BEYOND SINGLE VIDEOS: BENCHMARKING AND ACTIVE EVIDENCE SEEKING FOR E-COMMERCE CROSS-VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

E-commerce videos are information-dense and frequently compared in practice, as consumers compare products and merchants evaluate marketing tactics, making cross-video analysis highly valuable for real-world business analytics. Yet, existing multimodal models primarily focus on single video understanding, with limited ability to perform the cross-video comparisons needed for e-commerce video analysis. To bridge this gap, we introduce **AdsCVR**, the first e-commerce cross-video reasoning benchmark. After filtering, AdsCVR contains 2,483 videos and 6,110 QA pairs across six reasoning dimensions. Extending reasoning across multiple videos fundamentally magnifies the challenge of pinpointing fine-grained evidence amid massive redundant frames, especially in e-commerce videos where visual details, speech, and on-screen text are combined into densely packed content. We propose **AdSeek**, an agentic framework that dynamically allocates visual and audio tools during multi-turn exploration, replacing static uniform sampling with active evidence grounding. While reinforcement learning (RL) can optimize sequential tool use, sparse outcome rewards struggle with long-horizon credit assignment, providing limited guidance on where a trajectory goes wrong or how to correct it. To address this, we propose an **offline trajectory rectification mechanism** that identifies and corrects reasoning errors and missing multimodal evidence in RL-generated trajectories. It then uses them for SFT to strengthen the model's evidence gathering and reasoning abilities, thereby mitigating biases learned during RL. This mechanism anchors our **rectified bootstrapping pipeline**: an initial RL phase exposes reasoning bottlenecks, the subsequent SFT phase uses rectified data to calibrate model's thinking and action biases, and a final RL phase achieves optimal convergence. AdSeek achieves 74.30% accuracy on the AdsCVR test split, improving upon its backbone (Qwen3-VL-8B-Instruct) by 27.90%. It also generalizes well to the open-domain CrossVid benchmark, validating its effectiveness in active evidence gathering.

1 INTRODUCTION

Many similar products on e-commerce platforms are promoted through different videos, presenting different marketing strategies. These videos often pack product demonstrations, selling points, and promotional messages into just tens of seconds through visuals, speech, and on-screen text. Consumers and merchants often need to compare this dense multimodal information across videos to inform purchasing and marketing decisions. However, existing advertising benchmarks such as MM-AU, AdsQA, and E-VAds (Bose et al., 2023; Long et al., 2025; Liu et al., 2026) predominantly focus on single-video understanding and do not adequately capture this practical need or the challenges of identifying key differences across videos.

Therefore, we propose **AdsCVR**, the first e-commerce cross-video reasoning benchmark, evaluating models' ability to understand complex e-commerce videos and draw effective comparisons across multiple videos. Unlike general-domain CVR benchmarks such as CrossVid and MVU-Eval (Li et al., 2026; Peng et al., 2025), AdsCVR focuses on commercial reasoning over advertisements, where decisive evidence is scattered across videos, non-contiguous temporal intervals, and visual and audio modalities amid distracting dense content.

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.

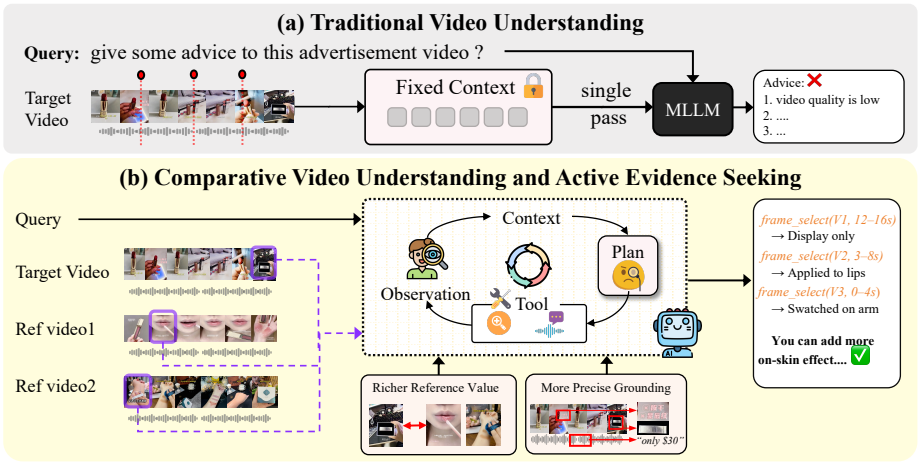

Figure 1: **Single-video understanding versus active cross-video understanding.** (a) Fixed-context, single-pass analysis may produce generic advice. (b) AdSeek leverages richer cross-video references and produce more precise evidence grounding for more specific and actionable advice.

To improve models' ability to understand and compare e-commerce videos, we propose **AdSeek**, an agentic framework for multi-turn active evidence acquisition. Unlike static uniform sampling (Wu et al., 2024), AdSeek dynamically selects which videos, temporal segments, modalities, and visual regions to inspect based on the comparison objective. It filters out irrelevant content and aligns evidence across videos before generating an answer.

Following tool-augmented language modeling (Yao et al., 2023; Schick et al., 2023; Yang et al., 2023; Hao et al., 2023), we build a multi-turn interaction environment for AdSeek and train its evidence-search policy through reinforcement learning (RL) (Qian et al., 2025). During RL training process, sparse outcome rewards struggle with long-horizon credit assignment (Arjona-Medina et al., 2019), failing to provide explicit feedback for intermediate tool calls or reasoning decisions. So, error patterns that emerge during exploration may persist even with continued RL training.

To address this challenge, we introduce an **offline trajectory rectification mechanism** . It uses a teacher model to identify intermediate deviations that remain undetected by outcome-based RL feedback and rectify them by supplying missing evidence and correcting faulty reasoning. The corrected trajectories directly supervise intermediate evidence acquisition and reasoning, helping the policy model internalize these corrections via SFT. Building on this, our **Rectified Bootstrapping Pipeline (RBP)** uses initial RL exploration to expose bottlenecks, offline trajectory rectification to correct persistent biases, and subsequent RL to further refine the policy through online exploration.

Our contributions include:

- We introduce **AdsCVR**, an e-commerce cross-video reasoning benchmark comprising 2,483 videos and 6,110 QA pairs spanning 86 product categories (see Table 1 for a comparison). It evaluates models' understanding of complex commercial analysis and their ability to identify fine-grained differences across information-dense videos.
- We propose **AdSeek**, an agentic framework that actively acquires and aligns multimodal evidence through multi-turn reasoning and tool use, adapting its search across videos, temporal segments, modalities, and visual regions to each comparison objective.
- We develop the **Rectified Bootstrapping Pipeline (RBP)**, which combines online RL exploration with offline trajectory rectification. By correcting biased trajectories and enabling further exploration, this training paradigm strengthens the model's ability to acquire multimodal evidence.
- AdSeek yields a 74.30% accuracy on the AdsCVR test split, outperforming the Qwen3-VL-8B-Instruct baseline by an absolute margin of 27.90%. Beyond e-commerce, it exhibits robust transferability on the open-domain CrossVid dataset, proving its broader capacity for proactive multimodal evidence collection.

2 Related Work

Video Understanding Benchmarks

Advances in MLLMs have driven video QA evaluations (Lin et al., 2024; Li et al., 2024) from single-video content and causal reasoning (Yu et al., 2019; Xiao et al., 2021) to long-context understanding (Fu et al., 2025; Zhou et al., 2025; Wu et al., 2024; Cheng et al., 2025) and egocentric tasks (Huang et al., 2024). Beyond single videos, benchmarks like CrossVid (Li et al., 2026), MVU-Eval (Peng et al., 2025), and CVBench (Zhu et al., 2025) evaluate joint understanding of video collections, requiring models to aggregate distributed evidence and compare entities across videos.

E-commerce Advertisement Video Understanding

Advertisement understanding requires connecting audiovisual content with persuasive intent. Early works focused on message symbolism (Hussain et al., 2017), temporal segmentation, and classification (Wang et al., 2021; Bose et al., 2023). Recent QA benchmarks further evaluate visual reasoning in fast-paced ads (Zhang et al., 2025), core messages and target audiences (Long et al., 2025), and conversion-oriented e-commerce elements (Liu et al., 2026). Crucially, these benchmarks evaluate videos individually. In contrast, AdsCVR connects advertisement understanding with cross-video reasoning, requiring models to align cross-ad evidence to compare product presentations.

Table 1: **Comparison of video understanding benchmarks.** Temporal and Intent denote explicit temporal and intent understanding coverage (✓); —: not explicitly targeted. MCQ: multiple-choice QA; Open: open-ended QA; Seg.: segmentation; Cls.: classification.

Benchmark	Domain	#Videos	#QA	Video Count	Task format	Temporal	Intent
MVU-Eval (Peng et al., 2025)	General	4,959	1,824	Multiple	MCQ	✓	—
CrossVid (Li et al., 2026)	General	5,331	9,015	Multiple	MCQ / Open	✓	—
Tencent-AVS (Wang et al., 2021)	Ads	12,000	N/A	Single	Seg. + Cls.	✓	—
MM-AU (Bose et al., 2023)	Ads	8,399	N/A	Single	Cls.	✓	✓
VideoAds (Zhang et al., 2025)	Ads	200	1,100	Single	MCQ	✓	✓
AdsQA (Long et al., 2025)	Ads	1,544	7,859	Single	Open	—	✓
E-VAds (Liu et al., 2026)	E-commerce	3,961	19,785	Single	Open	—	✓
AdsCVR	E-commerce	**2,483**	**6,110**	Multiple	MCQ	✓	✓

MLLMs for Video Understanding

Video MLLMs acquire temporal reasoning by training visual encoders and language models on multimodal instructions (Lin et al., 2024; Li et al., 2024; Zhang et al., 2024; Cheng et al., 2024; Wang et al., 2024b). However, early systems rely on uniform frame sampling, risking the omission of brief, decisive events. To address this, recent agentic and tool-using approaches make perception adaptive, iteratively retrieving or selecting keyframes based on query relevance and reasoning feedback (Wang et al., 2024a; Jeoung et al., 2024; Wang et al., 2025b; He et al., 2026; Zhang et al., 2026). AgentCVR extends this active acquisition to cross-video reasoning via script-simulated RL (Qiu et al., 2026). Together, these approaches shift video understanding from fixed frame sampling toward adaptive evidence acquisition guided by the question and intermediate reasoning.

3 The AdsCVR Benchmark

We introduce the AdsCVR benchmark to evaluate fine-grained e-commerce video understanding. We define a *comparable video group* as a set of videos sharing contextual dimensions (e.g., narrative subject or commercial intent) to expose subtle distinctions in advertising strategies. Figure 2 illustrates our three-stage construction pipeline.

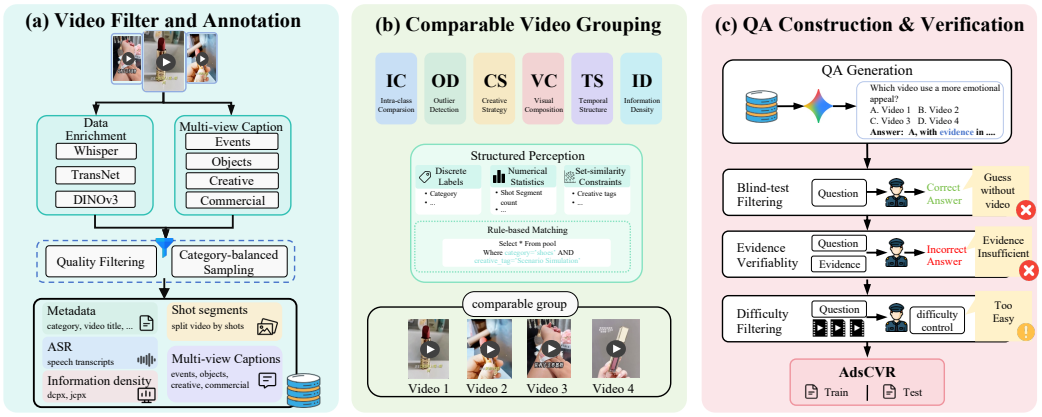

Figure 2: **AdsCVR construction pipeline.** (a) **Video Annotation:** Raw videos undergo multimodal feature extraction and multidimensional captioning, followed by quality and category-balanced filtering. (b) **Comparable Grouping:** Structured rules guide the assembly of 2–4 videos into comparable groups across six reasoning dimensions. (c) **QA Generation & Filtering:** LLMs generate multiple-choice questions for each group. A rigorous multi-stage filtering pipeline (blind-test, verifiability, and difficulty) yields 6,110 high-quality QA pairs (3,110 train, 3,000 test).

3.1 Video Filtering and Multi-View Annotation

Video filtering. To construct a diverse benchmark of manageable size, we first preprocess the videos to support subsequent filtering: We extract video metadata, speech transcripts (Whisper-large-v3), and shot boundaries (TransNetV2). To filter out outliers, we estimate temporal and spatial complexities using DINOv3 embeddings and JPEG compression (see Appendix C for metric details). Based on these measures, we discard videos with extreme durations or low visual variation. Finally, we balance the dataset by dropping long-tail categories (fewer than ten videos) and downsampling dominant ones, yielding 2,483 videos across 86 product categories (Figure 3a).

Multi-view annotation. To go beyond sparse metadata and capture fine-grained semantics, we prompt Gemini to generate dense, multidimensional captions from four expert perspectives: *events* (actions and demonstrations), *objects* (products and scenes), *creative structure* (narratives and filming techniques), and *commercial logic* (audiences and advertising strategies; see Figure 3b). The full multi-view annotation prompt is provided in Appendix A.2. These annotations provide the semantic foundation for the subsequent video grouping and QA generation.

3.2 Video Grouping

We form comparable groups of two to four videos using dimension-specific sampling rules (Table 2), covering visual execution and commercial intent.

Table 2: AdsCVR comparison dimensions and video-grouping rules.

Dimension	Comparison objective and sampling rule
Intra-class Comparison (IC)	Fine-grained discrimination: sample unique video combinations within the same first-level category.
Outlier Detection (OD)	Anomaly identification: combine $N-1$ reference videos with one outlier from a different second-level or first-level category, with a 50%/50% allocation.
Creative Strategy (CS)	Commercial intent: include at least two distinct commercial-strategy buckets per group.
Visual Composition (VC)	Visual presentation: require creative-tag Jaccard similarity < 0.5 between candidate and seed videos.
Temporal Structure (TS)	Narrative progression: use segment count as a proxy and span at least two ranges (1–3, 4–6, ≥ 7 segments) per group.
Information Density (ID)	Visual pacing: sample intra-category videos whose absolute shot counts differ by ≥ 3, independent of duration.

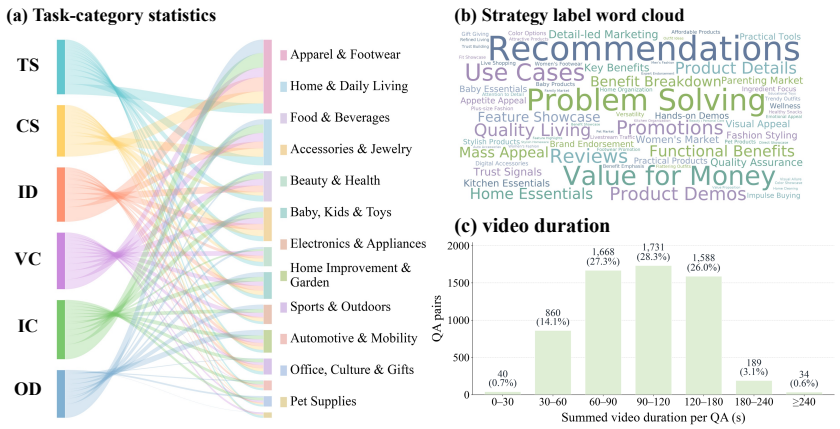

Figure 3: **AdsCVR dataset statistics.** (a) Associations between the six evaluation dimensions and 12 representative product categories, with flow width indicating video count. (b) Commercial-strategy word cloud. (c) Total video duration per QA pair across 6,110 examples.

3.3 QA Generation and Verification

To construct informative and challenging multiple-choice QA pairs, we employ prompt-based generation followed by a rigorous multistage filtering pipeline. First, for **QA generation**, Gemini-3.5-Flash generates a single-choice question, options, the correct answer, and supporting textual clues based on aggregated video metadata, multidimensional captions, and dimension-specific grouping rationales. Next, we apply **blind-test filtering** to mitigate language-prior leakage, discarding any QA pair that Qwen3VL-8B answers correctly in three consecutive attempts using exclusively the text question and options (without video inputs). To ensure **answer verifiability** and grounding in video evidence, we then require GPT-5.5 to correctly predict the reference answer relying solely on the generated textual clues. Finally, for **difficulty filtering**, we evaluate the remaining pairs using Qwen3-VL-8B-Instruct with uniformly sampled frames, randomly discarding 50% of the questions it answers correctly to calibrate the overall benchmark difficulty.

AdsCVR contains 6,110 QA pairs from 2,483 e-commerce videos across 86 top-level product categories, with QA pairs randomly split into 3,110 training and 3,000 test examples.

4 AdSeek

We introduce **AdSeek**, a framework for active multi-turn evidence acquisition trained through the Rectified Bootstrapping Pipeline, as illustrated in Figure 4.

4.1 Task Definition

Given a query q and a video collection $\mathcal{V} = \{V_j\}_{j=1}^{N}$, the task is to generate an answer $\hat{y}$ by iteratively acquiring cross-video evidence via perception tools.

Interaction process. The policy Φ initializes with context $h_1 = (q, \mathcal{M})$, where $\mathcal{M}$ contains lightweight metadata (e.g., duration, resolution). At round i, the context expands to:

$$h_i = (q, \mathcal{M}, t_1, a_1, o_1, \ldots, t_{i-1}, a_{i-1}, o_{i-1}), \tag{1}$$

where t_k, a_k, and o_k denote the reasoning trace, tool call, and returned textual observation at round k. Conditioned on h_i, Φ generates t_i followed by either a final answer $\hat{y}$ or a tool call $a_i \in \mathcal{A}$. Executing a_i yields observation o_i, updating the context as:

$$h_{i+1} = (h_i, t_i, a_i, o_i). \tag{2}$$

Perception tools. The action space comprises visual and audio tools:

$$\begin{aligned} \mathcal{A} &= \mathcal{A}_{\text{vis}} \cup \mathcal{A}_{\text{aud}}, \\ \mathcal{A}_{\text{vis}} &= \{\texttt{frame_select}(j, s, e, f)\}, \\ \mathcal{A}_{\text{aud}} &= \{\texttt{audio_select}(j, s, e)\}, \end{aligned} \tag{3}$$

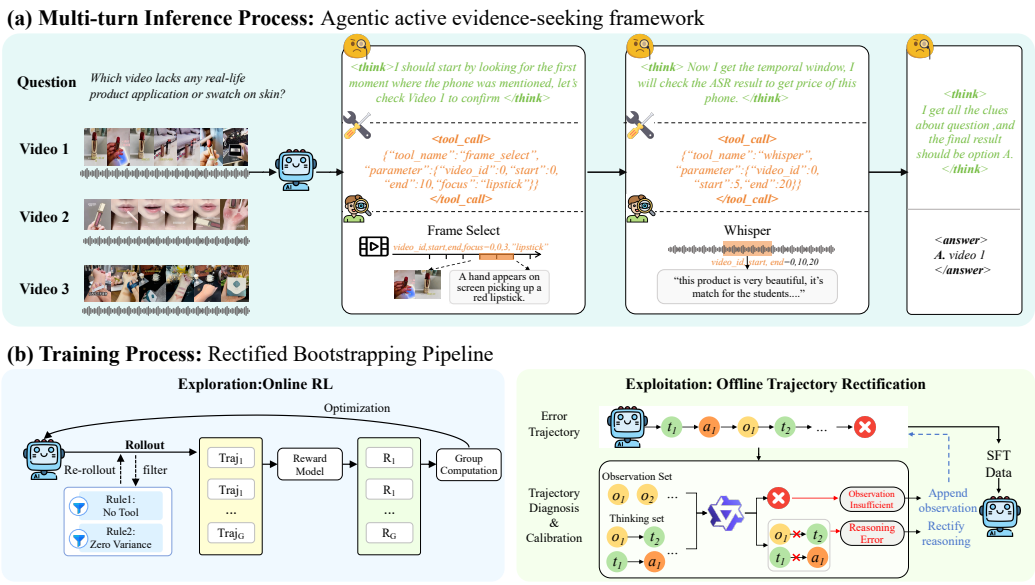

Figure 4: **AdSeek overview.** (a) AdSeek's multi-turn inference process, where reasoning guides tool use to actively acquire relevant evidence. (b) The two modules of the Rectified Bootstrapping Pipeline: online RL exploration with trajectory filtering, and offline teacher-guided trajectory rectification, which corrects observation and reasoning errors and provides targeted SFT supervision for subsequent exploration.

where $j \in \{1, \ldots, N\}$ identifies the video and $0 \le s < e \le D_j$ specifies a temporal interval. The prompt $f \in \mathcal{F}$ defines specific visual details to inspect. The visual and audio tools return a focused textual description and a speech transcript of the segment, respectively.

Trajectory. This process repeats until Φ produces an answer or hits the interaction limit, yielding a completed n-round trajectory:

$$\tau = (q, \mathcal{M}, t_1, a_1, o_1, \ldots, t_{n-1}, a_{n-1}, o_{n-1}, t_n, \hat{y}). \tag{4}$$

4.2 Exploration: Online Reinforcement Learning (RL)

4.2.1 Policy Optimization

We optimize the policy using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), implemented with ROLL (Wang et al., 2025a). For each input x, the policy samples G trajectories $\{\tau_i\}_{i=1}^{G}$. Each trajectory receives a reward combining final-answer correctness (w.r.t. the reference $y^\star$) and action-format compliance:

$$R(\tau_i) = R_{\text{acc}}(\hat{y}_i, y^\star) + 0.1 R_{\text{fmt}}(\tau_i). \tag{5}$$

We compute the group-normalized advantage as $\widehat{A}_i = (R(\tau_i) - \mu_R)/(\sigma_R + \epsilon_A)$, where μ_R and σ_R are the group reward mean and standard deviation. For a generated token $z_{i,k}$ with preceding context $c_{i,k}$, we define the probability ratio $r_{i,k}(\theta) = \pi_\theta(z_{i,k} \mid c_{i,k}) / \pi_{\theta_{\text{old}}}(z_{i,k} \mid c_{i,k})$. The token-level GRPO objective is then:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{i,k}\Big[& \min\left(r_{i,k}(\theta)\widehat{A}_i, \text{clip}(r_{i,k}(\theta), 1-\epsilon, 1+\epsilon)\widehat{A}_i\right) \\ & - \beta D_{\text{KL}}\left(\pi_\theta(\cdot \mid c_{i,k}) \| \pi_{\text{ref}}(\cdot \mid c_{i,k})\right)\Big], \end{aligned} \tag{6}$$

where ϵ controls policy-ratio clipping and β weights the KL penalty.

4.2.2 Rollout Filtering and Resampling

We apply two filtering rules to encourage evidence acquisition and retain informative reward comparisons: (1) **No-tool trajectories.** We discard trajectories with no effective tool calls, $N_{\text{action}} = 0$.

Since video content is accessible only through perception tools, these trajectories provide no evidence of video inspection. (2) **Zero-variance groups.** We discard groups $\mathcal{G}$ with $\mathrm{Var}(\{R(\tau) : \tau \in \mathcal{G}\}) = 0$. Such groups yield zero normalized advantages and therefore no reward-driven policy-gradient signal. They can arise when all trajectories receive the same correctness and format rewards. After filtering, we collect additional rollouts to replenish the training batch, allowing up to three (`MAX_TRY`=3) resampling attempts. Implementation details are provided in Appendix B.

4.3 Exploitation: Offline Trajectory Rectification

Outcome rewards provide limited guidance on intermediate failures (Lightman et al., 2024). Building on feedback-guided refinement (Madaan et al., 2023; Shinn et al., 2023; Havrilla et al., 2024), we introduce offline trajectory rectification to diagnose and correct deviations in the policy's exploration. This diagnosis targets two core issues: evidence sufficiency and reasoning validity.

Candidate selection. We select failed training trajectories from the RL-trained policy where a stronger teacher (Φ_{teacher}) succeeds in the same environment. The teacher's successful trajectory serves as a reference for identifying evidence gaps and guiding corrections.

Observation sufficiency check. For a trajectory with observations $\mathcal{O}_\tau = \{o_1, \ldots, o_{n-1}\}$, the teacher attempts to answer the query using only $\mathcal{O}_\tau$ (without tool calls). An incorrect answer acts as a diagnostic signal indicating a potential evidence gap (e.g., missing videos, intervals, or modalities).

Reasoning validity check. The teacher reviews each reasoning step against three criteria: (1) **Factual consistency** (claims must be grounded in observations, avoiding hallucinations), (2) **Reasoning completeness** (conclusions require sufficient evidence support), and (3) **Historical consistency** (alignment with prior rounds). Steps are flagged if any criterion's risk score (1–5) exceeds 3.

Targeted trajectory rectification. We apply corrections based on the diagnosis: (1) **Evidence omission:** We append missing observations from the teacher's reference trajectory to the policy's context, prompting Φ to regenerate intermediate reasoning. (2) **Reasoning deviation:** The teacher provides corrective comments on flagged steps to guide Φ's reasoning regeneration. By targeting specific failures within the student's own context (rather than replacing it entirely with the teacher's rollout), we produce tailored corrective demonstrations.

4.3.1 Supervised Policy Optimization

Following rationale bootstrapping (Zelikman et al., 2022), we construct $\mathcal{D}_{\text{SFT}}$ from initially successful and newly rectified trajectories, helping the policy internalize corrected evidence acquisition and reasoning.

For a tokenized trajectory $z_{1:M}$, a mask $m_j \in \{0, 1\}$ isolates assistant tokens (reasoning traces, tool calls, and final answers, where $m_j = 1$) from the context (queries, metadata, and observations, where $m_j = 0$). The objective is:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{\tau \sim \mathcal{D}_{\text{SFT}}} \left[\frac{\sum_{j=1}^{M} m_j \log \pi_\theta(z_j \mid z_{<j})}{\sum_{j=1}^{M} m_j} \right]. \tag{7}$$

The loss is averaged solely over assistant tokens. This applies direct supervision to intermediate steps on fixed trajectories, without invoking perception tools during training.

4.4 Training Paradigm: Rectified Bootstrapping Pipeline

We integrate these components into the *Rectified Bootstrapping Pipeline* (RBP), a three-stage paradigm: (1) **Initial RL** develops evidence acquisition and exposes trajectory errors; (2) **Offline Rectification & SFT** corrects missing evidence and faulty reasoning via teacher guidance, internalized through SFT; (3) **Refined RL** resumes online optimization. By complementing sparse outcome feedback with direct supervision of intermediate decisions, RBP strengthens the model's ability to locate, integrate, and compare evidence scattered across videos, temporal segments, and modalities, supporting a deeper understanding of information-dense e-commerce videos.

5 EXPERIMENTS

TRAINING AND EVALUATION SETTINGS

Our policy backbone is Qwen3-VL-8B-Instruct (Bai et al., 2025), using Qwen3-VL-4B-Instruct and Whisper large-v3 (Radford et al., 2023) for perception, and Qwen3.8-27B for offline attribution and evaluation. We use 16 PPUs for SFT via LLaMA-Factory (Zheng et al., 2024) and 32 PPUs for RL via ROLL (Wang et al., 2025a) (Appendix B). We evaluate accuracy on the AdsCVR test split (3,000 test / 3,110 train from 6,110 QA pairs) and CrossVid for out-of-domain generalization. Baselines include general MLLMs and AgentCVR (Qiu et al., 2026)[1], a prior cross-video RL framework.

5.1 IN-DOMAIN RESULTS ON ADSCVR

Table 3 highlights the difficulty of AdsCVR: evaluated open-source single-pass models achieve average accuracies of 46.40%–75.15%, leaving substantial room for improvement. Trained only on AdsCVR, AdSeek reaches 73.57%, outperforming its Qwen3-VL-8B backbone by 27.17 percentage points and our AgentCVR reproduction by 2.52 points, while approaching Qwen3.8-27B (75.15%). Incorporating public out-of-domain training data further improves accuracy to 74.30%; we analyze mixed-source training in Section 5.3.

Average accuracy improves across training stages, with the largest stage-to-stage gain occurring after RL resumes (+2.04 points). Relative to Stage 1, the final mixed-source model improves across all six dimensions, most notably Information Density (ID, +5.0 points) and Visual Composition (VC, +4.8 points), where answering requires locating fine-grained evidence in dense visual content or comparing details across shots. Together, these results suggest that AdSeek's SFT stage preserves overall performance while enabling further improvement during subsequent RL, and that targeted trajectory rectification contributes to higher accuracy on challenging tasks involving dense visual content and cross-shot comparisons.

Table 3: Accuracy (%) on the AdsCVR test split. All AdSeek variants use Qwen3-VL-8B. Bold indicates our best results. Average denotes the unweighted mean across dimensions.

Method	TS	CS	ID	IC	OD	VC	Average
General-purpose MLLM baselines							
Gemini 3.5 Flash	79.3	75.0	76.2	75.6	78.8	81.5	77.73
GPT-5.5	85.2	80.2	84.3	79.6	82.0	83.1	82.40
Qwen3.8-27B	79.7	74.3	74.9	77.1	72.2	72.7	75.15
Qwen3.5-4B	64.8	59.0	56.6	61.0	60.6	61.9	60.65
Qwen3.5-9B	68.9	64.6	67.1	64.0	64.9	64.5	65.67
Prior agentic baseline							
AgentCVR (2026)	73.7	71.1	68.7	70.3	73.6	68.9	71.05
AdSeek with Qwen3-VL-8B backbone							
Qwen3-VL-8B (backbone)	52.2	46.8	41.6	43.6	48.2	46.0	46.40
AdSeek (Stage 1)	74.6	72.4	65.8	72.8	72.6	69.4	71.27
AdSeek (Stage 2)	75.0	67.6	68.2	73.6	74.4	70.4	71.53
AdSeek (Stage 3, in-domain only)	75.0	74.4	70.6	72.8	74.6	74.0	73.57
AdSeek (Stage 3, mixed)	**77.0**	**73.4**	**70.8**	**75.6**	**74.8**	**74.2**	**74.30**
Gain vs. backbone (pp)	+24.8	+26.6	+29.2	+32.0	+26.6	+28.2	+27.90

5.2 OUT-OF-DOMAIN RESULTS ON CROSSVID

To assess cross-domain generalization, we evaluate AdSeek on ten tasks from CrossVid, a general-domain cross-video benchmark. As shown in Table 4, AdSeek achieves an average accuracy of 31.55%, exceeding its Qwen3-VL-8B backbone by 2.08 percentage points and our AgentCVR reproduction by 3.31 points. It improves over the backbone on six of the ten tasks. These results suggest that AdSeek's evidence-seeking capabilities transfer beyond e-commerce videos.

[1]Reproduced in our own environment using the same infrastructure, based on the code released by the authors.

Table 4: Out-of-domain performance (%) on CrossVid. Average is the unweighted mean over ten tasks; means and gains are computed before display rounding. Bold marks the highest average.

Method	BU	NC	CC	PEA	PI	FSA	MSR	MOC	PSS	CCQA	Average
Qwen3-VL-8B	23.5	27.4	40.7	34.9	62.5	5.71	34.5	27.2	7.22	31.2	29.47
AgentCVR (2026)	27.6	37.5	48.9	30.1	50.2	11.0	22.3	28.4	8.14	18.2	28.24
AdSeek (Stage 3, mixed)	38.9	38.7	47.7	37.5	45.4	16.1	26.4	37.8	6.61	20.4	**31.55**

5.3 Ablations on Training Data and Action Space

Training data. Table 5 shows that AdsCVR-only training outperforms MVU-Eval-only training by 9.50 percentage points (73.57% vs. 64.07%), supporting the value of domain-specific data for e-commerce video comparison. Our mixed-source dataset, AdSeek-train, supplements AdsCVR with MVU-Eval videos and script-simulated AgentCVR examples, further improving accuracy to 74.30%. The additional gains are concentrated in IC (+2.8 points) and TS (+2.0 points), while CS decreases by 1.0 point. Thus, mixed-source training provides a modest overall benefit, with effects that vary across comparison dimensions.

Action space. Restricting AdSeek to `frame_select` or `audio_select` reduces average accuracy to 63.97% and 65.03%, respectively, corresponding to drops of 10.33 and 9.27 points from the full model. The two modalities exhibit different strengths: visual-only access performs better on IC (63.8% vs. 58.2%), whereas speech-only access performs better on the other five dimensions. Nevertheless, the full model outperforms both restricted variants across all six dimensions, supporting the benefit of combining visual and speech evidence for cross-video comparison.

Table 5: Training-data and action-space ablations on AdsCVR test accuracy (%). Average denotes the unweighted mean across dimensions. MVU-Eval-only and restricted-action variants are evaluated on the 3,000-question test split, with unparsed answers counted as incorrect.

Variant	TS	CS	ID	IC	OD	VC	Average
Training data							
In-domain only (AdsCVR)	75.0	74.4	70.6	72.8	74.6	74.0	73.57
MVU-Eval only	67.4	60.8	60.8	63.6	65.6	66.2	64.07
Mixed	77.0	73.4	70.8	75.6	74.8	74.2	74.30
Action space							
`frame_select` only	66.2	63.8	61.2	63.8	65.8	63.0	63.97
`audio_select` only	67.4	65.6	65.6	58.2	66.2	67.2	65.03
AdSeek	**77.0**	**73.4**	**70.8**	**75.6**	**74.8**	**74.2**	**74.30**

6 Limitations

AdSeek involves computational trade-offs associated with multi-turn agentic reasoning. Sequential tool calls and repeated context processing can increase inference latency, while RL rollouts and teacher-guided trajectory rectification introduce additional training overhead. Future work could explore more efficient tool scheduling and lightweight perception models to better balance reasoning quality and computational cost, particularly for latency-sensitive applications.

7 Conclusion

We introduced AdsCVR, a comprehensive benchmark for fine-grained cross-video e-commerce comparison, alongside AdSeek, an agentic framework for active evidence acquisition. To optimize this framework, we developed the Rectified Bootstrapping Pipeline (RBP), which seamlessly integrates RL exploration with offline trajectory rectification via SFT. AdSeek achieves 74.30% accuracy on AdsCVR and demonstrates robust transferability to open-domain multimodal reasoning tasks.

AI USE STATEMENT

We used generative AI models as components of the reported research workflow. Gemini models were used to produce multi-view video annotations and candidate AdsCVR questions, answer options, and supporting clues. Qwen3-VL-8B was used for prompt-only and frame-based filtering, GPT-5.5 was used for evidence-based answer verification, and Qwen3.8-27B was used as the teacher for offline trajectory diagnosis and as an evaluation judge where described in the paper. These model outputs were processed by deterministic filters and checked on representative examples by the authors; model-generated outputs were not treated as ground truth without the stated verification steps. The authors reviewed the final manuscript, experimental claims, and released code, and take responsibility for the accuracy, integrity, and originality of all submitted material.

REFERENCES

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return Decomposition for Delayed Rewards. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1806.07857.

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. URL https://arxiv.org/abs/2511.21631.

Digbalay Bose, Rajat Hebbar, Tiantian Feng, Krishna Somandepalli, Anfeng Xu, and Shrikanth Narayanan. MM-AU: Towards Multimodal Understanding of Advertisement Videos. In *Proceedings of the ACM International Conference on Multimedia*, pp. 86–95, 2023. URL https://arxiv.org/abs/2308.14052.

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. URL https://arxiv.org/abs/2505.21374.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing Spatiotemporal Modeling and Audio Understanding in Video LLMs. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. ToolkenGPT: Augmenting Frozen Language Models with Massive Tools via Tool Embeddings. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45870–45894, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8fd1a81c882cd45f64958da6284f4a3f-Abstract-Conference.html.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. GLoRe: When, Where, and How to Improve LLM Reasoning via Global and Local Refinements. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/havrilla24a.html.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. FrameThinker: Learning to Think with Long Videos via Multi-Turn Frame Spotlighting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nsNpsCpVG1.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. EgoExoLearn: A Dataset for Bridging Asynchronous Ego- and Exo-Centric View of Procedural Activities in Real World. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22072–22086, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_EgoExoLearn_A_Dataset_for_Bridging_Asynchronous_Ego-_and_Exo-centric_View_CVPR_2024_paper.html.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic Understanding of Image and Video Advertisements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1705–1715, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Hussain_Automatic_Understanding_of_CVPR_2017_paper.html.

Sullam Jeoung, Goeric Huybrechts, Bhavana Ganesh, Aram Galstyan, and Sravan Bodapati. Adaptive Video Understanding Agent: Enhancing Efficiency with Dynamic Frame Sampling and Feedback-Driven Reasoning. *arXiv preprint arXiv:2410.20252*, 2024. URL https://arxiv.org/abs/2410.20252.

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. doi: 10.1609/aaai.v40i8.37550.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=A7x2z0qFUD.

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. doi: 10.18653/v1/2024.emnlp-main.342.

Xianjie Liu, Yiman Hu, Liang Wu, Ping Hu, Yixiong Zou, Jian Xu, and Bo Zheng. E-VAds: An E-Commerce Short Video Understanding Benchmark for MLLMs. *arXiv preprint arXiv:2602.08355*, 2026. URL https://arxiv.org/abs/2602.08355.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. AdsQA: Towards Advertisement Video Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23396–23407, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Long_AdsQA_Towards_Advertisement_Video_Understanding_ICCV_2025_paper.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Tianhao Peng, Haochen Wang, Yuanxing Zhang, Noah Wang, Zili Wang, Ge Zhang, Jian Yang, Shihao Li, Yanghai Wang, Xintao Wang, et al. MVU-Eval: Towards Multi-Video Understanding Evaluation for Multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1800.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tur, Gokhan Tur, and Heng Ji. ToolRL: Reward Is All Tool Learning Needs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 105523–105553, 2025. doi: 10.52202/085713-3524.

Yilun Qiu, Jiahe Wang, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Chun Yuan. AgentCVR: Active Multi-Agent Cross-Video Reasoning via Script-Simulated Reinforcement Learning. *arXiv preprint arXiv:2605.29643*, 2026. URL https://arxiv.org/abs/2605.29643.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023. URL https://proceedings.neurips.cc/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

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. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Weixun Wang, Shaopan Xiong, Gengru Chen, Wei Gao, Sheng Guo, Yancheng He, Ju Huang, Jiaheng Liu, Zhendong Li, Xiaoyang Li, et al. Reinforcement Learning Optimization for Large-Scale Learning: An Efficient and User-Friendly Scaling Library. *arXiv preprint arXiv:2506.06122*, 2025a. URL https://arxiv.org/abs/2506.06122.

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, 2024a. doi: 10.1007/978-3-031-72989-8_4.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. In *European Conference on Computer Vision*, pp. 396–416, 2024b. URL https://arxiv.org/abs/2403.15377.

Zhenzhi Wang, Liyu Wu, Zhimin Li, Jiangfeng Xiong, and Qinglin Lu. Overview of Tencent Multi-Modal Ads Video Understanding Challenge. *arXiv preprint arXiv:2109.07951*, 2021. URL https://arxiv.org/abs/2109.07951.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive Tree-Based Video Representation for LLM Reasoning on Long Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3282, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VideoTree_Adaptive_Tree-based_Video_Representation_for_LLM_Reasoning_on_Long_CVPR_2025_paper.html.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-Context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024. doi: 10.52202/079017-0907.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9772–9781, 2021. URL https://arxiv.org/abs/2105.08276.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. GPT4Tools: Teaching Large Language Models to Use Tools via Self-Instruction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71995–72007, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e393677793767624f2821cec8bdd02f1-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019. doi: 10.1609/aaai.v33i01.33019127.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping Reasoning with Reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022. URL https://papers.nips.cc/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Yaolun Zhang, Ruohui Wang, Jiahao Wang, Yepeng Tang, Xuanyu Zheng, Haonan Duan, Hao Lu, Hanming Deng, and Lewei Lu. EVA: Efficient Reinforcement Learning for End-to-End Video Agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12289–12299, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_EVA_Efficient_Reinforcement_Learning_for_End-to-End_Video_Agent_CVPR_2026_paper.html.

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. URL https://arxiv.org/abs/2410.02713.

Zheyuan Zhang, Wanying Dou, Linkai Peng, Hongyi Pan, Ulas Bagci, and Boqing Gong. VideoAds for Fast-Paced Video Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21812–21821, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_VideoAds_for_Fast-Paced_Video_Understanding_ICCV_2025_paper.html.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LLaMA-Factory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, 2024. doi: 10.18653/v1/2024.acl-demos.38.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking Multi-Task Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. URL https://arxiv.org/abs/2406.04264.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. CVBench: Benchmarking Cross-Video Synergies for Complex Multimodal Reasoning. *arXiv preprint arXiv:2508.19542*, 2025. URL https://arxiv.org/abs/2508.19542.

A Annotation Guidelines

AdsCVR annotations compare two to four e-commerce videos along six dimensions: TS, CS, ID, VC, IC, and OD. Data construction first applies category-balanced quality filtering and timestamped ASR/OCR extraction, then uses Gemini 3.5 Flash to generate structured descriptions and salient tags. Candidate questions are generated as (question, options, answer, clues) tuples. A hierarchical verifier uses Qwen3-VL-8B for prompt-only shortcut and difficulty checks, followed by GPT-5.5 checks for answer completeness, multi-video dependence, logical validity, and multimodal evidence requirements.

A.1 Qualitative Examples Across Dimensions

Figures 5–10 show selected cases for each of the six dimensions.

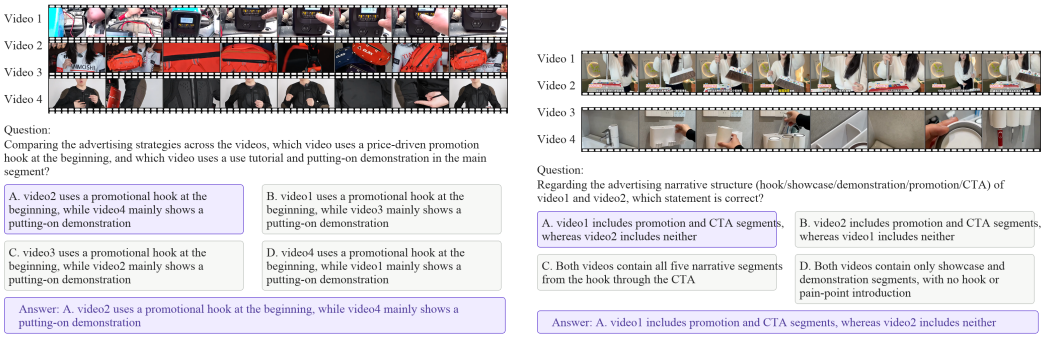

Figure 5: Temporal Structure (TS) examples comparing how advertisements organize events and product claims.

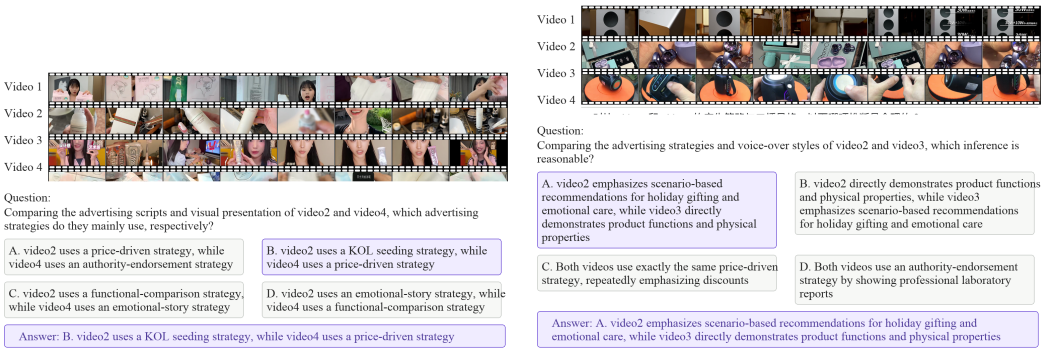

Figure 6: Creative Strategy (CS) examples comparing how advertisements attract attention and communicate product value.

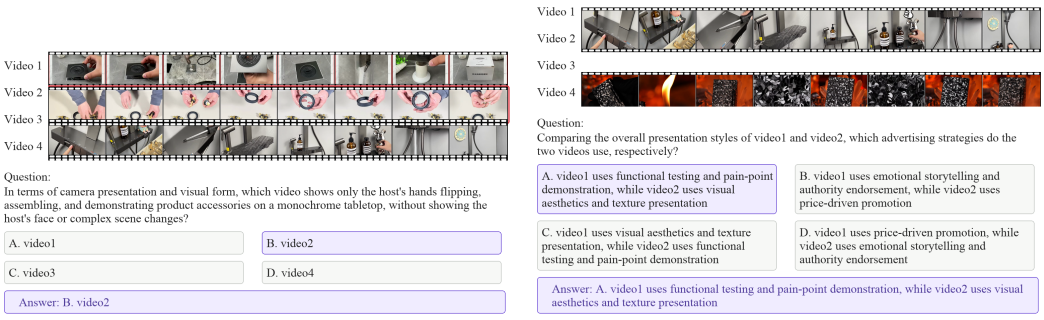

Figure 7: Outlier Detection (OD) examples comparing advertisements across product categories.

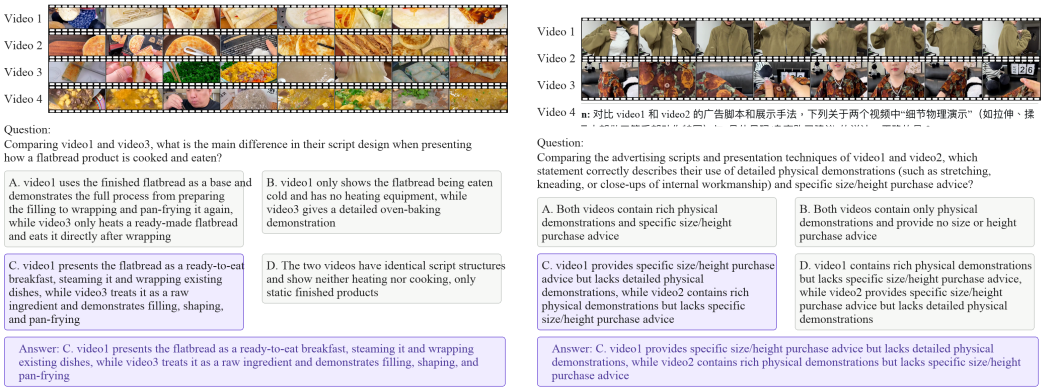

Figure 8: Intra-class Comparison (IC) examples comparing substantive similarities and differences among advertisements for related products.

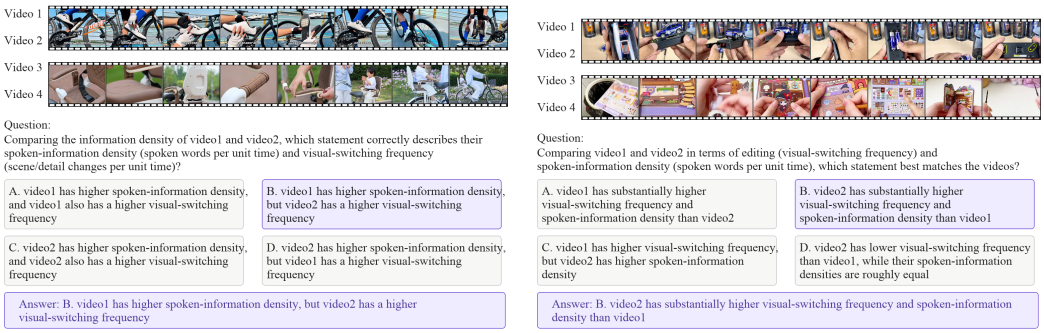

Figure 9: Information Density (ID) examples comparing the amount and pace of product-related spoken and visual information.

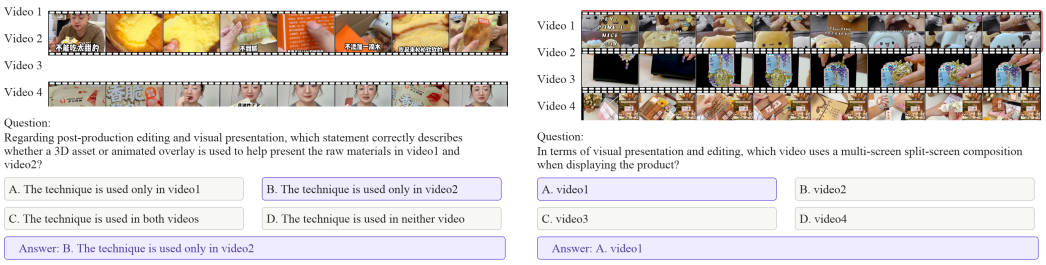

Figure 10: Visual Composition (VC) examples comparing framing, editing, transitions, and the functional role of individual shots.

A.2 MULTI-VIEW ANNOTATION PROMPT

The following prompt is used to generate the four-view annotations described in Section 3.3. The placeholders in the user prompt are filled with the corresponding metadata for each video.

System prompt.

```
SYSTEM_PROMPT_ALL = (
    "You are a multi-role video synthesis analyst. Analyze the advertising video
     in depth by acting as each of the following four roles.\n\n"

    "[General requirements (strictly followed by all roles)]\n"
    "- Ground every fact: describe only content clearly visible in the video or
     reasonably inferable; do not invent dialogue, background stories, or
     off-screen elements\n"
    "- Segmentation: prioritize the [shot time segmentation] in the auxiliary
     information. If it is unavailable, segment by action changes, shot
     transitions, or scene changes; uniform time segmentation is prohibited\n"
    "- Caption length for each segment: 20-50 words\n\n"

    "[Role 1: Event analyst (temporal-event view)]\n"
    "Describe the main events occurring in each shot:\n"
    "  Temporal dynamics: specific actions, motion trajectories, human
     interactions, object manipulation, and state changes\n"
    "  Spatial context: scene environment and positions of key objects; describe
     them only when they change\n\n"

    "[Role 2: Scene analyst (object-and-space view)]\n"
    "Describe the people, objects, and scene in each shot:\n"
    "  Spatial details: setting, spatial relations among objects (left/right,
     above/below, front/behind, near/far), people's positions and postures,
     color and material properties, and composition\n"
    "  Key objects: objects relevant to the video topic or the current frame\n\n"

    "[Role 3: Creative analyst (cinematographic view)]\n"
    "Analyze the filming and narrative techniques of each shot:\n"
    "  Cinematic language: camera movement (push/pull, pan, follow, or static),
     shot scale (close-up, medium, or wide), and camera angle (high, low, or
     eye-level)\n"
    "  Narrative structure: the function of the shot in the overall script
     (introduction, demonstration, conflict, transition, or ending)\n"
    "  Emotional tone: observable emotional style and pacing characteristics\n\n"

    "[Role 4: Commercial analyst (holistic view, no segmentation)]\n"
    "Make an overall commercial judgment about the video:\n"
    "  audience: target audience, formatted as [price positioning + group], no
     more than 20 words and without a qualifier prefix\n"
    "    Correct example: [mid-to-low-priced consumer goods, students and
     ordinary office workers]\n"
    "    Incorrect example: [consumers who value the user experience and dislike
     a sticky feeling]\n"
    "  ad_logic: core marketing logic, explaining in one sentence how the
     advertisement drives the purchase decision\n\n"

    "Output 3-5 additional tags for each dimension: short phrases, each no more
     than 10 words, representing the core category or key point of that
     dimension.\n"
    "Output all four dimensions at once, strictly in JSON format, with no content
     outside the JSON."
)
```

User prompt.

```
USER_PROMPT_ALL = """The following auxiliary information is provided for the video:

ASR transcript: {asr}
Product title: {item_title}
Video title: {content_title}
Merchant name: {shop_name}
Video description: {summary}
Shot time segmentation: {shot_detect}
Product category: {item_cate_level1_name} / {item_cate_level1_name2}

Watch the video frame by frame. Strictly follow the role requirements in the
system instruction and use shot time segmentation as the basis for segmentation.
If shot time segmentation is unavailable, use action changes, shot transitions,
or scene changes. Output the analysis for all four dimensions at once.
```

```
Output format (strict JSON, with no other content):
{{"events":{{"segments":[{{"start":0,"end":5,"caption":"Person A walks to the display,
 picks up the product, and the camera follows the hand movement"}}],"tags":["product
 demonstration","human interaction"]}},"objects":{{"segments":[{{"start":0,"end":5,
"caption":"A white display is centered, the product is on the left side of the
 surface, and a person stands in the right foreground"}}],"tags":["indoor setting",
"product close-up"]}},"creative":{{"segments":[{{"start":0,"end":5,"caption":"A
 static medium shot from an eye-level angle introduces the product"}}],"tags":["static
 camera","product introduction"]}},"commercial":{{"audience":"Mid-to-low-priced consumer
 goods for students and ordinary office workers","ad_logic":"A realistic usage scenario
 lowers the decision barrier and encourages impulse purchase","tags":["mass-market
 consumption","scenario-based promotion"]}}}}
"""
```

A.3 Dataset Split and Video-Overlap Audit

AdsCVR is split randomly into training and test partitions. To obtain a stricter test subset for cross-video transfer, we separately extracted test examples and performed a video-overlap audit against the training partition before evaluation. The resulting strictly non-overlapping subset contains 1k samples. Table 6 aligns the audited test-subset accuracy with the mixed-source AdSeek result across the six dimensions. The high accuracy retained by the audited subset, together with the explicit overlap review, supports the robustness of our method to cross-video transfer.

Table 6: Accuracy (%) on the strictly audited test subset and with mixed-source AdSeek across the six AdsCVR dimensions. AdsCVR is randomly split, and the separately extracted test subset is audited for video overlap with the training partition before evaluation.

Variant	TS	CS	ID	IC	OD	VC	Average
Strictly audited test subset	86.7	78.5	78.5	75.6	74.6	78.5	78.73
AdSeek (mixed)	77.0	73.4	70.8	75.6	74.8	74.2	74.30

B Implementation Details

Table 7 summarizes the SFT hyperparameters used for AdsCVR, and Table 8 reports the corresponding RL configuration. The default SFT launch uses 16 PPUs; the RL run uses 32 PPUs.

Table 7: SFT hyperparameters for AdsCVR.

Hyperparameter	Value
Base model	Qwen3-VL-8B-Instruct
Fine-tuning	Full-parameter SFT; FlashAttention 2; Liger kernel
Template	`qwen3_vl_nothink`
Cutoff length	51,200 tokens
Per-device batch size	1
Gradient accumulation	2
Learning rate	1×10^{-5}
Scheduler / warmup	Cosine / warmup ratio 0.1
Epochs	5
Precision / memory	BF16; gradient checkpointing
Checkpointing	Save every 10 steps; keep all checkpoints
Distributed training	16 PPUs

Table 8: RL hyperparameters for AdsCVR.

Hyperparameter	Value
Base model	Qwen3-VL-8B-Instruct
Training steps / checkpoint interval	500 / 20 steps
Rollout / validation batch size	512 / 256
Learning rate / scheduler	5×10^{-7} / constant
Weight decay / warmup	1×10^{-2} / 50 steps
Per-device batch / accumulation	1 / 16
Advantage estimator	GRPO; no advantage whitening
PPO epochs	1
KL regularization	KL loss coefficient 10^{-2}; initial KL coefficient 0.001
Entropy coefficient	0.01
Reward weights	Accuracy 1.0; format 0.1
Rollout groups	group size 8
Filtering / retries	`MAX_TRY` $= 3$
Precision	BF16
Distributed training	32 PPUs

C Video Complexity Metrics

We detail the temporal and spatial complexity proxies used for video filtering in Section 3.1.

Frame sampling and temporal complexity. We sample frames at approximately 1 fps and uniformly retain at most 16 frames per video. For the T selected frames, let $\mathbf{u}_t = \mathbf{z}_t/(\|\mathbf{z}_t\|_2 + \delta)$ be the normalized DINOv3 feature, with $\delta = 10^{-12}$. We compute:

$$\begin{aligned} c_t &= 1 - \frac{\sum_{j \in \mathcal{N}_t} w_{tj} \mathbf{u}_t^\top \mathbf{u}_j}{\sum_{j \in \mathcal{N}_t} w_{tj} + \delta}, \\ \mathrm{Dcpx}(V) &= \sum_{t=1}^{T} c_t, \qquad \mathrm{Vcpx}(V) = \frac{100}{T} \mathrm{Dcpx}(V), \end{aligned} \tag{8}$$

where $\mathcal{N}_t = \{j : 1 \le j \le T,\ 0 < |j - t| \le d\}$, $w_{tj} = \exp(-|j - t|/(2d))$, and $d = 8$. Both metrics are set to zero for $T \le 1$. Higher Dcpx and Vcpx indicate greater cumulative and average visual variation, respectively.

Spatial complexity. For the same T selected frames $I_1, \ldots, I_T$, we estimate spatial complexity using the average JPEG-encoded bytes per pixel:

$$\mathrm{Jcpx}(V) = \frac{1}{T} \sum_{t=1}^{T} \frac{B_{\mathrm{JPEG}}(I_t)}{H_t W_t}, \tag{9}$$

where $B_{\mathrm{JPEG}}(I_t)$ is the encoded size in bytes and $H_t \times W_t$ is the frame resolution. We use quality 95, 4:2:0 chroma subsampling, and optimized coding. Higher Jcpx indicates lower spatial compressibility under these fixed encoding settings.