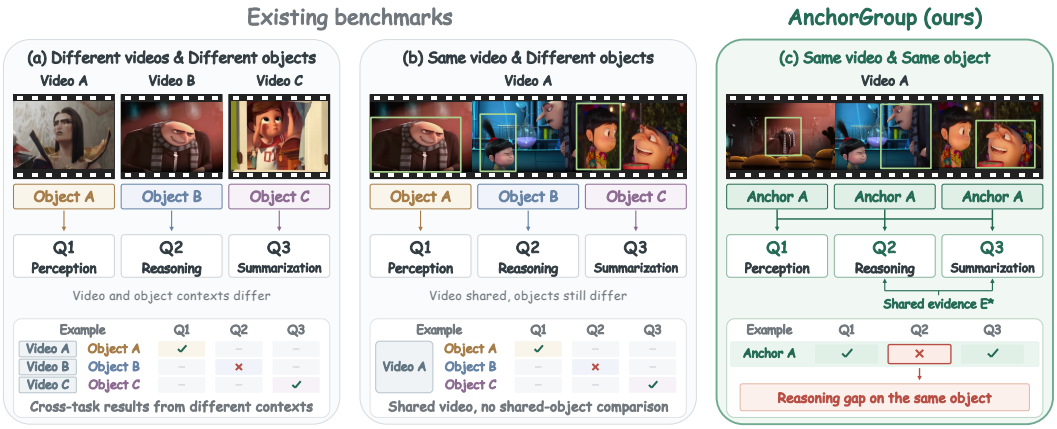

Figure 1: **The difference between AnchorGroup and other benchmarks.** AnchorGroup provides QA groups with different tasks within the same anchor object.

To examine model performance across tasks within a shared object context, we introduce **AnchorGroup**, a video understanding benchmark organized into object-anchored QA groups. As illustrated in Fig. 1(c), each group contains perception, reasoning, and summarization questions about the same anchor object in the same video. The reasoning and summarization questions also share an evidence segment that connects a specific inference with the content being summarized. Grouping these questions allows us to examine associations between task outcomes for the same object. For example, we compare reasoning accuracy between QA groups with correct and incorrect perception answers. The comparison shows whether reasoning success is associated with perception correctness for the same object. Separate task averages do not reveal this relationship. Constructing these groups requires checks of both individual QA quality and group-level object and evidence constraints. To meet these requirements, we introduce **A^{3}Loop**, an anchor-guided agentic annotation loop that coordinates generators and verifiers through iterative feedback. Generators produce QA pairs and their supporting evidence and can request additional video clips when needed. Verifiers check factual correctness, answerability, and compliance with task-specific and group-level requirements. Video evidence supports both generation and verification. When a check fails, the loop automatically returns the annotation to the relevant generation stage with targeted feedback. Generators revise the QA pairs or reselect candidates, and verifiers check the updated annotations again. The feedback loop supports joint construction and verification of QA groups at scale.

We evaluate 35 video MLLMs across multiple model families and sizes. Gemini 3.7 Flash achieves the highest average accuracy of 78.72%. Gemma 4 31B leads the evaluated open-source models with an average accuracy of 61.39%. However, only one model answers all three questions correctly in more than half of the QA groups. Perception has the lowest average accuracy among the three tasks. Reasoning and summarization accuracy is generally higher in groups with correct perception answers. Averaged across models, the conditional accuracy gaps are 5.44 percentage points for reasoning and 4.61 points for summarization. These results reveal associations between task outcomes within a shared object context, although their strength varies across models.

In summary, **AnchorGroup** organizes questions at different task levels into object-anchored QA groups with **A^{3}Loop** to construct and verify these groups. Our main contributions are as follows:

1. We introduce **AnchorGroup**, a video understanding benchmark covering 137.9 hours of video and containing 1,875 QA pairs in 625 QA groups. Each group includes perception, reasoning, and summarization questions about the same object. The reasoning and summarization questions also share an evidence segment. The groups provide a common object context for comparing performance across tasks.

2. We propose **A^{3}Loop**, an anchor-guided agentic annotation loop with video-grounded generation and verification. It constructs QA groups from summarization to reasoning and then perception. Failed annotations automatically return to the relevant stage for revision or reselection, followed by re-verification against task-specific and group-level constraints.

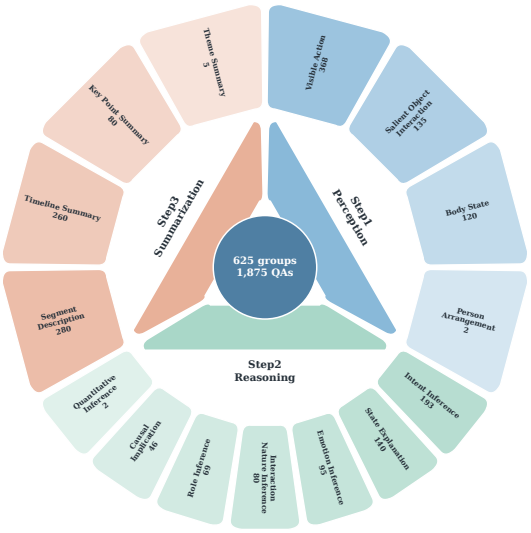

Figure 2: Task category distribution.

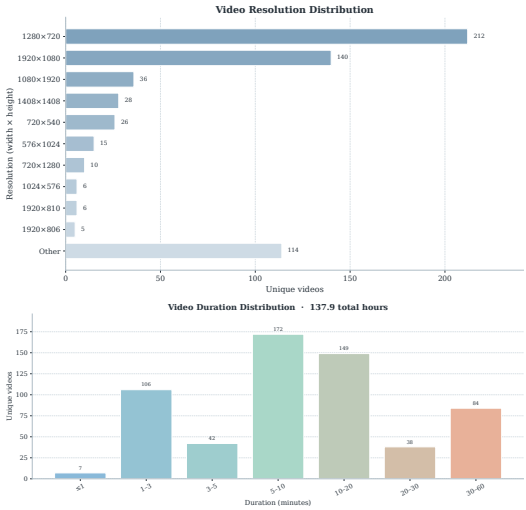

Figure 3: Video length and resolution statistics.

3. We evaluate representative video MLLMs on **AnchorGroup** and analyze task-level accuracy, joint success within QA groups, and cross-task associations. Average accuracy and the all-three-correct rate can rank models differently. Group-level evaluation therefore provides information beyond separate task averages. Correct perception answers are generally associated with higher reasoning and summarization accuracy.

2 Related Work

Video Understanding Benchmarks. Video understanding benchmarks assess MLLMs across diverse capabilities (Li et al., 2024; Cao et al., 2026; Hong et al., 2025; Liu et al., 2025) and application scenarios (Hu et al., 2026a; Zhao et al., 2025; Fu et al., 2025a). Other benchmarks focus on video-specific settings, including long-video understanding (Wang et al., 2025a; Wu et al., 2024; Chen et al., 2025a; Ma et al., 2025; Zhou et al., 2025) and streaming video understanding (Zeng et al., 2025; Fu et al., 2025b; Chen et al., 2025b). Recent studies provide finer-grained assessment through task taxonomies (Meng et al., 2026; Swetha et al., 2026) or multi-level evaluation protocols (Fu et al., 2026; Shaar et al., 2026). However, questions across task categories do not necessarily concern the same object or share supporting evidence. In contrast, **AnchorGroup** organizes questions from different task levels into QA groups around a shared object anchor. The reasoning and summarization questions also share one evidence segment. The shared anchor provides a common reference for comparing performance across tasks and examining associations between their outcomes.

Agentic Annotation Pipelines. Video benchmark construction has traditionally relied on human annotation to address the complexity of video content (Fu et al., 2025a; 2026; Cao et al., 2026). To assist annotation, recent benchmarks use MLLM-generated captions as references for human QA construction and verification (Meng et al., 2026; Yu et al., 2025). Agent-assisted annotation and data synthesis methods incorporate feedback at different levels. Batch-level feedback uses accumulated errors to guide subsequent generations (Wu et al., 2026). Sample-level feedback supports iterative verification and refinement of individual annotations (Alam et al., 2026; Hu et al., 2026b). Pipeline-level feedback optimizes data generation recipes (Kulikov et al., 2026). Constructing object-anchored QA groups additionally requires joint checks of shared object references and specified evidence relations. Our **A^{3}Loop** combines video-grounded generation and verification in an agentic feedback loop for QA-group construction. Failed annotations automatically return to the relevant generation stage for revision or candidate reselection and are then verified again.

Table 1: **Comparison of video understanding benchmarks. #Videos** and **#QA Pairs** denote the numbers of videos and QA pairs, respectively. **A** and **M** denote automatic and manual annotation. **QA Type** specifies the answer format. **Task Level** reports the number of predefined task levels. **Group** indicates whether QA pairs are grouped by video. **Anchor** indicates whether QA pairs share the same object instance.

Benchmark	#Videos	#QA Pairs	Anno. Type	QA Type	#Task Level	Group	Anchor
Video-MME (Fu et al., 2025a)	900	2,700	M	MCQ	3	✔	✘
Video-MME-v2 (Fu et al., 2026)	800	3,200	M	MCQ	3	✔	✘
CG-Bench (Chen et al., 2025a)	1,219	12,129	M	MCQ&Open	3	✔	✘
MLVU (Zhou et al., 2025)	1,730	3,102	A&M	MCQ	3	✘	✘
LVBench (Wang et al., 2025a)	103	1,549	M	MCQ	6	✔	✘
Video-MMMU (Hu et al., 2026a)	300	900	M	MCQ	3	✔	✘
MMVU (Zhao et al., 2025)	1,529	3,000	M	MCQ&Open	6	✔	✘
MVBench (Li et al., 2024)	3,641	4,000	A&M	MCQ	9	✘	✘
AnchorGroup (Ours)	598	1875	A&M	MCQ	3	✔	✔

3 AnchorGroup

3.1 Overview

AnchorGroup is a video understanding benchmark organized into object-anchored QA groups. Each video is associated with at least one QA group. Each group contains three QA pairs about the same object: one each for perception, reasoning, and summarization. The reasoning and summarization questions also share an evidence segment. All questions are presented in a four-option multiple-choice format. The dataset contains 598 videos totaling 137.9 hours and 625 QA groups comprising 1,875 QA pairs. The videos are up to 60 minutes long and have a resolution of at least 480p. Fig. 3 shows the distributions of video duration and resolution. Fig. 2 presents the distribution of task types within each task level. Table 1 compares AnchorGroup with other video understanding benchmarks.

3.2 Design Principles and Structure

To compare model performance across tasks for the same object, AnchorGroup organizes questions into object-anchored QA groups. Each group focuses on an identifiable object instance in a video. The anchor maintains a consistent identity across the relevant video segments. Each group contains three QA pairs covering perception, reasoning, and summarization.

Perception asks about directly observable actions, interactions, physical states, or spatial arrangements involving the anchor. The anchor is the main subject of each question. Reasoning requires the model to infer an answer from visual evidence about the anchor, rather than directly report an observed fact. Summarization asks for an integrated account of multiple events involving the anchor. Our task categories are consistent with those commonly used in video understanding benchmarks (Fu et al., 2025a; Wu et al., 2024). Detailed task types and definitions are provided in Appendix A.2.

All three questions share the same anchor, while reasoning and summarization additionally share an evidence segment. The shared segment connects a specific inference to part of the content covered by the summarization task. Object-anchored grouping allows us to examine whether a model succeeds across all three tasks for the same object. It also supports analysis of associations between task outcomes that are not captured by separate task averages.

3.3 Evaluation

We evaluate each QA pair independently. In the video-input setting, each query contains frames sampled from the full video, one question, and its four answer options. The model selects one option. An answer is correct if the selected option matches the reference answer. We summarize the video-input predictions using two complementary reporting schemes.

In **task-level reporting**, we report accuracy separately for perception, reasoning, and summarization. Overall accuracy is the mean of the three task accuracies. We additionally report **All3**, the percentage of QA groups with all three answers correct under video input. All3 measures joint suc-

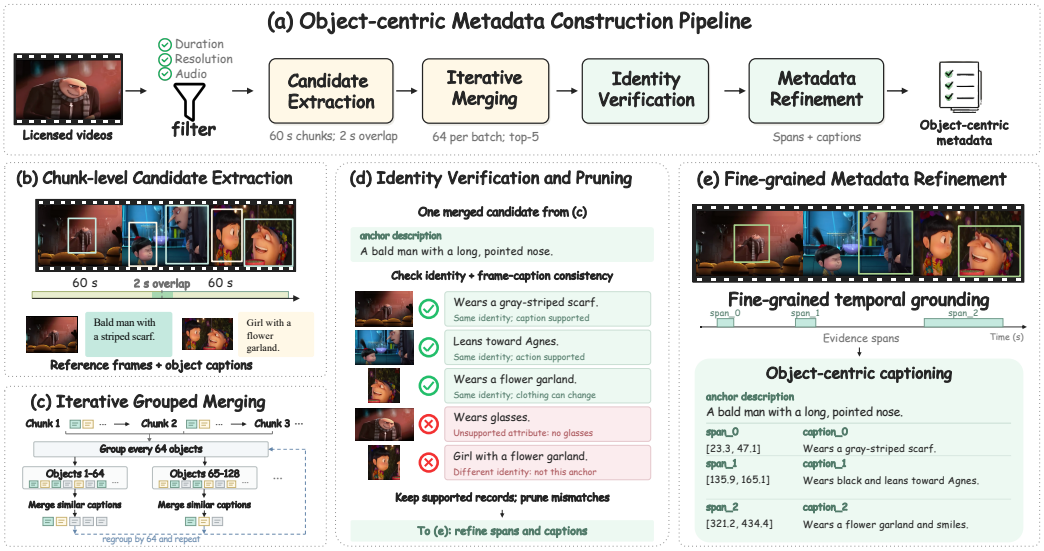

Figure 4: **Object-centric metadata construction.** (a) Overview of video filtering and metadata construction. (b) Object captions and reference frames are extracted from overlapping video chunks. (c) Candidate objects are iteratively merged in batches. (d) Object identity and caption support are verified, and inconsistent annotations are pruned. (e) Fine-grained temporal grounding and object-centric captioning refine the video spans and captions for each candidate anchor.

cess within each group. For comparison, we report overall accuracy for a text-only baseline using the same questions and options without video input.

In **conditional reporting**, we reuse the same video-input predictions to examine associations between task outcomes within QA groups. Let $P, R, S \in \{0, 1\}$ denote the correctness of perception, reasoning, and summarization answers, respectively. A value of 1 indicates a correct answer, and 0 indicates an incorrect answer. For each model, we partition QA groups according to perception correctness. We then compute reasoning and summarization accuracy within the perception-correct and perception-incorrect subsets. We define the perception-conditioned accuracy gaps as:

$$\Delta_{P \to R} = \Pr(R = 1 \mid P = 1) - \Pr(R = 1 \mid P = 0), \tag{1}$$

$$\Delta_{P \to S} = \Pr(S = 1 \mid P = 1) - \Pr(S = 1 \mid P = 0). \tag{2}$$

For example, $\Pr(R = 1 \mid P = 0)$ is reasoning accuracy in perception-incorrect groups. Gaps are reported in percentage points. Positive gaps indicate higher accuracy in perception-correct groups. These gaps capture cross-task associations beyond separate task accuracies.

4 A^3 Loop

A^{3}Loop is an anchor-guided agentic annotation loop for constructing QA groups under task-specific and group-level constraints. The core loop is preceded by metadata construction and followed by distractor construction.

4.1 Video Collection and Metadata Construction

We collect videos from existing benchmarks under appropriate licenses to cover diverse video types. We filter the videos by duration, resolution, and audio. The retained videos serve as input to the metadata construction pipeline, as shown in Fig. 4.

AnchorGroup requires anchor objects that can be identified consistently across relevant video segments. The metadata pipeline therefore searches for frequently occurring objects with usable video

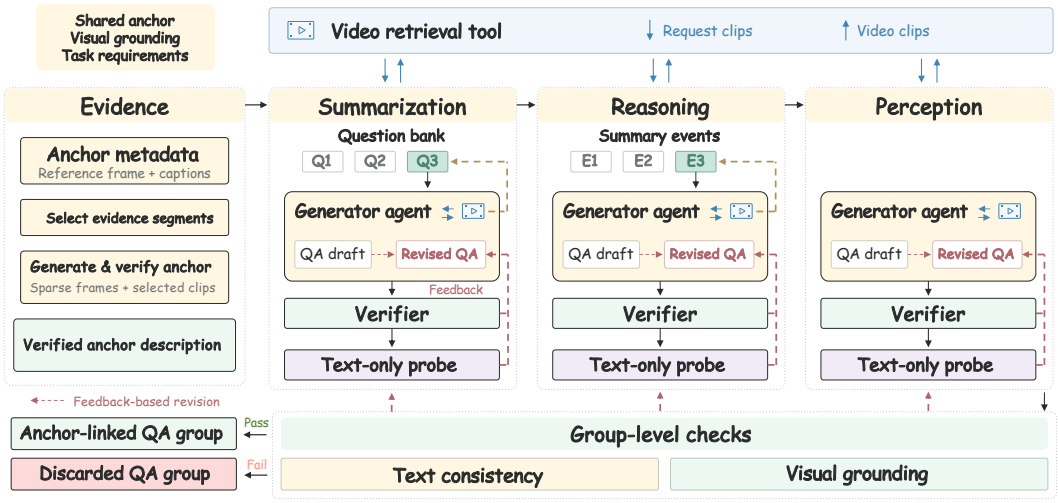

Figure 5: **Overview of A^{3}Loop.** The loop constructs object-anchored QA groups from summarization to reasoning and then perception. Generators and verifiers use video evidence, while feedback guides revision, candidate reselection, and re-verification.

spans to form a candidate anchor pool. We prioritize annotation reliability over yield, even at the cost of retaining fewer candidates.

We first divide each video into 60-second chunks with a 2-second overlap between adjacent chunks. For each chunk, the pipeline generates object captions and corresponding reference frames. The identified objects form the initial candidate anchor pool.

The pipeline iteratively merges candidates for the same object across chunks using captions and reference frames. Each batch retains its highest-confidence candidates for the next round. The merged annotations are verified, and failed records are removed. Finally, denser frame sampling within each span supports fine-grained temporal grounding and object-centric captioning.

The resulting metadata forms the candidate anchor pool for A^{3}Loop. Each candidate anchor is associated with a reference frame from the source video and multiple video spans. Each span has a caption focused on the anchor. Detailed filtering rules, retry limits, and construction statistics are provided in Appendix A.4.

4.2 A^{3}Loop: Anchor-guided Agentic Annotation Loop

A^{3}Loop jointly constructs QA groups around a shared anchor. It supports on-demand access to video clips during generation and verification rather than relying solely on dense captions. Existing annotation loops (Alam et al., 2026; Hu et al., 2026b) iteratively generate and verify individual QA pairs rather than jointly construct QA groups. A straightforward extension to our three-task setting is to generate perception, reasoning, and summarization QAs one at a time. However, individually valid QA pairs do not necessarily satisfy the shared-anchor and evidence constraints of AnchorGroup. If the summarization QA is generated last, it must cover multiple anchor-related events while incorporating the evidence already selected for reasoning. An event that supports a specific inference may not readily fit into a coherent summary of multiple events. Finding suitable additional events while retaining the selected reasoning evidence can therefore complicate summarization construction.

To address these constraints, A^{3}Loop begins with a summarization QA covering multiple events involving the same anchor. The loop then selects one of these events to construct a reasoning QA that shares an evidence segment with the summarization QA. Finally, it generates a perception QA about the same anchor. The order is designed to ease summarization construction by establishing its content before selecting the reasoning evidence. The loop coordinates generators, verifiers, and quality checkpoints through iterative feedback, as shown in Fig. 5. Feedback guides revision, candidate reselection, and re-verification under task-specific and group-level constraints.

Table 2: **Performance of MLLMs on AnchorGroup.** Text reports overall accuracy without video input. Video reports results with video input. P, R, and S denote perception, reasoning, and summarization. P^+ and P^- denote groups with correct and incorrect perception answers, respectively. Each Δ is the difference between the corresponding P^+ and P^- conditional accuracies, reported in percentage points. **All3** reports the percentage of QA groups with all three answers correct.

	Text	Video										
Model	Avg.	Avg.	P	R	S	$R \mid P^+$	$R \mid P^-$	$\Delta_{P\to R}$	$S \mid P^+$	$S \mid P^-$	$\Delta_{P\to S}$	All3
Proprietary Models												
Doubao 2.0 (Seed, 2026)	29.71	71.68	65.12	73.12	76.80	75.68	68.35	7.33	78.38	73.85	4.53	39.20
Doubao 2.1 (Bytedance Seed, 2026)	34.13	68.96	64.48	68.16	74.24	72.46	60.36	12.10	74.44	73.87	0.57	34.88
Gemini 3.6 Flash (Google DeepMind, 2026a)	36.53	73.23	67.20	75.04	77.44	78.57	67.80	10.77	77.86	76.59	1.27	42.56
Gemini 3.7 Flash (Google DeepMind, 2026b)	43.89	78.72	73.28	78.56	84.32	81.44	70.66	10.78	85.81	80.24	5.57	51.84
Gemini 3.8 Flash (Google DeepMind, 2026c)	45.71	77.44	72.16	77.12	83.04	80.49	68.39	12.10	84.92	78.16	6.76	49.92
GLM 5.3 Flash (Z.ai, 2026)	29.71	67.79	62.40	69.60	71.36	73.33	63.40	9.93	74.87	65.53	9.34	34.56
GLM 5.3 Flashx (Z.ai, 2026)	30.03	67.89	63.20	68.00	72.48	69.87	64.78	5.09	75.19	67.83	7.36	35.04
Kimi K2.6 (Moonshot AI, 2026)	30.56	70.45	63.36	71.84	76.16	75.25	65.94	9.31	79.55	70.31	9.24	38.56
Kimi K3 (Kimi Team, 2026)	32.11	74.83	68.80	76.16	79.52	77.67	72.82	4.85	81.63	74.87	6.76	44.16
MiMo V2.6 Flash (LLM-Core Xiaomi, 2026)	28.75	63.73	59.68	63.84	67.68	68.10	57.54	10.56	68.63	66.27	2.36	29.12
MiMo V2.6 Pro (LLM-Core Xiaomi, 2026)	28.21	63.95	58.24	64.48	69.12	65.66	62.84	2.82	72.25	64.75	7.50	28.80
Muse Spark 1.1 (Meta Superintelligence Labs, 2026a)	35.89	74.13	67.20	74.88	80.32	76.90	70.73	6.17	82.38	76.10	6.28	43.52
Muse Spark 1.2 (Meta Superintelligence Labs, 2026b)	34.51	74.40	67.84	75.04	80.32	77.36	70.15	7.21	80.90	79.10	1.79	43.68
Muse Spark 1.3 (Meta Superintelligence Labs, 2026c)	35.63	74.61	68.00	75.84	80.00	76.00	75.50	0.50	80.00	80.00	0.00	42.56
Qwen3.8 Max (Qwen Team, 2026d)	29.49	74.24	66.40	74.88	81.44	78.55	67.62	10.94	83.61	77.14	6.47	44.80
Step 5 preview (StepFun, 2026)	27.73	68.69	61.60	69.28	75.20	70.91	66.67	4.24	76.88	72.50	4.38	34.24
Open-Source Models												
Gemma4 12B it (Gemma Team, 2026)	26.56	38.61	34.72	39.04	42.08	45.16	35.78	9.38	40.09	43.14	-3.05	7.04
Gemma4 26B-A4B it (Gemma Team, 2026)	25.55	46.56	43.68	46.72	49.28	49.45	44.60	4.85	48.35	50.00	-1.65	11.36
Gemma4 31B it (Gemma Team, 2026)	28.59	61.39	56.32	64.80	63.04	65.34	64.10	1.24	64.77	60.81	3.97	24.80
InternVL3.5 4B (Wang et al., 2025b)	21.17	40.43	38.08	43.04	40.16	42.02	43.67	-1.65	44.54	37.47	7.07	6.88
InternVL3.5 8B (Wang et al., 2025b)	21.01	46.24	42.08	47.84	48.80	50.95	45.58	5.37	52.09	46.41	5.68	11.04
InternVL3.5 30B-A3B (Wang et al., 2025b)	24.21	47.09	39.36	49.92	52.00	49.59	50.13	-0.54	52.44	51.72	0.72	10.72
MiMoVL 7B RL (Xiaomi, 2025)	20.48	39.89	34.40	42.40	42.88	47.44	39.76	7.69	46.05	41.22	4.83	8.96
MiMoVL 7B SFT (Xiaomi, 2025)	21.49	38.19	32.96	42.88	38.72	46.12	41.29	4.83	42.23	36.99	5.24	5.92
Qwen3VL 2B Instruct (Bai et al., 2025)	21.60	40.53	39.68	40.48	41.44	41.53	39.79	1.74	43.15	40.32	2.83	8.16
Qwen3VL 4B Instruct (Bai et al., 2025)	21.76	47.95	46.56	49.60	47.68	53.26	46.41	6.86	52.92	43.11	9.81	14.24
Qwen3VL 8B Instruct (Bai et al., 2025)	20.96	51.95	46.88	53.76	55.20	59.04	49.10	9.95	55.29	55.12	0.17	15.52
Qwen3VL 30B-A3B Instruct (Bai et al., 2025)	21.81	57.81	53.12	60.00	60.32	62.35	57.34	5.01	65.06	54.95	10.11	22.40
Qwen3.5 2B (Qwen Team, 2026a)	22.56	35.73	41.28	36.32	29.60	33.72	38.15	-4.43	32.95	27.25	5.70	4.96
Qwen3.5 4B (Qwen Team, 2026a)	20.53	50.35	46.24	50.56	54.24	49.48	51.49	-2.01	57.09	51.79	5.31	13.60
Qwen3.5 9B (Qwen Team, 2026a)	23.63	52.69	46.40	53.28	58.40	55.52	51.34	4.17	61.03	56.12	4.92	16.32
Qwen3.5 35B-A3B (Qwen Team, 2026a)	24.91	56.69	54.72	57.12	58.24	57.60	56.54	1.07	62.28	53.36	8.92	19.36
Qwen3.6 27B (Qwen Team, 2026b)	27.57	57.44	53.28	59.36	59.68	62.16	56.16	6.00	62.16	56.85	5.31	20.96
Qwen3.6 35B-A3B (Qwen Team, 2026c)	25.17	58.24	53.60	62.08	59.04	63.58	60.34	3.24	60.60	57.24	3.36	21.92
Qwen3.8 27B (Qwen Team, 2026d)	27.31	57.07	54.24	60.64	56.32	61.95	59.09	2.86	57.23	55.24	1.98	20.48

Evidence selection and anchor description. To limit video tokens in the model context, the model first receives an anchor reference frame and captions for the associated video segments. It uses these inputs to select evidence segments for subsequent video input. The loop then uses sparsely sampled frames from the full video and the selected clips to generate and verify an anchor description. The description identifies the same object across relevant segments. The verified anchor description provides a common object reference for subsequent QA generation and verification. The selected video segments provide the initial visual context for QA construction.

Generators and verifiers. Each task has a generator and a verifier. Generators construct visually grounded QA pairs about the shared anchor. They can request specific clips from the source video when the initially selected segments provide insufficient information. Verifiers check visual grounding, semantic correctness, and task requirements. Failed QA pairs receive targeted feedback and undergo revision and re-verification. For summarization, the generator first builds a bank of candidate questions. When revision is needed, it selects another candidate from the bank.

Quality checkpoints. A text-only probe follows each task-specific verifier to assess whether the question can be answered without video input. Questions flagged by the probe return to the corresponding generator for revision. Two final verifiers check textual consistency and visual grounding across the three QA pairs. The group-level checks assess whether all three QAs refer to the same anchor and whether the reasoning and summarization QAs share the required evidence segment.

Loop mechanism. Verification failures automatically return annotations to the generation stage responsible for the error. Feedback guides QA revision or candidate reselection, followed by re-

verification. For reasoning, the generator can select another event from the same summarization QA if the current event cannot support a suitable question. Each task stage allows up to three retries after the initial attempt. If the stage still fails, the loop stops QA construction for the current metadata input. Groups that pass the task-specific and group-level checks are retained as complete open-ended QA groups for subsequent quality control and distractor construction.

4.3 Quality Control

A^3Loop is designed to construct QA groups under task-specific, shared-anchor, and evidence constraints. Its automated checks may still miss factual errors or ambiguities. Context length limits also restrict how densely the loop can sample frames from the full video. We therefore apply additional quality control to the generated QA groups.

We use GPT-5.6 as an additional text-only verifier to filter out QA pairs with vague answers or ambiguous wording. We convert the open-ended QAs produced by A^3Loop into four-option multiple-choice questions. Details of distractor construction are provided in Appendix A.5. We then apply a text-only test to identify questions that can be answered without video input.

Finally, Human annotators inspect relevant video segments and review the questions and answer options. They revise ambiguous question wording, cross-question answer leakage, and unsuitable distractors. Groups with factual errors are discarded. The final dataset contains 625 complete QA groups. Detailed review guidelines and outcomes are provided in Appendix A.6.

5 Experiments

We evaluate proprietary and open-source video MLLMs on AnchorGroup across multiple model families and sizes. The open- source models include Qwen3-VL (Bai et al., 2025), Qwen3.5 (Qwen Team, 2026a), Qwen3.6 (Qwen Team, 2026b;c), Qwen3.8 (Qwen Team, 2026d), Gemma 4 (Gemma Team, 2026), InternVL 3.5 (Wang et al., 2025b), and MiMo-VL (Xiaomi, 2025). The proprietary models include Qwen3.8 Max (Qwen Team, 2026d), Gemini 3.6 Flash (Google DeepMind, 2026a), Gemini 3.7 Flash (Google DeepMind, 2026b), Gemini 3.8 Flash (Google DeepMind, 2026c), Doubao 2.0 (Seed, 2026), Doubao 2.1 (Bytedance Seed, 2026), GLM 5.3 Flash (Z.ai, 2026), GLM 5.3 FlashX (Z.ai, 2026), Kimi K2.6 (Moonshot AI, 2026), Kimi K3(Kimi Team, 2026), MiMo V2.6 Pro (LLM-Core Xiaomi, 2026), MiMo V2.6 Flash (LLM-Core Xiaomi, 2026), Muse Spark 1.1 (Meta Superintelligence Labs, 2026a), Muse Spark 1.2 (Meta Superintelligence Labs, 2026b), Muse Spark 1.3 (Meta Superintelligence Labs, 2026c), and Step 5 Preview (StepFun, 2026).

5.1 Experimental Settings

To improve comparability, we use 64 sampled frames per video for all models. We disable thinking when supported and otherwise use the lowest available reasoning intensity. We deploy open-source models with vLLM using their recommended inference settings. For proprietary models, we use default API settings except for the reasoning settings described above. Detailed inference and model-specific image preprocessing settings are provided in Appendix A.3.

5.2 Main Results

Proprietary Models. Under video input, proprietary models achieve an average accuracy of 71.55%, with individual scores ranging from 63.73% to 78.72%. Gemini 3.7 Flash achieves the highest overall accuracy at 78.72%, followed by Gemini 3.8 Flash at 77.44%. Averaged across models, perception, reasoning, and summarization accuracies are 65.56%, 72.24%, and 76.84%, respectively. The average conditional accuracy gaps are +7.79 percentage points for $\Delta_{P\to R}$ and +5.01 points for $\Delta_{P\to S}$. The gap magnitudes vary considerably across models.

Open-Source Models. Under video input, open-source models achieve an average accuracy of 48.68%, with individual scores ranging from 35.73% to 61.39%. Gemma 4 31B achieves the highest accuracy at 61.39%, followed by Qwen3.6 35B-A3B at 58.24%. Averaged across models, perception, reasoning, and summarization accuracies are 45.14%, 50.52%, and 50.37%, respectively. The average conditional accuracy gaps are +3.45 percentage points for $\Delta_{P\to R}$ and +4.28 points for

$\Delta_{P\to S}$. Several models have near-zero or negative gaps, indicating that associations with perception correctness vary across models.

Proprietary vs. Open-Source Models. The evaluated proprietary models outperform the open-source models by an average of 22.87 percentage points under video input (71.55% vs. 48.68%). The lowest proprietary-model accuracy, 63.73%, exceeds the highest open-source accuracy, 61.39%. The average differences are 20.42 points for perception, 21.72 points for reasoning, and 26.47 points for summarization. The largest difference is in summarization. Proprietary models also have a larger average $\Delta_{P\to R}$, while the difference between the two model groups is smaller for $\Delta_{P\to S}$.

5.3 Further Analysis

Gains from Video Input. Proprietary models achieve higher text-only accuracy than open-source models, averaging 33.29% compared with 23.52%. The open-source average is slightly below the 25% random-choice baseline. The higher text-only scores may reflect language priors or cues in the questions and answer options. Adding video input increases average accuracy by 38.26 percentage points for proprietary models and 25.16 points for open-source models. These results show substantial benefits from video input, even for models with stronger text-only performance.

Joint Success across Tasks. All3 measures the percentage of QA groups in which all three questions are answered correctly. Proprietary and open-source models achieve average All3 scores of 39.84% and 13.93%, respectively. Gemini 3.7 Flash achieves the highest score of 51.84% and is the only evaluated model to exceed 50%. The strongest open-source model, Gemma 4 31B, achieves 24.80%. Model rankings also differ between average accuracy and All3. Muse Spark 1.3 has higher average accuracy than Qwen3.8 Max (74.61% vs. 74.24%) but lower All3 (42.56% vs. 44.80%). Higher average accuracy therefore does not necessarily imply a higher rate of success across all three tasks for the same object. All3 complements task-level accuracy by measuring whether models answer all three questions correctly for the same object.

Perception Performance and Cross-Task Associations. Perception has the lowest average accuracy in both model groups. Proprietary models average 65.56% on perception, 72.24% on reasoning, and 76.84% on summarization. All 16 proprietary models follow this ordering. Open-source models average 45.14%, 50.52%, and 50.37% on the three tasks, respectively. Reasoning accuracy is generally higher in QA groups with correct perception answers. The mean $\Delta_{P\to R}$ is +7.79 percentage points for proprietary models and +3.45 points for open-source models. However, the association varies across models. Gemini 3.8 Flash has a gap of +12.10 points, whereas Muse Spark 1.3 has a gap of only +0.50 points. Several open-source models have negative gaps. Perception is the lowest-scoring task on average, but its association with reasoning differs across models. Task averages alone do not reveal these differences in cross-task performance.

Frame Count and Cross-Task Performance. We evaluate 14 Qwen and Gemma models with 64, 128, and 256 input frames while keeping other evaluation settings unchanged. Increasing the frame count from 64 to 256 improves average accuracy by 5.76 percentage points. Mean All3 rises from 15.79% to 21.23%, with all 14 models showing higher joint success. However, the mean $\Delta_{P\to R}$ decreases from 3.57 to 1.75 percentage points, and the mean $\Delta_{P\to S}$ decreases from 4.12 to 2.91 points. More input frames can therefore improve joint success without increasing perception-conditioned accuracy gaps. Detailed results are provided in Tables 9 and 10.

6 Conclusion

In this work, we introduce AnchorGroup, a video understanding benchmark for evaluating video MLLMs across different tasks within a shared object context. We propose A^{3}Loop, an agentic annotation loop that constructs QA groups from summarization to reasoning and then perception. The loop uses video evidence during generation and verification. Feedback guides revision, candidate reselection, and re-verification under task-specific and group-level constraints. Our experiments show that perception has the lowest average accuracy among the three tasks. Reasoning and summarization accuracy is generally higher in groups with correct perception answers, although the associations vary across models. AnchorGroup supports analysis of these cross-task associations beyond separate task averages. We hope the benchmark and annotation framework support further research on object-anchored video evaluation and video-grounded QA-group construction.

AI USE STATEMENT

In this work, we used generative AI tools for "generate synthetic data sets, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis". We have not used generative AI tools for "help develop interpret results", and "theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, assisting in the writing of proofs" are not applicable to this work. Additionally, we used generative AI tools for "create or modify scientific figures or images, edit a research paper to improve readability". We have reviewed all AI-assisted work. We checked every LLM-generated text or data content before we decided to adopt them. Especially for our benchmark data generated by LLM in loop, we manually check every QA pairs to ensure the correctness and quality of our benchmark. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics and all authors have read and adhered to the Code of Ethics. In this study, no human subjects is involved. The use of all datasets follows the relevant usage guidelines and public licenses, ensuring no violation of privacy. We have been careful to avoid any biased or discriminatory results during our research process. No personally identifiable information is used, and no privacy or security concerns will be raised due to our experiments. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. All code and datasets have been made publicly available in an anonymous repository to facilitate replication and verification. The experimental setup, including annotation model configurations, evaluated model configurations, and annotation loop details, is described in detail in the paper. We have also provided a full description of AnchorGroup and A^3 Loop to assist others in reproducing our experiments. Additionally, the datasets used in our experiments are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

REFERENCES

Samiul Alam, Shakhrul Iman Siam, Michael J Proulx, James Fort, Richard Newcombe, Hyo Jin Kim, and Mi Zhang. Supermemory-vqa: An egocentric visual question-answering benchmark for long-horizon memory. *arXiv preprint arXiv:2606.00825*, 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.

Bytedance Seed. Seed2.1 Model Card: Agentic Intelligence for Productivity. `https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2.1/Seed2_1_Model_Card.pdf`, 2026. Model card. Accessed: 2026-09-26.

Meng Cao, Pengfei Hu, Yingyao Wang, Jihao Gu, Haoran Tang, Haoze Zhao, Chen Wang, Jiahua Dong, Wangbo Yu, Ge Zhang, Xiang Li, Ian Reid, and Xiaodan Liang. Video simpleqa: Towards factuality evaluation in large video language models. In *AAAI*, pp. 2616–2624. AAAI Press, 2026.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025a.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video LLM with streaming speech transcription at scale. In *CVPR*, pp. 29083–29095. Computer Vision Foundation / IEEE, 2025b.

Xinlong Chen, Jiafu Tang, Yue Ding, Yizhuo Jia, Bozhou Li, Bohan Zeng, Yang Shi, Shihao Li, Yiyan Ji, Qiang Liu, Weihong Lin, Yuanxing Zhang, Pengfei Wan, Liang Wang, and Tieniu Tan. Capricorn-1k: A comprehensive benchmark for video captioning and subject referential consistency across temporal scales. *CoRR*, abs/2606.21949, 2026.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Aditya Chetan, Eric Cai, Peeyush Kushwaha, Bharath Raj Nagoor Kani, Utkarsh Mall, Qianqian Wang, Noah Snavely, and Bharath Hariharan. Flat-pack bench: Evaluating spatio-temporal understanding in large vision-language models through furniture assembly. *CoRR*, abs/2605.21625, 2026.

OmniCapBench Contributors. Omnicapbench: A benchmark for structured multi-level temporal video captioning. https://huggingface.co/datasets/OmniCapBench/OmniCapBench, 2026.

Shaoyang Cui, Lingbei Meng, Yaodi Luo, and Peize He. Vidnum: Diagnosing vlm failure modes in video-grounded numerical reasoning, 2026. URL https://arxiv.org/abs/2604.03701.

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. *CoRR*, abs/2510.08559, 2025.

Yiming Ding, Siyu Cao, Luyuan Jiao, Yixuan Li, Zitong Wang, Zhiyong Liu, and Lu Zhang. Retrieving any relevant moments: Benchmark and models for generalized moment retrieval. *CoRR*, abs/2605.02623, 2026.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. In *NeurIPS*, 2024.

Hengyi Feng, Hao Liang, Mingrui Chen, Bohan Zeng, Meiyi Qiang, Zhengyang Zhao, Zimo Meng, Zeang Sheng, and Wentao Zhang. TraceAV-Bench: Benchmarking Multi-Hop Trajectory Reasoning over Long Audio-Visual Videos. *arXiv preprint arXiv:2605.07593*, 2026a. URL https://arxiv.org/abs/2605.07593.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026b.

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

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos. In *ICCV*, pp. 21778–21788. IEEE, 2025b.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *CVPR*, pp. 18959–18969. Computer Vision Foundation / IEEE, 2025.

Google DeepMind. Gemini 3.6 Flash Model Card. `https://deepmind.google/models/model-cards/gemini-3-6-flash/`, 2026a. Accessed: 2026-09-26.

Google DeepMind. Gemini 3.7 Flash Model Card. `https://deepmind.google/models/model-cards/gemini-3-7-flash/`, 2026b. Accessed: 2026-09-26.

Google DeepMind. Gemini 3.8 Flash Model Card. `https://deepmind.google/models/model-cards/gemini-3-8-flash/`, 2026c. Accessed: 2026-09-26.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *CVPR*, pp. 8450–8460. Computer Vision Foundation / IEEE, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27798–27828, 2026a.

Rongsheng Hu, Runwei Guan, Yicheng Di, Jiayu Bao, and Yuan Liu. Autovqa-g: Self-improving agentic framework for automated visual question answering and grounding annotation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12312–12316. IEEE, 2026b.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *CVPR*, pp. 3328–3338. Computer Vision Foundation / IEEE, 2025.

Kimi Team. Kimi K3: open frontier intelligence. *CoRR*, abs/2607.24653, 2026.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. Siv-bench: A video benchmark for social interaction understanding and reasoning. In *ACL (Findings)*, pp. 37379–37403. Association for Computational Linguistics, 2026.

Ilia Kulikov, Chenxi Whitehouse, Tianhao Wu, Yixin Nie, Swarnadeep Saha, Eryk Helenowski, Weizhe Yuan, Olga Golovneva, Jack Lanchantin, Yoram Bachrach, et al. Autodata: An agentic data scientist to create high quality synthetic data. *arXiv preprint arXiv:2606.25996*, 2026.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, Ke Wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi Li, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *CoRR*, abs/2510.10689, 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 *AAAI*, pp. 6244–6252. AAAI Press, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pp. 22195–22206. IEEE, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Xianjie Liu, Yiman Hu, Liang Wu, Ping Hu, Yixiong Zou, Jian Xu, and Bo Zheng. E-vads: An e-commerce short videos understanding benchmark for mllms. *CoRR*, abs/2602.08355, 2026.

Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y. Charles, Xinyu Zhou, and Xu Sun. Videoreasonbench: Can mllms perform vision-centric complex video reasoning? *CoRR*, abs/2505.23359, 2025.

LLM-Core Xiaomi. MiMo-V2.6: Scaling Reinforcement Learning Towards Self-Improvement. https://huggingface.co/XiaomiMiMo/MiMo-V2.6-Pro-RL/blob/main/MiMo_V2_6_technical_report.pdf, 2026. Technical report. Accessed: 2026-09-26.

Lidong Lu, Guo Chen, Zhu Wei, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms. In *CVPR*, pp. 33477–33487. Computer Vision Foundation, 2026.

Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. Videoeval-pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*, 2025.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. Videozerobench: Probing the limits of video mllms with spatio-temporal evidence verification. *arXiv preprint arXiv:2604.01569*, 2026.

Meta Superintelligence Labs. Introducing Muse Spark 1.1. https://research.meta.ai/blog/introducing-muse-spark-meta-model-api, 2026a. Accessed: 2026-09-26.

Meta Superintelligence Labs. Introducing Muse Code and Muse Spark 1.2. https://research.meta.ai/blog/introducing-muse-code-and-muse-spark-1-2, 2026b. Accessed: 2026-09-26.

Meta Superintelligence Labs. Introducing Muse Spark 1.3. https://research.meta.ai/blog/introducing-muse-spark-1-3, 2026c. Accessed: 2026-09-26.

Moonshot AI. Kimi K2.6: Advancing Open-Source Coding. https://www.kimi.com/en/blog/kimi-k2-6, 2026. Technical blog. Accessed: 2026-09-26.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. Ovo-bench: How far is your video-llms from real-world online video understanding? In *CVPR*, pp. 18902–18913. Computer Vision Foundation / IEEE, 2025.

Tianhao Peng, Haochen Wang, Yuanxing Zhang, Noah Wang, Zili Wang, Ge Zhang, Jian Yang, Shihao Li, Yanghai Wang, Xintao Wang, Houyi Li, Wei Ji, Pengfei Wan, Wenhao Huang, Zhaoxiang Zhang, and Jiaheng Liu. Mvu-eval: Towards multi-video understanding evaluation for multimodal llms. In *NeurIPS*, 2025.

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

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

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

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

Hanoona Abdul Rasheed, Abdelrahman M. Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman H. Khan, and Fahad Shahbaz Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *CoRR*, abs/2506.05349, 2025.

ByteDance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *CoRR*, abs/2607.00248, 2026.

Shaden Shaar, Bradon Thymes, Sirawut Chaixanien, Claire Cardie, and Bharath Hariharan. Movierecapsqa: A multimodal open-ended video question-answering benchmark. *arXiv preprint arXiv:2601.02536*, 2026.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, pp. 18221–18232. IEEE, 2024.

StepFun. Step 5 Preview: Advancing the Pareto Frontier. `https://www.stepfun.com/step-5-preview`, 2026. Accessed: 2026-09-26.

Peiwen Sun, Xudong Lu, Huadai Liu, Yang Bo, Dongming Wu, Huankang Guan, Minghong Cai, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Rui Liu, and Xiangyu Yue. X-stream: Exploring mllms as multiplexers for multi-stream understanding. *CoRR*, abs/2606.02482, 2026a.

Shangkun Sun, Ruyang Liu, Haoran Tang, Yixiao Ge, Haibo Lu, Jiankun Yang, and Chen Li. Ppllava: Varied video sequence understanding with prompt guidance. In *International Conference on Learning Representations*, volume 2026, pp. 139734–139751, 2026b.

Sirnam Swetha, Rohit Gupta, Parth Parag Kulkarni, David G Shatwell, Jeffrey A Chan Santiago, Nyle Siddiqui, Joseph Fioresi, and Mubarak Shah. Vrr-qa: Visual relational reasoning in videos beyond explicit cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32840–32849, 2026.

Jiahao Wang, An Ping, Yanghai Wang, Yuanxing Zhang, Shihao Li, Hanyan Bian, Yichi Ren, Yize Zhang, Han Wang, Haowen Chen, Junze Li, Jiaqi Wang, Yiyang Hu, Zhuze Xu, Zijie Zhang, and Jiaheng Liu. Omnicap-if: Benchmarking and improving instruction following abilities for omni-video captioning. *CoRR*, abs/2606.08572, 2026a.

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

Yun Wang, Junbin Xiao, Han Lyu, Yifan Wang, Jing Zuo, Zhanjie Zhang, Hong Huang, Dapeng Wu, and Angela Yao. Keep it in mind: User centric continual spatial intelligence reasoning in egocentric video streams. *CoRR*, abs/2606.15200, 2026b.

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.

Xuecheng Wu, Dingkang Yang, Danlei Huang, Xinyi Yin, Yifan Wang, Jia Zhang, Jiayu Nie, Liangyu Fu, Yang Liu, Junxiao Xue, Hadi Amirpour, and Wei Zhou. emotions: A large-scale dataset and audio-visual fusion network for emotion analysis in short-form videos. *CoRR*, abs/2508.06902, 2025.

Zhengqian Wu, Zhixian Liu, Aodong Chen, Jingyang Zhang, Ruizhe Li, Hanlin Ge, Zhongyuan Wang, Chunxia Xiao, and Chao Liang. Storyvideoqa: Scaling deep video understanding with a large-scale, multi-genre and auto-generated dataset. *International Journal of Computer Vision*, 134(6):308, 2026.

Junbin Xiao, Shenglang Zhang, Pengxiang Zhu, and Angela Yao. Ego-grounding for personalized question-answering in egocentric videos. In *CVPR*, pp. 40537–40547. Computer Vision Foundation, 2026.

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL `https://arxiv.org/abs/2506.03569`.

Yicheng Xu, Yue Wu, Jiashuo Yu, Ziang Yan, Tianxiang Jiang, Yinan He, Qingsong Zhao, Kai Chen, Yu Qiao, Limin Wang, Manabu Okumura, and Yi Wang. ExpVid: A Benchmark for Experiment Video Understanding & Reasoning. In *International Conference on Learning Representations*, pp. 83310–83348, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/86833e97827ff52d297a9fbde5842a81-Paper-Conference.pdf.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, Zhenxiang Li, Zhongying Tu, Conghui He, Yu Qiao, Yali Wang, Yi Wang, and Limin Wang. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *ICCV*, pp. 21655–21666. IEEE, 2025.

Z.ai. GLM-5.3-Flash/FlashX. https://docs.z.ai/guides/vlm/glm-5.3-flash, 2026. Model documentation. Accessed: 2026-09-26.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. Streamforest: Efficient online video understanding with persistent event memory. In *NeurIPS*, 2025.

Ruixu Zhang, Deyi Ji, Lanyun Zhu, Xuanyi Liu, Yuxin Meng, Ruihang Chu, and Yujiu Yang. Videozero: Self-evolution video understanding. *arXiv preprint arXiv:2605.14733*, 2026.

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

Ruixiang Zhao, Jie Yang, Zijie Xin, Tianyi Wang, Fengyun Rao, Jing Lyu, and Xirong Li. Omnipro: A comprehensive benchmark for omni-proactive streaming video understanding. *CoRR*, abs/2605.18577, 2026.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: measuring expert-level multi-discipline video understanding. In *CVPR*, pp. 8475–8489. Computer Vision Foundation / IEEE, 2025.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*, 2024.

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

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. MMR-V: what's left unsaid? A benchmark for multimodal deep reasoning in videos. *CoRR*, abs/2506.04141, 2025.

Yuanhao Zou, Arthad Kulkarni, Lucas Tonanez, Lincoln Spencer, Guangyu Sun, Tianxingjian Ding, Andong Deng, Yi Li, Shuangjun Liu, Yuan Li, Dashan Gao, Ning Bi, Taotao Jing, Shuai Zhang, and Chen Chen. Conditional multi-event temporal grounding in long-form video. *CoRR*, abs/2606.15320, 2026.

A Appendix

Appendix is organized as follows:

- SectionA.1: Video Source and Categories Distribution
- SectionA.2: Detailed Task Definition
- SectionA.3: Detailed Settings of Evaluation
- SectionA.4: Detailed Loop Agent Settings and Statistics
- SectionA.5: Detailed Process of Distractor Construction

A.1 Video Collection

Table 3 reports the video sources of AnchorGroup. Fig. 6 presents the video categories defined in this work. Both distributions count QA groups rather than unique videos. Each video is counted once for every associated QA group.

Table 3: **Video source distribution.** Counts are based on the 625 QA groups in AnchorGroup.

Source	Count	Source	Count
CrossVid (Li et al., 2026)	86	MMBench-Video (Fang et al., 2024)	9
MLVU (Zhou et al., 2025)	83	OmniPro (Zhao et al., 2026)	9
CapRiCorn-1K (Chen et al., 2026)	68	ExpVid (Xu et al., 2026)	8
VideoMME-v2 (Fu et al., 2026)	40	LongVALE (Geng et al., 2025)	8
AV-Reasoner (Lu et al., 2026)	39	MMR-VBench (Zhu et al., 2025)	7
EMOTIONS (Wu et al., 2025)	30	E-VADS (Liu et al., 2026)	6
TraceAV-Bench (Feng et al., 2026a)	28	OVBench (Huang et al., 2025)	6
VideoEval-Pro (Ma et al., 2025)	27	Flat-Pack-Bench (Chetan et al., 2026)	2
CoMET-Bench (Zou et al., 2026)	22	OmniCapBench (Contributors, 2026)	2
SIV-Bench (Kong et al., 2026)	20	SciVideoBench (Deng et al., 2025)	2
CG-Bench (Chen et al., 2025a)	19	LiveSports-3K (Chen et al., 2025b)	1
Video-MME (Fu et al., 2025a)	18	LVBench (Wang et al., 2025a)	1
UCS-Bench (Wang et al., 2026b)	17	MVU-Eval (Peng et al., 2025)	1
X-STREAM (Sun et al., 2026a)	16	MovieChat-1K-Test (Song et al., 2024)	1
LongVideoBench (Wu et al., 2024)	13	OmniCap-IF (Wang et al., 2026a)	1
MyEgo (Xiao et al., 2026)	11	OVO-Bench (Niu et al., 2025)	1
Soccer-GMR (Ding et al., 2026)	11	VideoMathQA (Rasheed et al., 2025)	1
OmniVideoBench (Li et al., 2025)	10	VidNum-1.4K (Cui et al., 2026)	1
		Total	**625**



Figure 6: **Video category distribution.** Counts are based on the 625 QA groups in AnchorGroup.

A.2 Task Description

Table 4 lists the task types and definitions for perception, reasoning, and summarization.

A.3 Implementation Details

For proprietary models, we use default inference settings except for reasoning intensity. We disable thinking when supported and otherwise use the lowest available reasoning intensity. The latter applies to Gemini 3.7 Flash, Gemini 3.8 Flash, the evaluated GLM and Muse Spark models, and Step 5 Preview.

For open-source models, we follow their recommended configurations on Hugging Face and disable thinking, as detailed in Table 5. Original-resolution inputs for InternVL 3.5 and MiMo-VL could not be processed in our evaluation setup. We therefore reduce their frame resolutions following their official examples.

A.4 A^{3}Loop Settings and Process

Settings. We use Qwen3.8-Flash-Next as the base model for all agents in A^{3}Loop. Each agent uses a role-specific prompt and inference configuration. Table 6 provides the detailed settings.

Construction and screening. We initially collect approximately 300,000 videos. About 90,000 remain after metadata-based filtering and enter the metadata construction pipeline. For span filtering, we use a sampling rate of 32 frames per 60 seconds and remove spans containing fewer than three sampled frames. We retain anchors with at least three remaining spans. We also discard anchors whose combined span duration exceeds 80% of the source video, counting overlapping portions only once. The resulting 23,415 candidate metadata inputs are used for QA generation.

A^{3}Loop produces 8,392 QA groups that pass its internal checks. An additional review using GPT-5.6 retains 6,423 groups. We then iteratively refine the GPT-5.6 review prompt based on human spot checks and apply further screening. Further screening retains 1,095 QA groups for subsequent manual inspection. Manual inspection retains 625 complete QA groups. The loop output rate is 35.84%, while the final retention rate is 2.67% of the 23,415 inputs.

Agent call statistics. Table 7 summarizes the recorded agent calls. The full run contains 1,160,966 calls, averaging 49.58 calls per input. The 625 retained QA groups are associated with 29,905

Table 4: **Task taxonomy of AnchorGroup.** Each QA group contains one perception, one reasoning, and one summarization QA pair.

Task Level	Task Type	Description
Perception	Visible Action	Identify a directly observable action or gesture performed by the anchor in a specified scene.
	Salient Object Interaction	Identify the salient object that the anchor holds, uses, manipulates, or physically interacts with.
	Body State	Recognize the anchor's visible posture, physical condition, orientation, or body configuration.
	Person Arrangement	Determine the anchor's position in the scene or relative to other people.
Reasoning	Intent Inference	Infer the anchor's likely goal, purpose, or intention from observed behavior and context.
	State Explanation	Infer an unobserved state or circumstance that explains an observed condition of the anchor or an entity it interacts with.
	Emotion Inference	Infer the anchor's emotional state or reaction from behavioral and contextual evidence.
	Interaction Nature Inference	Infer the nature or quality of the interaction between the anchor and another entity.
	Role Inference	Infer the anchor's functional or social role from observed behavior and context.
	Causal Implication	Infer a causal relation, consequence, or implication connecting observed events.
	Quantitative Inference	Derive a count, quantity, or numerical relation by integrating visual evidence.
Summarization	Segment Description	Describe the relevant actions or events involving the anchor within a local temporal segment.
	Timeline Summary	Summarize the events involving the anchor in chronological order across the video.
	Key-Point Summary	Summarize the anchor's major activities or key events without requiring a complete chronological account.
	Theme Summary	Summarize the overarching theme or central idea conveyed by anchor-related events.

calls, averaging 47.85 calls per group. The retained-group statistics exclude calls associated with discarded inputs and do not represent the overall production cost per retained group.

Loop yield and failure analysis. Table 8 summarizes terminal outcomes, recorded failure categories, and termination stages. The outer technical retry budget allows up to three retries. This budget is separate from task-level revision limits and the global 100-step loop limit. Of the 23,415 inputs, 8,392 produce QA groups that pass the loop checks, 14,646 are discarded during construction or quality gating, and 377 terminate after exhausting the outer technical retry budget. Technical retry exhaustion accounts for 2.51% of the 15,023 unsuccessful inputs. The two largest recorded failure categories are builder failure (6,578 inputs) and policy rejection (5,636 inputs), including rejection by text-only leakage checks.

Summarization is the most frequent termination stage among the 14,646 discarded inputs, accounting for 7,079 cases (48.33%). Of these, 6,336 terminate after exhausting the summarization candidate bank. Other frequent terminal paths include unresolved joint text-only leakage (1,943), exhausted perception search (1,789), the global 100-step loop limit (1,582), and repeated duplicate

Table 5: **Inference settings for the evaluated open-source models.** All listed models use 64 frames sampled uniformly from the full video. The maximum output length is 4,096 tokens, and the random seed is 42.

Model Family	Evaluated Checkpoints	Video Input	Inference Mode	Temp.	Top-p	Top-k
Gemma 4	12B-it; 26B-A4B-it; 31B-it	64 uniformly sampled JPEG frames	Non-thinking	1.0	0.95	64
InternVL 3.5	4B; 8B; 30B-A3B	64 interleaved JPEG frames, resized to 448×448	Non-thinking	0.6	0.95	50
MiMo-VL	7B-SFT-2508; 7B-RL-2508	64 uniformly sampled JPEG frames, at most 200,704 pixels per frame	Non-thinking (`/no_think`)	0.3	0.95	–
Qwen3-VL	2B-Instruct; 4B-Instruct; 8B-Instruct; 30B-A3B-Instruct	64 uniformly sampled JPEG frames	Non-thinking	0.7	0.80	20
Qwen3.5	2B; 4B; 9B; 35B-A3B	64 uniformly sampled JPEG frames	Non-thinking	0.7	0.80	20
Qwen3.6	27B; 35B-A3B	64 uniformly sampled JPEG frames	Non-thinking	0.7	0.80	20
Qwen3.8	27B	64 uniformly sampled JPEG frames	Non-thinking	0.7	0.80	20

generation (1,097). These counts describe where construction terminates, rather than failure rates conditional on reaching each stage.

The recorded failures mainly concern candidate construction, quality checks, and search limits rather than exhaustion of technical retries. The final retention rate additionally reflects the subsequent model-based screening and manual inspection.

A.5 Distractor Construction

In this section, we provide detailed steps for constructing distractor. **Stage-Aware Distractor Construction.** We construct distractors separately for Perception, Reasoning, and Summarization rather than applying a single generic generation strategy. In all cases, distractors are conditioned on the public Question and the semantic form of the reference answer, but they are generated without access to the video. Each distractor must answer the same requested variable, match the Gold answer in grammatical form, specificity, granularity, and information load, and remain independently plausible from the Question alone. At the same time, it must be mutually exclusive with the Gold answer in the exact scope of the Question. We exclude synonyms, compatible details, subtype–supertype relations, explicit negations, and malformed or conspicuously implausible alternatives.

Task-Specific Hard Negatives. For Perception, we use an answer-conditioned role-inversion strategy: the generator is instructed to treat the provided answer candidate as incorrect and produce three coordinate alternatives in the same visual slot and ontology. A separate verifier then treats the provided candidate as Gold and checks whether every alternative is natural, distinct, and mutually exclusive with it. For Reasoning, type-specific generators construct three competing values for the same latent semantic variable, such as intent, emotion, role, interaction nature, or state. The alternatives preserve all Question-specified actors, actions, circumstances, and temporal constraints while changing only the requested inference value. For multi-event Summarization, we identify two independent binary factual axes, treat Gold as code 00, and generate 10, 01, and 11 alternatives. This balanced 2×2 construction ensures that every original and altered attribute appears equally often, while all four options preserve the same event structure, chronology, detail level, and linguistic form.

Table 6: **Agent configurations for A^{3}Loop.** All listed agents use Qwen3.8-Flash-Next through vLLM. Shared settings are min-$p = 0$, repetition penalty 1.0, presence penalty 0, a 300-second timeout, and seeded decoding.

Agent	Function	Input	Temp.	Top-p	Top-k	Max Tokens
Evidence Selector	Select candidate occurrences and check preliminary identity	Video	0.2	0.8	20	16,384
Referring Builder	Construct the shared anchor referring expression	Video	0.2	0.8	20	16,384
Referring Verifier	Verify anchor identity and referring-expression validity	Video	0.2	0.8	20	16,384
Summarization Bank Builder	Propose grounded summarization candidates	Video	0.2	0.8	20	16,384
Summarization Bank Verifier	Audit candidate coverage, core precision, and answer neutrality	Video	0.0	0.8	20	16,384
Summarization Builder	Construct a QA pair from a verified candidate	Video	0.2	0.8	20	16,384
Open Text Probe	Answer an individual question without video	Text	0.7	0.8	20	16,384
Text Probe Judge	Judge whether a text-only answer recovers the reference meaning	Text	0.0	0.8	20	16,384
Summarization Verifier	Visually verify facts, answer, rubric, and summarization capability	Video	0.2	0.8	20	16,384
Reasoning Builder	Construct a local semantic-inference QA	Video	0.2	0.8	20	16,384
Reasoning Verifier	Verify visual support and inference validity	Video	0.2	0.8	20	16,384
Reasoning–Summarization Pair Verifier	Check semantic compatibility between the reasoning and summarization tasks	Text	0.2	0.8	20	16,384
Joint Text Probe	Answer multiple public questions jointly without video	Text	0.7	0.8	20	16,384
Joint Text Probe Judge	Judge cross-question text-only leakage	Text	0.0	0.8	20	16,384
Perception Builder	Construct a directly observable perception QA	Video	0.2	0.8	20	16,384
Perception Verifier	Verify direct visual support and perceptual distinctness	Video	0.2	0.8	20	16,384
Group Verifier	Check anchor consistency and cross-question contradictions	Text	0.2	0.8	20	16,384
Final Quality Verifier	Perform the final independent visual quality audit	Video	0.1	0.8	20	16,384

Shortcut Filtering and Final Review. To reduce option-only shortcuts, we generate three candidate option sets using seeds 43, 44, and 45. Each set is evaluated from the Question and options alone under four answer-position rotations, with three leakage votes per rotation evaluation. We select a leakage-clean variant independently for each stage, preventing one weak option set from invalidating the complete chain. The correct answer and distractors are then deterministically shuffled to avoid position bias. Persistent leakage, stylistic imbalance, or construction failures are routed to targeted GPT-5.6 repair, which preserves the Gold semantics while replacing defective distractors. Finally, human review verifies the correct option, mutual exclusivity, plausibility, parallel structure, and absence of superficial cues before inclusion in the benchmark.

A.6 Manual Quality Inspection

We recruit three annotators, each holding at least a bachelor's degree. All annotators receive training and follow standardized review and revision guidelines.

Table 7: **Agent invocation statistics.** The full run covers 23,415 candidate metadata inputs. The retained-group columns report calls associated with the 625 final QA groups. A call denotes one telemetry-recorded schema attempt. A video call carries at least one video or frame payload. Calls/Input and Calls/Group use 23,415 and 625 as their respective denominators.

Agent	Full Production Run			Retained 625 QA Groups		
	Calls	Video Calls	Calls/Input	Calls	Video Calls	Calls/Group
Evidence Selector	34,432	34,432	1.47	732	732	1.17
Referring Builder	39,882	39,882	1.70	838	838	1.34
Referring Verifier	39,706	39,706	1.70	838	838	1.34
Summarization Bank Builder	36,516	36,516	1.56	777	777	1.24
Summarization Bank Verifier	33,621	33,621	1.44	718	718	1.15
Summarization Builder	43,225	42,763	1.85	879	876	1.41
Open Text Probe	177,203	0	7.57	4,128	0	6.60
Text Probe Judge	179,205	0	7.65	4,096	0	6.55
Summarization Verifier	38,734	38,734	1.65	976	976	1.56
Reasoning Builder	115,237	94,635	4.92	2,275	1,929	3.64
Reasoning Verifier	59,866	59,866	2.56	1,353	1,353	2.16
Reasoning–Summarization Pair Verifier	47,965	0	2.05	1,145	0	1.83
Joint Text Probe	99,239	0	4.24	3,385	0	5.42
Joint Text Probe Judge	101,334	0	4.33	3,431	0	5.49
Perception Builder	48,615	33,407	2.08	1,277	1,006	2.04
Perception Verifier	25,049	25,049	1.07	928	928	1.48
Group Verifier	22,434	0	0.96	860	0	1.38
Final Quality Verifier	18,703	18,703	0.80	1,269	1,269	2.03
Total	**1,160,966**	**497,314**	**49.58**	**29,905**	**12,240**	**47.85**

Table 8: Yield and failure distribution of the QA construction loop. Shares are computed within the denominator stated for each analysis level.

Analysis Level	Category	Count	Share
Terminal outcome (denominator: 23,415 candidates)			
	Successful QA group	8,392	35.84%
	Quality-gated discard	14,646	62.55%
	Technical retry exhausted	377	1.61%
Root cause (denominator: 15,023 unsuccessful candidates)			
	Builder failure	6,578	43.79%
	Policy rejection	5,636	37.52%
	Evidence-selection insufficiency	1,214	8.08%
	Verifier disagreement	843	5.61%
	Source-sample insufficiency	375	2.50%
	Technical retry exhausted	377	2.51%
Failed stage (denominator: 14,646 discarded candidates)			
	Summarization	7,079	48.33%
	Perception	5,748	39.25%
	Reasoning	1,002	6.84%
	Evidence or group-level gate	817	5.58%

Annotators inspect relevant segments of the source videos, review the questions and all answer options, and identify supporting evidence segments. The review guidelines are as follows:

1. Check whether the anchor description uniquely identifies an object instance in the video. Discard the entire QA group if the description is ambiguous.

2. Verify the factual correctness and uniqueness of each reference answer. Discard the entire group if any QA pair contains a factual error. If multiple answers are valid because the question scope is unclear, attempt to narrow the scope. Discard the group if the ambiguity cannot be resolved.

Table 9: **Effect of input frame count on AnchorGroup.** Accuracy (%) is reported on the same 625 QA groups using 64, 128, and 256 frames. All other evaluation settings remain unchanged. P, R, and S denote perception, reasoning, and summarization. The final three columns report gains in overall accuracy, in percentage points.

	64 Frames				128 Frames				256 Frames				Average Gain		
Model	Avg.	P	R	S	Avg.	P	R	S	Avg.	P	R	S	Δ_{128-64}	Δ_{256-64}	$\Delta_{256-128}$
Qwen3VL 2B Instruct	40.53	39.68	40.48	41.44	43.95	43.36	43.68	44.80	45.71	45.92	46.24	44.96	+3.41	+5.17	+1.76
Qwen3VL 4B Instruct	47.95	46.56	49.60	47.68	51.31	50.08	53.76	50.08	54.77	54.56	55.04	54.72	+3.36	+6.83	+3.47
Qwen3VL 8B Instruct	51.95	46.88	53.76	55.20	56.37	52.96	56.16	60.00	58.19	54.88	59.84	59.84	+4.43	+6.24	+1.81
Qwen3VL 30B-A3B Instruct	57.81	53.12	60.00	60.32	60.75	57.76	58.56	65.92	64.11	60.16	63.68	68.48	+2.93	+6.29	+3.36
Qwen3.5 2B	35.73	41.28	36.32	29.60	39.41	44.96	39.04	34.24	39.52	46.08	40.16	32.32	+3.68	+3.79	+0.11
Qwen3.5 4B	50.35	46.24	50.56	54.24	54.88	51.68	54.88	58.08	57.60	54.88	57.60	60.32	+4.53	+7.25	+2.72
Qwen3.5 9B	52.69	46.40	53.28	58.40	55.04	50.88	56.64	57.60	56.48	51.20	57.44	60.80	+2.35	+3.79	+1.44
Qwen3.5 35B-A3B	56.69	54.72	57.12	58.24	63.31	60.48	64.16	65.28	65.23	61.12	66.24	68.32	+6.61	+8.53	+1.92
Qwen3.6 27B	57.44	53.28	59.36	59.68	62.99	57.92	64.96	66.08	68.11	63.20	68.64	72.48	+5.55	+10.67	+5.12
Qwen3.6 35B-A3B	58.24	53.60	62.08	59.04	60.80	55.84	64.00	62.56	61.28	55.52	64.80	63.52	+2.56	+3.04	+0.48
Qwen3.8 27B	57.07	54.24	60.64	56.32	61.07	59.04	63.68	60.48	61.23	56.80	63.84	63.04	+4.00	+4.16	+0.16
Gemma4 12B it	38.61	34.72	39.04	42.08	42.72	39.52	43.36	45.28	42.51	38.88	44.96	43.68	+4.11	+3.89	-0.21
Gemma4 26B-A4B it	46.56	43.68	46.72	49.28	48.64	44.16	49.92	51.84	49.76	50.56	49.60	49.12	+2.08	+3.20	+1.12
Gemma4 31B it	61.39	56.32	64.80	63.04	63.09	58.08	64.96	66.24	69.17	66.40	69.92	71.20	+1.71	+7.79	+6.08

3. Check whether all three questions refer to the same anchor and whether any question reveals an answer to another question. Discard groups with inconsistent anchor references. For cross-question answer leakage, attempt to revise the question wording. Discard the group if the leakage cannot be resolved.
4. Review the distractors and revise unsuitable options.

We distribute 1,095 automatically generated and screened QA groups among the three annotators. Each annotator reviews approximately one-third of the groups. Cases that annotators cannot resolve are discussed with an author to reach a final decision. Reviewing each individual QA group takes 3-6 minutes. The review process is completed within four working days.

Of the 1,095 QA groups, 75 are accepted without changes, 550 are accepted after revision, and 470 are rejected. The final dataset therefore contains 625 complete QA groups. Most rejections involve cases that require substantial manual effort for factual verification or distractor revision. We discard these groups to limit the manual review workload.

We evaluate text-only performance using Qwen3.8-Flash-Next on the 1,095 pre-review QA groups and the 625 retained groups after review. Post-review accuracy is lower by 3.97 percentage points for perception, 3.82 points for reasoning, and 6.64 points for summarization, averaging 4.81 points. The final mean text-only accuracy is 26.24%, close to the 25% random-choice baseline. The comparison includes both sample filtering and annotation revision.

A.7 Effect of Input Frame Count

Task-level accuracy and average accuracy. We evaluate 14 Qwen and Gemma models with 64, 128, and 256 input frames. Table 9 reports overall and task-level accuracy under each setting.

Increasing the input frame count from 64 to 128 improves overall accuracy for all 14 models, with an average gain of 3.66 percentage points. Using 256 frames yields an average gain of 5.76 points over 64 frames. Perception shows the largest average gain (+6.39 points), followed by summarization (+5.59) and reasoning (+5.30). The results indicate that all three tasks benefit from denser video sampling, with perception showing the greatest average sensitivity to frame count.

The average gain from 128 to 256 frames is smaller at 2.10 percentage points, but individual models show different trends. Qwen3.6 27B, Qwen3.5 35B-A3B, and Gemma 4 31B improve by 10.67, 8.53, and 7.79 points from 64 to 256 frames, respectively. Qwen3.5 2B and Qwen3.8 27B gain only 0.11 and 0.16 points from 128 to 256 frames, while Gemma 4 12B decreases by 0.21 points. Increasing the frame count therefore does not produce uniform gains across models, and higher frame counts do not always improve accuracy.

Joint success and conditional accuracy. Table 10 reports perception-conditioned accuracy and All3 under different frame counts. Averaged across the 14 models, All3 increases from 15.79% at 64 frames to 19.07% at 128 frames and 21.23% at 256 frames. All 14 models achieve higher All3

Table 10: **Perception-conditioned accuracy and joint success under different frame counts.** Conditional accuracies and All3 are reported as percentages. Values in parentheses indicate the numbers of QA groups with all three answers correct. $\Delta_{P \to R}$ and $\Delta_{P \to S}$ are reported in percentage points.

Model	Frames	$R \mid P^+$	$R \mid P^-$	$\Delta_{P \to R}$	$S \mid P^+$	$S \mid P^-$	$\Delta_{P \to S}$	All3
Qwen3VL 2B Instruct	64	41.53	39.79	+1.74	43.15	40.32	+2.83	8.16 (51)
	128	45.39	42.37	+3.01	47.60	42.66	+4.95	9.92 (62)
	256	49.13	43.79	+5.34	48.08	42.31	+5.78	12.80 (80)
Qwen3VL 4B Instruct	64	53.26	46.41	+6.86	52.92	43.11	+9.81	14.24 (89)
	128	56.23	51.28	+4.95	52.40	47.76	+4.64	15.04 (94)
	256	54.84	55.28	-0.44	56.89	52.11	+4.78	18.24 (114)
Qwen3VL 8B Instruct	64	59.04	49.10	+9.95	55.29	55.12	+0.17	15.52 (97)
	128	58.01	54.08	+3.92	58.61	61.56	-2.95	17.76 (111)
	256	59.77	59.93	-0.16	58.60	61.35	-2.75	20.64 (129)
Qwen3VL 30B-A3B Instruct	64	62.35	57.34	+5.01	65.06	54.95	+10.11	22.40 (140)
	128	60.66	55.68	+4.98	70.64	59.47	+11.17	25.60 (160)
	256	63.83	63.45	+0.38	68.88	67.87	+1.01	27.52 (172)
Qwen3.5 2B	64	33.72	38.15	-4.43	32.95	27.25	+5.70	4.96 (31)
	128	43.06	35.76	+7.30	35.59	33.14	+2.45	7.84 (49)
	256	39.24	40.95	-1.71	32.29	32.34	-0.05	7.52 (47)
Qwen3.5 4B	64	49.48	51.49	-2.01	57.09	51.79	+5.31	13.60 (85)
	128	54.18	55.63	-1.45	62.23	53.64	+8.59	18.40 (115)
	256	55.98	59.57	-3.60	60.64	59.93	+0.71	20.64 (129)
Qwen3.5 9B	64	55.52	51.34	+4.17	61.03	56.12	+4.92	16.32 (102)
	128	55.97	57.33	-1.35	59.43	55.70	+3.73	18.08 (113)
	256	56.56	58.36	-1.80	62.81	58.69	+4.12	19.68 (123)
Qwen3.5 35B-A3B	64	57.60	56.54	+1.07	62.28	53.36	+8.92	19.36 (121)
	128	67.99	58.30	+9.69	68.25	60.73	+7.53	28.96 (181)
	256	68.06	63.37	+4.69	68.59	67.90	+0.69	29.28 (183)
Qwen3.6 27B	64	62.16	56.16	+6.00	62.16	56.85	+5.31	20.96 (131)
	128	66.57	62.74	+3.84	68.23	63.12	+5.11	27.52 (172)
	256	70.63	65.22	+5.42	74.18	69.57	+4.61	33.44 (209)
Qwen3.6 35B-A3B	64	63.58	60.34	+3.24	60.60	57.24	+3.36	21.92 (137)
	128	62.18	66.30	-4.13	66.48	57.61	+8.87	24.80 (155)
	256	66.86	62.23	+4.63	64.84	61.87	+2.97	24.32 (152)
Qwen3.8 27B	64	61.95	59.09	+2.86	57.23	55.24	+1.98	20.48 (128)
	128	65.04	61.72	+3.32	65.04	53.91	+11.13	25.76 (161)
	256	64.23	63.33	+0.89	65.07	60.37	+4.70	25.28 (158)
Gemma4 12B it	64	45.16	35.78	+9.38	40.09	43.14	-3.05	7.04 (44)
	128	49.39	39.42	+9.97	47.77	43.65	+4.12	9.12 (57)
	256	50.21	41.62	+8.58	43.21	43.98	-0.77	8.16 (51)
Gemma4 26B-A4B it	64	49.45	44.60	+4.85	48.35	50.00	-1.65	11.36 (71)
	128	54.35	46.42	+7.93	52.54	51.29	+1.25	12.16 (76)
	256	49.37	49.84	-0.47	53.48	44.66	+8.82	14.72 (92)
Gemma4 31B it	64	65.34	64.10	+1.24	64.77	60.81	+3.97	24.80 (155)
	128	67.22	61.83	+5.39	66.67	65.65	+1.02	26.08 (163)
	256	70.84	68.10	+2.75	73.25	67.14	+6.11	35.04 (219)

with 256 frames than with 64 frames. However, four models show small decreases from 128 to 256 frames, indicating that additional frames do not always improve joint success.

The conditional accuracy gaps follow a different pattern. From 64 to 256 frames, the mean $\Delta_{P \to R}$ decreases from 3.57 to 1.75 percentage points, while the mean $\Delta_{P \to S}$ decreases from 4.12 to 2.91 points. For example, Qwen3VL 8B Instruct improves its All3 from 15.52% to 20.64%, while its reasoning gap changes from +9.95 to -0.16 points. Higher joint success therefore does not neces-

sarily coincide with larger perception-conditioned gaps. All3 measures joint correctness, whereas the gaps describe differences in task accuracy between perception-correct and perception-incorrect groups.

A.8 SAMPLE

Fig. 7 shows an example QA group from AnchorGroup.



Figure 7: **An example QA group from AnchorGroup.** All three questions concern the girl in the pink knit hat. (a) Perception identifies her action during the carnival game. (b) Reasoning infers her emotional reaction after the game. (c) Summarization covers key events involving her across the video. Reasoning and summarization share evidence from the carnival-game event. Correct answers are highlighted in green.