VideoSailor: Navigating Video Deep Research via Trajectory-to-Policy Flywheel

Meng Cao[1]*, **Pengfei Hu**[1,2]*, **Yingyao Wang**[2], **Chen Wang**[2], **Pi Bu**[2], **Jun Song**[2], **YuCheng**[2], **Bo Zheng**[2], **Ian Reid**[1], **Xiaodan Liang**[1,3]†
[1]Mohamed bin Zayed University of Artificial Intelligence [2]Alibaba Group [3]Sun Yat-sen University

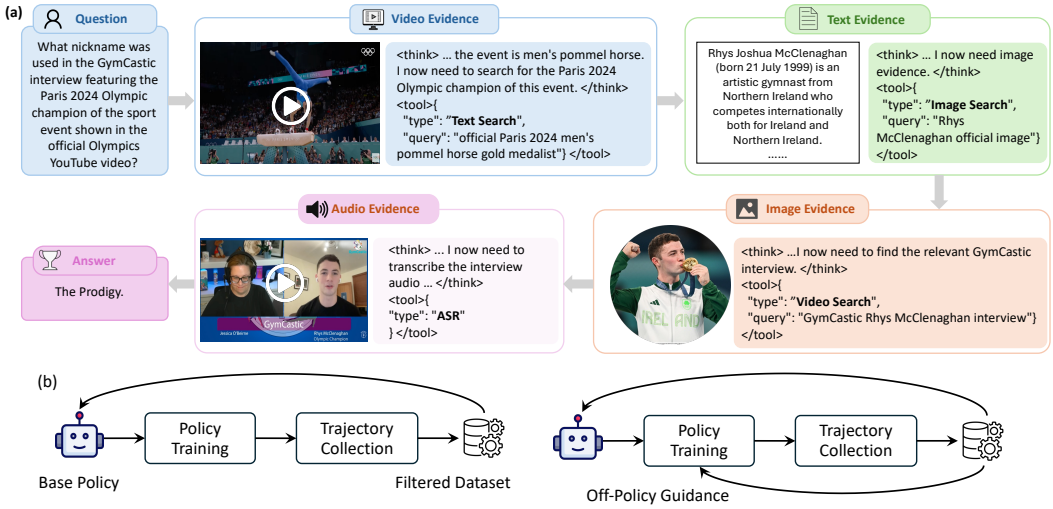


Figure 1: **(a) An exemplar case of our VideoSailor**, demonstrating cross-video and omni-modal reasoning through multi-hop evidence aggregation. (b) **Conceptual comparisons** between (left) the traditional data flywheel and (right) our trajectory-to-policy flywheel.

Abstract

Recent advances in Multi-modal Large Language Models (MLLMs) have significantly improved video understanding. However, real-world scenarios often require going beyond isolated video inputs, giving rise to the emerging paradigm of *video deep research*, which involves open-world reasoning through integration of external knowledge. Despite this progress, existing approaches are limited by their inability to perform effective inter-video interaction and their reliance on static, one-shot datasets that yield noisy and suboptimal supervision. To address these challenges, we propose **VideoSailor**, an open-world video deep research agent powered by a *trajectory-to-policy flywheel* that iteratively converts generated reasoning trajectories into high-quality *off-policy* guidance to mitigate the limitations of low-quality *on-policy* rollouts. For systematic evaluations, we further introduce **Omni-BrowseComp**, a benchmark that emphasizes omni-modal evidence aggregation, cross-video reasoning, and fine-grained temporal grounding with reasoning-relevant segment annotations. Extensive experiments demonstrate that VideoSailor significantly improves performance on complex multi-hop and cross-video reasoning tasks, establishing a scalable and effective paradigm for open-world video deep research. The project is available at `https://mengcaopku.github.io/VideoSailor`.

* Equal contributions.
† Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: **Comparisons with existing video reasoning benchmarks.** "Cross-video" indicates whether tasks require reasoning across multiple videos. "Temp. Anno." (Temporal Annotation) denotes the availability of fine-grained segment-level annotations. ✓ denotes full support, ✗ denotes no support, and △ denotes partial support. In Video-BrowseComp [38], △ indicates that only a small subset (23 samples) involves cross-video reasoning.

Benchmarks	Video-involved	Open-world	Omni-modal	Cross-video	Temp. Anno.
OmniVideoBench [31]	✓	✗	✓	✗	✗
VDR-Bench [74]	✗	✓	✗	✗	✗
MM-BrowseComp [33]	✗	✓	✗	✗	✗
VideoSimpleQA [6]	✓	✓	✗	✗	✓
Video-BrowseComp [38]	✓	✓	✗	△	✗
VideoDR [39]	✓	✓	✗	✗	✗
SAGE-Bench [29]	✓	✓	✗	✗	✗
Omni-BrowseComp (Ours)	✓	✓	✓	✓	✓

1 Introduction

Recent advances in Multi-modal Large Language Models (MLLMs) [43, 1, 3] have evolved into general-purpose assistants for video understanding, including long-horizon comprehension [53, 65], complex reasoning [15, 35], and temporal grounding [67, 26]. Despite this progress, real-world video understanding extends beyond isolated video inputs, necessitating the integration of world knowledge and interaction with external systems, as exemplified by the emerging paradigm of video deep research [72, 38, 39, 77, 37, 29].

Current video deep research approaches can be broadly categorized into two paradigms: *inference-time agentic methods* [77, 72, 80], where MLLMs act as orchestrators for tool use and external retrieval, and *training-based methods* [29, 75, 58, 72, 24], which enhance reasoning through supervised fine-tuning (SFT) or reinforcement learning (RL) on curated datasets. However, both paradigms suffer from two fundamental limitations: 1) **Limited inter-video interaction.** Current methods are largely confined to reasoning over a *single* video, neglecting the need for cross-video retrieval and interaction in open-world settings, where relevant evidence is often distributed across multiple videos. For example, as illustrated in Figure 1 (a), answering the query requires aggregating evidence across multiple videos and modalities, including video clips, textual descriptions, images, and audio transcripts. This necessitates holistic reasoning and evidence aggregation beyond isolated sources; **2) Static one-shot training data.** These methods rely on static, one-shot generated training datasets, whose quality is constrained by the current deep research capabilities of MLLMs, particularly in complex video scenarios. Without iterative refinement or feedback from model execution, such datasets are prone to noisy reasoning traces and suboptimal supervision, limiting their ability to support complex, long-horizon reasoning. As illustrated in Figure 2 (a), we empirically observe that simply increasing the amount of one-shot RL training data leads to rapid performance saturation and even slight degradation beyond a certain scale.

To address these limitations, we propose **VideoSailor**, an open-world video deep research agent that actively seeks, aligns, and integrates information across inter-video and omni-modal evidence. Unlike prior approaches that rely on static pre-generated datasets or passively reuse model-generated data, we introduce a **trajectory-to-policy flywheel**, which iteratively converts generated reasoning trajectories into high-quality off-policy guidance for policy optimization. Specifically, we observe that on-policy rollouts in RL often suffer from low-quality trajectory generation, especially in complex video reasoning scenarios. Therefore, the proposed trajectory-to-policy flywheel explicitly selects and refines generated trajectories to construct a high-quality off-policy dataset, which is then used as additional guidance for policy optimization (*c.f*. the right panel of Figure 1 (b)). Through this process, VideoSailor establishes a self-reinforcing learning loop that progressively improves both data quality and policy performance, enabling more scalable learning of long-horizon, multi-hop reasoning behaviors in open-world video environments (*c.f*. Figure 2 (b)).

To systematically evaluate open-world video reasoning, we introduce **Omni-BrowseComp**, a benchmark designed to capture the core challenges of video deep research. As shown in Table 1, Omni-BrowseComp exhibits three distinctive features: *1) Omni-modality reasoning*. Tasks require evidence

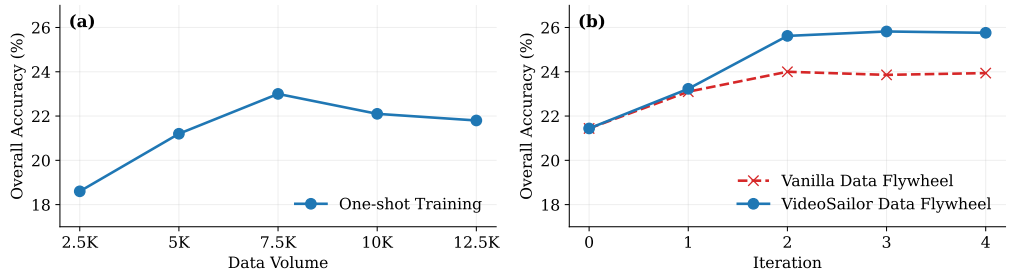


Figure 2: **(a)** Overall accuracy with respect to RL data volume under one-shot training on Omni-BrowseComp benchmark. **(b)** Overall accuracy across flywheel iterations for both the vanilla data flywheel and our proposed trajectory-to-policy flywheel.

extraction across heterogeneous modalities, including text, image, video, and audio, enabling unified omni-modal reasoning; *2) Cross-video reasoning*. Tasks are explicitly constructed to require retrieval and interaction across multiple videos. While the recent benchmark Video-BrowseComp [38] includes cross-video reasoning instances, these are restricted to the "Level-3" difficulty category defined in their work and comprise only 23 samples, a scale that is far too small for reliable and systematic evaluation. In contrast, our Omni-BrowseComp systematically incorporates cross-video reasoning as a fundamental requirement for each case; *3) Temporal grounding*. We provide fine-grained supervision by annotating precise, reasoning-relevant temporal segments for each video in the reasoning chain, enabling frame-level grounding and discouraging reliance on coarse video-level cues.

In summary, our contributions are in three-folds:

- We propose VideoSailor, a video deep research framework built on a trajectory-to-policy flywheel that leverages the filtered trajectories as high-quality off-policy guidance for policy optimization.
- We introduce Omni-BrowseComp, a benchmark for open-world video reasoning that emphasizes omni-modal evidence extraction, cross-video reasoning, and fine-grained temporal grounding.
- Extensive experiments demonstrate that our VideoSailor outperforms state-of-the-art tool-free and tool-augmented MLLMs across both our Omni-BrowseComp benchmark and existing benchmarks.

2 Related Work

Video Deep Research Agents. The emergent reasoning capability of MLLMs [62, 10, 43] has transformed video understanding from direct question answering into problem-solving with interpretable long chain-of-thought reasoning. Early approaches [61, 23, 78] primarily rely on SFT with curated datasets to instill step-by-step reasoning grounded in video content. Subsequent works [15, 47, 75] introduce RL with verifiable objectives to refine reasoning through exploration and outcome-based feedback. However, these methods are largely confined to reasoning within individual videos and fail to address open-world scenarios that require integration of external knowledge.

To bridge this gap, recent efforts extend video reasoning toward Video Deep Research [72, 38, 39, 77, 37, 29], where video understanding is no longer limited to passive perception but involves actively seeking, retrieving, and integrating information from external sources in an iterative manner [64, 33, 74, 28, 68, 30]. Early works adopt agentic workflows [77, 72, 80], where MLLMs are instructed to follow a *reasoning-then-tool-call* [70] pipeline to solve complex tasks. However, such approaches rely heavily on predefined execution patterns and fail to internalize deep research capabilities, limiting their ability to generalize beyond fixed workflows. More recent efforts shift toward training-based paradigms [29, 75, 58, 72] by constructing SFT or RL datasets. For instance, SAGE [29] trains an any-horizon agent equipped with web search tools to perform knowledge-driven multi-turn reasoning. Despite these advances, these methods remain limited in *single*-video understanding and are unable to effectively handle cross-video evidence aggregation. Moreover, the reliance on static, one-shot generated training datasets often leads to suboptimal data quality, which in turn hinders the scalability of video deep research agents. To address these challenges, we propose VideoSailor, a framework that enables inter-video interaction and iterative data refinement through a trajectory-to-policy flywheel, allowing scalable learning of open-world video reasoning.

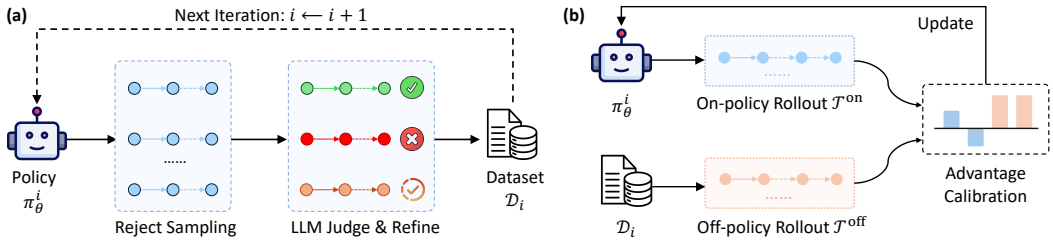


Figure 3: **(a) Overview of the trajectory-to-policy flywheel.** In i-th iteration, the current policy π_θ^i generates candidate trajectories via reject sampling, which are then refined by an LLM-based judge to form a high-quality off-policy dataset $\mathcal{D}_i$. **(b) Mix-policy optimization.** The policy is updated by jointly leveraging on-policy trajectories and off-policy data $\mathcal{D}_i$, with a calibrated advantage mechanism to ensure stable integration (*c.f.* Eq. (4)).

Algorithm 1: Trajectory-to-Policy Flywheel

Input: Seed dataset $\mathcal{D}_0$, policy π_θ, judge model $\mathcal{J}$, iteration number N
Output: Updated policy π_θ
Warm up π_θ with SFT and GRPO on $\mathcal{D}_0$;
for $i = 1$ **to** N **do**
  Sample reasoning trajectories $\mathcal{T}_i$ from π_θ (*c.f.* Eq. (1));
  Use $\mathcal{J}$ to classify trajectories $\mathcal{T}_i$ into correct-and-valid, correct-but-flawed, and incorrect;
  Refine correct-but-flawed trajectories with judge model $\mathcal{J}$;
  Construct filtered dataset $\mathcal{D}_i$ from correct-and-valid and refined trajectories (*c.f.* Eq. (2));
  Compute calibrated advantages for on-policy and off-policy trajectories (*c.f.* Eq. (3) and Eq. (4));
  Update π_θ using mix-policy optimization (*c.f.* Eq. (5));
return π_θ

Data Flywheel. Originating from the "flywheel effect" [9], the data flywheel [16, 4, 56, 14] describes a self-reinforcing closed-loop learning paradigm in which models iteratively generate and leverage new data for training, resulting in continuous and accelerating performance improvements. Building upon the success of data flywheel approaches in the NLP domain [73, 20, 54, 41, 36, 11], recent studies have extended this paradigm to multi-modal settings, including agentic learning [50, 79, 40], MLLM reasoning [60, 12, 76, 59, 13], and embodied intelligence [63, 81, 17, 19]. Arena Learning [40] constructs a competitive training environment where models iteratively generate new training data by competing with stronger opponents. DexFlyWheel [81] enables scalable and autonomous data generation for robotic manipulation via iterative rollout and augmentation cycles. OpenVLThinker [13] iteratively alternates SFT and RL, with the RL-improved model generating higher-quality reasoning trajectories to construct new training data for subsequent iterations. However, these methods merely treat the generated data as additional training samples, lacking a more structured and effective utilization of such data. In contrast, our approach explicitly leverages the generated data as off-policy signals in the RL process, enabling more effective and robust policy optimization.

3 VideoSailor

We propose VideoSailor, an open-world video deep research framework built on a trajectory-to-policy flywheel. As shown in Figure 3 and Algorithm 1, the pipeline consists of two stages including a warm-up stage that constructs a seed dataset to initialize the policy via SFT and RL, and a trajectory-to-policy flywheel based on the proposed mix-policy optimization.

3.1 Warm-up Stage

Seed Dataset Curation. To bootstrap the trajectory-to-policy flywheel, we construct a seed dataset $\mathcal{D}_0$ through a structured multi-stage data construction pipeline. As shown in Figure 4, the whole pipeline includes the follow parts: 1) **Unified captioning**: We begin by collecting diverse multi-modal data from existing datasets, including video datasets (ActivityNet-QA [71], How2QA [32], Video SimpleQA [6]) and image datasets (YFCC100M [57], SimpleVQA [8]). We then unify all



Figure 4: Overview of the seed dataset construction pipeline.

modalities into textual representations through captioning via GPT-4o [43], producing structured semantic descriptions that serve as a common interface for downstream reasoning; 2) **Multi-hop chain construction**. We construct multi-hop reasoning chains by organizing the captioned results into structured trajectories, where each caption serves as a node. A trajectory begins with a starting observation node, followed by intermediate nodes that progressively introduce necessary information, and ends with a final node from which the answer can be derived; 3) **QA generation**. Based on the constructed reasoning chains, we generate the question-answer pairs by prompting GPT-4o. The question is carefully designed to avoid exposing intermediate reasoning steps, ensuring that the model must perform multi-hop reasoning to arrive at the answer; 4) **Quality control**. We perform quality control using a judge LLM [43] that filters generated reasoning trajectories to meet the following criteria: *i) Non-revealing input*: the input modality only provides partial information and does not directly reveal the answer; *ii) Step necessity*: each intermediate reasoning step is necessary; *iii) Compositional reasoning*: the final answer requires the composition of multiple reasoning steps; *iv) Semantic coherence*: the reasoning chain within each domain is coherent and natural; *v) Shortcut-free*: no shortcut exists where intermediate nodes can directly yield the answer without full reasoning.

We split the constructed dataset into two parts for SFT and RL. The former uses the full QA pairs along with reasoning trajectories, while the latter uses only the QA pairs based on the outcome rewards via GRPO [21]. Detailed statistics are provided in Figure 5 (right).

Initial Policy Optimization. We initialize the policy model using a combination of SFT for learning structured reasoning and RL with GRPO for policy exploration. We adopt a reward formulation including a formatting reward and an accuracy reward. Specifically, the model is required to generate outputs in a structured format with `<think></think>`, `<tool></tool>`, and `<ans></ans>` fields, where `<think></think>` contains the step-by-step reasoning process, `<tool></tool>` specifies the supporting tool including text search, image search, video search and ASR (*c.f.* Sec A.2 for more details), and `<ans></ans>` provides the final answer. The accuracy reward is defined based on the correctness of the final answer against the ground-truth answer using lexical matching [15].

3.2 Trajectory-to-Policy Flywheel

This stage implements a closed-loop trajectory-to-policy flywheel mechanism that iteratively improves reasoning capability by transforming generated trajectories into policy optimization signals. At each iteration $i \in \{1, 2, \ldots, n-1\}$, the flywheel performs two key operations:

Trajectory Refinement. We refine generated reasoning trajectories using an LLM-based judge [43] to improve trajectory quality and recover useful supervision signals from imperfect rollouts.

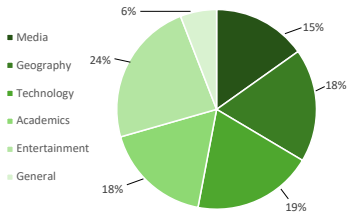
Given the current frozen policy π_θ^i, we sample a set of reasoning trajectories for each input query x:

$$\mathcal{T}_i = \{\tau \sim \pi_\theta^i(\cdot \mid x)\}, \tag{1}$$

where each trajectory τ represents a multi-hop reasoning process composed of structured intermediate steps. Due to the complexity of video deep reasoning, these trajectories may contain incorrect answers or flawed reasoning. To address this, we employ an MLLM-based judge $\mathcal{J}$ [43] to evaluate each trajectory with respect to answer correctness and reasoning quality. Specifically, trajectories are divided into three groups: 1) *correct and valid*, where both the final answer and the reasoning process are correct; 2) *correct but flawed*, where the final answer is correct but the reasoning trajectory contains errors or inconsistencies; and 3) *incorrect*, where the final answer itself is wrong.

Instead of discarding imperfect trajectories, we specifically refine those in the correct but flawed group using the judge model, correcting their reasoning paths while preserving the correct answers. The *incorrect* trajectories are discarded. The refined trajectories $\mathcal{T}_{\text{refined}}$, together with the correct and valid ones $\mathcal{T}_{\text{correct}}$, form a high-quality trajectory set:

$$\mathcal{D}_i = \mathcal{T}_{\text{correct}} \cup \mathcal{T}_{\text{refined}}. \tag{2}$$



Statistics	SFT	RL	Eval
# Samples	3129	7032	1003
Avg. videos/sample	4.35	4.62	4.08
Max videos/sample	8	9	7
Avg. reasoning hops	5.89	6.21	5.37
Max reasoning hops	18	20	17
Avg. segment dur. (s)	–	–	21.22

Figure 5: **Left: Category distribution** of the proposed Omni-BrowseComp benchmark. **Right: Dataset statistics.** SFT and RL report statistics of the seed dataset $\mathcal{D}_0$ (*c.f.* Sec 3.1), while Eval reports statistics from the Omni-BrowseComp benchmark (*c.f.* Sec 4).

Mix-Policy Optimization. Due to the inherent complexity of video deep reasoning, *on-policy* rollouts in GRPO [21] often fail to generate correct reasoning trajectories, leading to weak or unstable optimization signals. To address this limitation, we propose a mix-policy optimization framework that directly incorporates refined trajectories $\mathcal{D}_i$ obtained from the previous data flywheel process as *off-policy* rollouts. These trajectories are of higher quality, as they are filtered and refined by an external judge, providing more reliable supervision signals.

To achieve this, a straightforward approach is to treat off-policy trajectories equally with on-policy samples when computing advantages for RL optimization. However, this naive mixing strategy introduces a significant distributional bias: off-policy trajectories, being of higher quality, consistently receive the highest rewards, which shifts the group statistics (mean and variance) and causes most on-policy samples to obtain negative or near-zero advantages. As a result, gradient updates become dominated by off-policy samples, suppressing useful exploration of on-policy samples.

Therefore, we propose a *calibrated advantage* strategy for *mixed-policy* optimization. Specifically, the advantage computation for on-policy trajectories remains unchanged, *i.e.*, it is computed based on the statistics solely from on-policy samples:

$$A(\tau) = \frac{r(\tau) - \mu}{\sigma}, \quad \tau \in \mathcal{T}^{\text{on}}, \tag{3}$$

where μ and σ denote the mean and standard deviation of rewards $r(\tau)$ for on-policy samples $\mathcal{T}^{\text{on}}$.

For off-policy trajectories $\tau \in \mathcal{D}_i$, instead of including them in normalization computation, we calibrate their advantages using the on-policy scale:

$$A(\tau) = \lambda \cdot \max_{\tau' \in \mathcal{T}^{\text{on}}} A(\tau'), \quad \tau \in \mathcal{D}_i, \tag{4}$$

where λ is a fixed scaling factor. This mechanism preserves the high-quality guidance from off-policy trajectories while preventing them from dominating the optimization process, thereby maintaining stable and consistent policy gradient updates. Using these calibrated advantages, we optimize the policy with a mixed-policy objective that jointly incorporates on-policy and off-policy trajectories:

$$\begin{aligned} \mathcal{L}_{\text{Mix-GRPO}}(\theta) = \frac{1}{|\mathcal{T}^{\text{on}}| + |\mathcal{D}_i|} \sum_{\tau \in \mathcal{T}^{\text{on}} \cup \mathcal{D}_i} \min \Bigg(& \frac{\pi_\theta(\tau \mid x)}{\pi_{\theta_{\text{old}}}(\tau \mid x)} A(\tau), \\ & \text{clip}\left(\frac{\pi_\theta(\tau \mid x)}{\pi_{\theta_{\text{old}}}(\tau \mid x)}, 1-\epsilon, 1+\epsilon \right) A(\tau) \Bigg) - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}). \end{aligned} \tag{5}$$

4 Omni-BrowseComp Benchmark

Design Principles. As shown in Figure 5, we introduce Omni-BrowseComp, a benchmark designed to capture the core challenges of video-centric deep research: *1) Omni-modality reasoning.* Tasks require evidence extraction across heterogeneous modalities, including text, image, video, and audio; *2) Cross-video reasoning.* We design tasks to explicitly require retrieval and reasoning across multiple videos, rather than being solvable within a single video. Unlike prior benchmarks (*e.g.*, Video-BrowseComp [38]) where cross-video reasoning appears only sparsely, our Omni-BrowseComp treats it as a core requirement, ensuring that multi-video interaction is consistently necessary for

problem solving; *3) Temporal grounding.* We provide fine-grained supervision by annotating precise temporal segments for each video involved in the reasoning chain. This enables frame-level grounding and discourages models from relying on coarse video-level cues.

Annotation Pipeline. We adopt a reverse construction paradigm to build complex multi-hop reasoning tasks, where the process starts from identifying a target answer and progressively expands into a compositional query: *1) Answer-first construction.* We first select an answer that is unique, verifiable, and temporally stable, ensuring that it can be grounded in publicly available video content; *2) Multi-hop chain design.* Based on the selected answer, annotators construct an initial simple question, which is then expanded into a multi-hop reasoning problem by introducing intermediate constraints and retrieval steps. These steps form a structured reasoning chain, where each hop corresponds to a necessary operation, often requiring evidence extraction from multimodal sources (*e.g.*, video, text, and web content); *3) Checklist-based supervision.* To facilitate explicit reasoning supervision, each sample is further annotated with a structured checklist that decomposes the solution process into executable steps, including retrieval actions, reasoning operations, and video-based evidence localization; *4) Temporal annotation.* Finally, we provide fine-grained temporal annotations by marking the start and end timestamps of key video segments, enabling precise evaluation of temporal grounding and ensuring that the reasoning process is fully traceable.

Quality Control. To ensure the quality and reliability of the dataset, we enforce a set of strict constraints through a cross-annotator verification process, where multiple annotators independently review and validate each sample: *1) Video dependency.* Each task must require multiple video-based reasoning steps, ensuring that the solution critically depends on video content and cannot be derived through text-only reasoning. *2) Non-triviality.* Tasks are validated using strong models (*e.g.*, GPT-5.5 [46], Gemini 2.5 Pro [10]) to ensure that they cannot be solved correctly in a single attempt, thereby preventing trivial or memorization-based solutions. *3) Answer validity.* Answers must be unique, verifiable, and temporally stable. All answers are checked against authoritative sources to ensure consistency and enable reliable automatic evaluation. *4) Reasoning integrity.* The reasoning chain must be non-reducible, *i.e.*, each intermediate step is necessary and cannot be bypassed. This constraint ensures that tasks require genuine multi-hop reasoning rather than shortcut-based inference.

Evaluation Metrics. We follow MM-BrowseComp [33] to adopt three evaluation metrics including Overall Accuracy (OA), Strict Accuracy (SA), and Average Checklist Score (AVG CS).

5 Experiments

5.1 Main Results

Implementation Details. Our implementation is based on Qwen2.5-VL-7B-Instruct [3]. We use a batch size of 8, with learning rates of 5×10^{-6} for SFT and 1×10^{-6} for RL. The on-policy rollout group size is set to 8. All experiments are conducted on 8 A100 GPUs. The scaling factor λ, number of flywheel iterations N, and the number of off-policy trajectories $|\mathcal{D}_i|$ are set to 1.2, 3, and 4, respectively. Refer to Table A4 for more details.

Performance comparison on Omni-BrowseComp. As shown in Table 2, we conduct a comprehensive evaluation on Omni-BrowseComp across tool-free MLLMs, tool-augmented MLLMs, and open-source agents, and derive the following key findings: **i) Tool augmentation is beneficial.** Tool-augmented methods (*e.g.*, GPT-5.5 [46], Gemini-3 [18]) consistently outperform tool-free MLLMs, demonstrating the importance of external tool use for complex reasoning and information retrieval. However, their improvements in Strict Accuracy (SA) remain limited, suggesting that these methods still struggle with end-to-end correctness. **ii) Agent-based reasoning is beneficial yet limited.** Compared to tool-free MLLMs, agentic methods (*e.g.*, Agent-R1 [7], WebDancer [66]) achieve substantially higher AVG CS scores, indicating better sub-step reasoning and partial task completion. **iii) Holistic reasoning remains challenging.** Across all methods, SA is significantly lower than AVG CS, indicating that models can often complete individual reasoning steps but fail to produce fully correct step-by-step reasoning and final answers. **iv) Elevated difficulty of Omni-BrowseComp.** Compared to the results on Video-BrowseComp [38], almost all the methods experience a noticeable performance drop on Omni-BrowseComp in terms of the overall accuracy. This suggests that our benchmark introduces additional complexity, making it a more demanding testbed for real-world agent capabilities. **v) The state-of-the-art performance of VideoSailor.** Our VideoSailor achieves the highest OA (25.82%) and SA (11.96%) among all methods, significantly outperforming both tool-

Table 2: **Performance comparison** on our Omni-BrowseComp and Video-BrowseComp [38] benchmarks. On Omni-BrowseComp, we report Overall Accuracy (OA), Strict Accuracy (SA), and Average Checklist Score (AVG CS), while on Video-BrowseComp we report OA, difficulty-level accuracies (L1–L3), and Calibration Error (CE↓). **Bold** and underlined values denote the best and second-best performance in each column, respectively.

Model	Omni-BrowseComp			Video-BrowseComp [38]				
	OA ↑	SA ↑	AVG CS ↑	OA ↑	L1 ↑	L2 ↑	L3 ↑	CE ↓
Tool-Free MLLMs								
o4-mini [45]	18.35	0.26	5.47	16.67	25.60	6.45	8.70	36.84
GPT-5.1 [44]	10.57	0.77	5.13	15.24	21.60	6.45	4.35	30.20
GPT-4o-2024-11-20 [43]	10.05	1.55	7.50	17.62	28.00	3.23	0.00	58.81
GPT-4o-mini [43]	7.13	0.98	10.47	9.52	16.00	0.00	0.00	63.55
Claude Sonnet 4.5 [2]	8.53	1.81	12.02	14.29	22.40	4.84	4.35	44.18
Qwen2.5-VL-72B-Instruct [3]	10.03	1.85	9.12	13.81	22.40	3.23	4.35	47.62
Qwen2.5-VL-32B-Instruct [3]	9.18	1.54	8.11	11.43	18.40	3.23	0.00	52.18
Qwen2.5-VL-7B-Instruct [3]	6.34	0.71	5.32	8.10	12.80	1.61	0.00	61.35
Tool-Augmented MLLMs								
GPT-5.5 [46]	24.33	1.10	5.86	26.67	40.00	8.06	13.04	25.84
o3 [45]	23.45	0.77	4.21	25.24	38.40	6.45	13.04	29.72
Gemini-2.5-Pro-Preview [10]	17.48	5.14	17.16	23.81	37.60	4.84	0.00	31.45
Gemini-3-Flash-Preview [18]	22.74	7.24	21.55	20.95	32.80	4.84	0.00	35.98
Open-source Agents								
Agent-R1 [7]								
Gemini-2.5-Flash-Preview	23.25	10.68	**23.41**	25.71	36.80	6.45	**21.74**	32.46
GPT-4o-2024-11-20	16.91	4.37	13.45	24.76	34.40	6.45	**21.74**	34.92
Qwen2.5-VL-72B-Instruct	14.39	4.15	12.28	20.48	29.60	4.84	13.04	38.76
OWL [25]								
Gemini-2.5-Flash-Preview	20.14	8.73	20.38	23.33	33.60	6.45	17.39	35.12
GPT-4o-2024-11-20	12.37	3.18	10.07	21.90	31.20	4.84	17.39	38.41
Qwen2.5-VL-72B-Instruct	13.42	3.04	11.25	18.57	27.20	4.84	8.70	42.63
DeerFlow [5]								
Gemini-2.5-Flash-Preview	18.37	6.14	18.45	22.38	32.00	6.45	17.39	36.64
GPT-4o-2024-11-20	11.45	2.53	9.15	20.95	29.60	4.84	17.39	40.28
Qwen2.5-VL-72B-Instruct	11.07	1.93	9.22	17.62	25.60	4.84	8.70	44.15
WebDancer [66]								
Gemini-2.5-Flash-Preview	18.14	5.73	17.33	21.43	31.20	4.84	13.04	37.85
GPT-4o-2024-11-20	11.37	2.08	8.31	19.52	28.80	4.84	8.70	41.56
Qwen2.5-VL-72B-Instruct	11.08	1.97	9.36	18.10	26.40	4.84	8.70	43.47
WebDancer-32B	17.37	5.41	16.98	18.57	27.20	4.84	8.70	42.18
VideoSailor (Ours)	**25.82**	**11.96**	22.74	**30.00**	**41.60**	**11.29**	17.39	**24.48**

augmented MLLMs and existing agents. While its AVG CS is slightly lower than the top-performing agent-based method, this suggests that VideoSailor prioritizes end-to-end correctness over partial sub-step completion, leading to more reliable final answers.

Performance comparison on Video-BrowseComp [38]. Our VideoSailor consistently achieves the best performance across most metrics on Video-BrowseComp. Specifically, it attains the highest overall accuracy on Level-1 and Level-2 tasks, while achieving the lowest calibration error, demonstrating not only strong predictive performance but also more reliable and well-calibrated confidence estimates. For Level-3, VideoSailor achieves the second-best performance. This slight disadvantage is likely due to the limited number of samples in Level-3 (only 23), where the gap to the best model corresponds to just one additional correct prediction (5 *vs*. 4). Such a small sample size is insufficient to reliably reflect performance differences in this highly challenging cross-video reasoning setting. In contrast, our Omni-BrowseComp explicitly enforces cross-video reasoning and therefore provides a more reliable and robust evaluation.

5.2 Ablation Studies on Training Paradigms

Ablation on SFT and RL. As shown in Table 3 (left), based on Qwen2.5-VL-7B [3], both SFT and RL with the seed dataset lead to consistent performance gains. Compared to the base model, SFT alone significantly improves performance (16.82% *vs*. 6.34% on overall accuracy), demonstrating the effectiveness of structured supervision in learning basic reasoning patterns. Further incorporating

Table 3: **Ablation studies** on (left) training paradigms and (right) tool repositories of our VideoSailor.

Training Variant	OA	SA	AVG CS
Base MLLM (Qwen2.5-VL-7B)	6.34	0.71	5.32
SFT only	16.82	5.34	14.76
SFT + RL	19.26	6.92	16.84
SFT + RL + Vanilla Flywheel	20.34	7.78	17.92
SFT + RL + Trajectory-to-Policy Flywheel	**25.82**	**11.96**	**22.74**

Tool Variant	OA	SA	AVG CS
Full VideoSailor	**25.82**	**11.96**	**22.74**
w/o Video Search	12.84	3.92	11.36
w/o Text Search	13.67	4.28	12.14
w/o Image Search	20.91	8.46	18.73
w/o ASR	21.48	8.91	19.26

Table 4: **Ablation studies** on our trajectory-to-policy flywheel.

Training Variant	Refine	Off-policy	Calib. Adv.	OA	SA	AVG CS
Full VideoSailor	✓	✓	✓	**25.82**	**11.96**	**22.74**
w/o off-policy guidance	✓	×	×	20.34	7.78	17.92
Naive mixed normalization	✓	✓	×	23.03	9.17	20.51
w/o trajectory refinement	×	✓	✓	22.13	8.97	18.86

RL yields additional improvements (19.26% *vs*. 16.82% on overall accuracy), indicating that policy optimization enhances exploration and reasoning robustness beyond supervised learning.

Vanilla Flywheel *vs*. Trajectory-to-Policy Flywheel. As shown in Table 3 (left), we notice that the vanilla flywheel (*i.e*., removing the off-policy guidance in our trajectory-to-policy flywheel), yields consistent but moderate gains (*e.g*. 20.34% *vs*. 19.26% on overall accuracy). It suggests that iterative data reuse provides additional supervision, yet is limited by the quality of generated trajectories. In contrast, our proposed trajectory-to-policy flywheel achieves substantially larger improvements (*e.g*. 25.82% *vs*. 19.26% on overall accuracy). This manifests that explicitly using refined trajectories as high-quality off-policy guidance enables more efficient policy optimization.

Ablation on Tool Repositories. Table 3 (right) evaluates the contribution of different tools in VideoSailor. Removing any tool leads to a noticeable performance drop, highlighting the importance of multi-modal evidence aggregation. In particular, removing video search leads to the largest performance drop (25.82% *vs*. 12.84% on overall accuracy), underscoring the critical role of video evidence in task completion.

5.3 Ablation Studies on Trajectory-to-Policy Flywheel

Naive Mixing *vs*. Calibrated Advantage. We design an advantage calibration scheme for off-policy rollouts (*c.f*. Eq. (4)). To demonstrate its effectiveness, we replace the proposed calibrated advantage with a *naive mixing* strategy that directly mixes on-policy and off-policy samples in the same normalization process. Detailed formulations of the naive mixing strategy are provided in Appendix A.2. The results in Table 4 show that naive mixing leads to noticeable performance degradation (25.82% *vs*. 23.03% on overall accuracy), which may be due to the fact that high-quality off-policy samples tend to receive disproportionately large rewards, dominating the normalization process and suppressing informative on-policy signals, thereby resulting in unstable optimization. Refer to Figure A2 for detailed training stability comparisons.

Ablation on Trajectory Refinement. Trajectory refinement is designed to recover useful supervision from imperfect rollouts by correcting flawed reasoning paths. We evaluate whether refinement provides additional benefits beyond simple filtering. We remove the refinement step and only retain trajectories that are classified as correct-and-valid, discarding all flawed trajectories without correction. As shown in Table 4, removing refinement leads to a noticeable performance drop. This suggests that refined trajectories provide valuable guidance signals for effective policy exploration.

6 Conclusion

In this paper, we introduce VideoSailor, an open-world video deep research framework that addresses two fundamental limitations of existing approaches: limited inter-video interaction and reliance on static, one-shot training data. By leveraging a trajectory-to-policy flywheel, VideoSailor iteratively refines generated reasoning trajectories into high-quality off-policy supervision, enabling more effective policy optimization for complex multi-hop reasoning. Extensive experiments demonstrate that VideoSailor consistently outperforms state-of-the-art MLLMs on both our Omni-BrowseComp benchmark and the existing Video-BrowseComp benchmark.

Acknowledgements and Disclosures

This work was supported by the startup funding of Prof. Xiaodan Liang at Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). The authors declare no competing interests.

References

[1] Anthropic. Claude 3.5 sonnet, 2024.

[2] Anthropic. Claude sonnet 4.5. `https://platform.claude.com/docs/en/about-claude/models/overview`, 2025.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Manish Bhattarai, Ryan Barron, Maksim Eren, Minh Vu, Vesselin Grantcharov, Ismael Boureima, Valentin Stanev, Cynthia Matuszek, Vladimir Valtchinov, Kim Rasmussen, et al. Heal: hierarchical embedding alignment loss for improved retrieval and representation learning. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 205–214, 2025.

[5] ByteDance. DeerFlow. GitHub repository. Accessed: 2026-10-31.

[6] Meng Cao, Pengfei Hu, Yingyao Wang, Jihao Gu, Haoran Tang, Haoze Zhao, Chen Wang, Jiahua Dong, Wangbo Yu, Ge Zhang, et al. Video simpleqa: Towards factuality evaluation in large video language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2616–2624, 2026.

[7] Mingyue Cheng, Jie Ouyang, Shuo Yu, Ruiran Yan, Yucong Luo, Zirui Liu, Daoyu Wang, Qi Liu, and Enhong Chen. Agent-r1: Training powerful llm agents with end-to-end reinforcement learning. *arXiv preprint arXiv:2511.14460*, 2025.

[8] Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646, 2025.

[9] Jim Collins. *Good to Great: Why Some Companies Make the Leap and Others Don't*. HarperBusiness, 2001.

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[11] Caia Costello, Simon Guo, Anna Goldie, and Azalia Mirhoseini. Think, prune, train, improve: Scaling reasoning without scaling models. *arXiv preprint arXiv:2504.18116*, 2025.

[12] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*, 2024.

[13] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv e-prints*, pages arXiv–2503, 2025.

[14] Xiangjue Dong, Maria Teleki, and James Caverlee. A survey on llm inference-time self-improvement. *arXiv preprint arXiv:2412.14352*, 2024.

[15] 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. *arXiv preprint arXiv:2503.21776*, 2025.

[16] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

[17] Seyed Kamyar Seyed Ghasemipour, Ayzaan Wahid, Jonathan Tompson, Pannag Sanketi, and Igor Mordatch. Self-improving embodied foundation models. *arXiv preprint arXiv:2509.15155*, 2025.

[18] Google. Gemini 3 flash preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview`, 2025.

[19] Jennifer Grannen, Michelle Pan, Kenneth Llontop, Cherie Ho, Mark Zolotas, Jeannette Bohg, and Dorsa Sadigh. Robot-powered data flywheels: Deploying robots in the wild for continual data collection and foundation model adaptation. *arXiv preprint arXiv:2511.19647*, 2025.

[20] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

[21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[22] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[23] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191, 2025.

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

[25] Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, et al. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation. *arXiv preprint arXiv:2505.23885*, 2025.

[26] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024.

[27] Hsiu-Yuan Huang, Chenming Tang, Weijie Liu, Clive Bai, Saiyong Yang, and Yunfang Wu. Think outside the policy: In-context steered policy optimization. *arXiv preprint arXiv:2510.26519*, 2025.

[28] Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, et al. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. *arXiv preprint arXiv:2601.22060*, 2026.

[29] Jitesh Jain, Jialuo Li, Zixian Ma, Jieyu Zhang, Chris Dongjoo Kim, Sangho Lee, Rohun Tripathi, Tanmay Gupta, Christopher Clark, and Humphrey Shi. Sage: Training smart any-horizon agents for long video reasoning with reinforcement learning. *arXiv preprint arXiv:2512.13874*, 2025.

[30] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.

[31] Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025.

[32] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020.

[33] Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, et al. Mm-browsecomp: A comprehensive benchmark for multimodal browsing agents. *arXiv preprint arXiv:2508.13186*, 2025.

[34] Siheng Li, Zhanhui Zhou, Wai Lam, Chao Yang, and Chaochao Lu. Repo: Replay-enhanced policy optimization. *arXiv preprint arXiv:2506.09340*, 2025.

[35] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[36] Yiming Liang, Ge Zhang, Xingwei Qu, Tianyu Zheng, Jiawei Guo, Xinrun Du, Zhenzhu Yang, Jiaheng Liu, Chenghua Lin, Lei Ma, et al. I-sheep: Self-alignment of llm from scratch through an iterative self-enhancement paradigm. *arXiv preprint arXiv:2408.08072*, 2024.

[37] Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Paolo Rota, Nicu Sebe, Zheng Liu, and Lizi Liao. Video-browsecomp: Benchmarking agentic video research on open web. *arXiv preprint arXiv:2512.23044*, 2025.

[38] Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Nicu Sebe, Zheng Liu, and Lizi Liao. Video-browser: Towards agentic open-web video browsing, 2026.

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

[40] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Qingwei Lin, Jianguang Lou, Shifeng Chen, Yansong Tang, and Weizhu Chen. Arena learning: Build data flywheel for llms post-training via simulated chatbot arena. *arXiv preprint arXiv:2407.10627*, 2024.

[41] Alisia Lupidi, Carlos Gemmell, Nicola Cancedda, Jane Dwivedi-Yu, Jason Weston, Jakob Foerster, Roberta Raileanu, and Maria Lomeli. Source2synth: Synthetic data generation and curation grounded in real data sources. *arXiv preprint arXiv:2409.08239*, 2024.

[42] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[43] OpenAI. Hello gpt-4o. *OpenAI Blog*, 2024.

[44] OpenAI. Gpt-5.1: Openai api model documentation. `https://developers.openai.com/api/docs/models/gpt-5.1`, 2025.

[45] OpenAI. Introducing openai o3 and o4-mini. `https://openai.com/blog`, 2025. OpenAI Blog.

[46] OpenAI. Introducing gpt-5.5. `https://openai.com/index/introducing-gpt-5-5/`, 2026.

[47] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *arXiv preprint arXiv:2506.07464*, 2025.

[48] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[49] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[50] J Rosser and Jakob Nicolaus Foerster. Agentbreeder: mitigating the ai safety impact of multi-agent scaffolds via self-improvement. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025.

[51] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[52] SerpApi. Serpapi: Google search api. `https://serpapi.com/`, 2026.

[53] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169, 2025.

[54] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

[55] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[56] Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*, 2024.

[57] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[58] Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025.

[59] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[60] Xiao Wang, Jianlong Wu, Zijia Lin, Fuzheng Zhang, Di Zhang, and Liqiang Nie. Video dataflywheel: Resolving the impossible data trinity in video-language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[61] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 92–101, 2024.

[62] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

[63] Zun Wang, Jialu Li, Yicong Hong, Songze Li, Kunchang Li, Shoubin Yu, Yi Wang, Yu Qiao, Yali Wang, Mohit Bansal, et al. Bootstrapping language-guided navigation learning with self-refining data flywheel. *arXiv preprint arXiv:2412.08467*, 2024.

[64] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

[65] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024.

[66] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, et al. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025.

[67] Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

[69] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

[70] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[71] 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, pages 9127–9134, 2019.

[72] Huaying Yuan, Zheng Liu, Junjie Zhou, Ji-Rong Wen, and Zhicheng Dou. Videodeepresearch: Long video understanding with agentic tool using. *arXiv e-prints*, pages arXiv–2506, 2025.

[73] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[74] Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, et al. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026.

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

[76] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*, 2024.

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

[78] Yongheng Zhang, Xu Liu, Ruihan Tao, Qiguang Chen, Hao Fei, Wanxiang Che, and Libo Qin. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5267–5276, 2025.

[79] Cen Zhao, Tiantian Zhang, Hanchen Su, Yufeng Zhang, Shaowei Su, Mingzhi Xu, Yu Liu, Wei Han, Jeremy Werner, Claire Na Cheng, et al. Agent-in-the-loop: A data flywheel for continuous improvement in llm-based customer support. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1919–1930, 2025.

[80] Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. Reagent-v: A reward-driven multi-agent framework for video understanding. *arXiv preprint arXiv:2506.01300*, 2025.

[81] Kefei Zhu, Fengshuo Bai, YuanHao Xiang, Yishuai Cai, Xinglin Chen, Ruochong Li, Xingtao Wang, Hao Dong, Yaodong Yang, Xiaopeng Fan, et al. Dexflywheel: A scalable and self-improving data generation framework for dexterous manipulation. *arXiv preprint arXiv:2509.23829*, 2025.

A Appendix

A.1 Experimental Results

Hyper-parameter Analysis. We conduct the ablation studies on key hyper-parameters in the trajectory-to-policy flywheel, including the scaling factor λ and the number of off-policy trajectories. As shown in Figure A1 (a), increasing λ from 0.5 to 1.2 leads to consistent performance improvements, indicating that stronger off-policy guidance helps improve policy learning. However, further increasing λ beyond 1.2 results in slight performance degradation, suggesting that overly large scaling factors may overemphasize off-policy signals and suppress useful on-policy exploration. Figure A1 (b) shows that increasing the number of off-policy trajectories significantly improves performance in the early stage, highlighting the benefit of incorporating more high-quality supervision. However, the gains quickly saturate beyond 4 trajectories, indicating diminishing returns as additional off-policy samples provide redundant information.

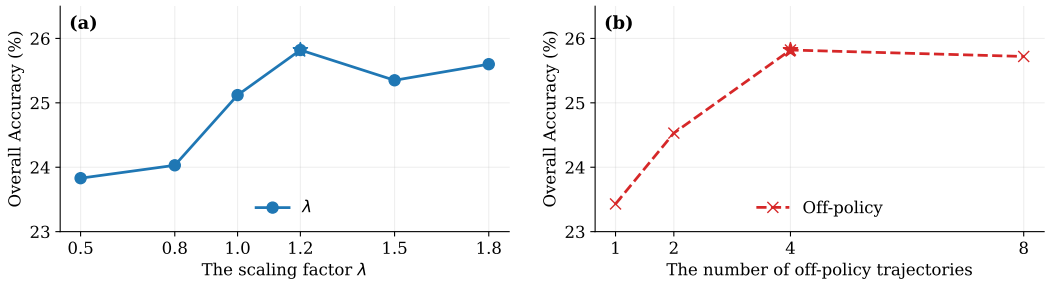


Figure A1: **Hyper-parameter ablations** including the scaling factor λ and the number of off-policy trajectories in our trajectory-to-policy flywheel.

Robust Analysis of Judge MLLMs. We replace the judge MLLM GPT-4o with GPT-5.5 and GPT-4o-mini, respectively. The results are shown in Table A1. We observe that VideoSailor maintains consistent performance improvements under different judge models, demonstrating its robustness to variations in judge quality. While stronger judges yield slightly better results, the overall gains remain stable even with weaker and more efficient models, indicating that the effectiveness of the trajectory-to-policy flywheel does not critically depend on a specific judge model.

Table A1: **Robust Analysis** of Judge MLLMs.

Judge MLLM	OA	SA	AVG CS
GPT-4o	25.82	11.96	22.74
GPT-5.5	**26.31**	**12.42**	**23.28**
GPT-4o-mini	25.21	11.38	22.11

Flywheel Filtering Dynamics. Table A2 reports the judge-based filtering statistics of on-policy rollouts across flywheel iterations. We observe that the rejected ratio decreases from 57.6% in Iter. 1 to 47.5% in Iter. 3, while the correct-and-valid ratio increases from 14.8% to 28.4%, indicating that the policy gradually produces higher-quality trajectories. In Iter. 4, the correct-and-valid ratio slightly drops to 27.6% and the rejected ratio increases to 48.6%, suggesting that the benefit of further flywheel iterations begins to saturate.

Table A2: Iteration-wise filtering statistics of on-policy rollouts.

Iteration	Correct-and-valid (%)	Correct-but-flawed (%)	Rejected (%)
Iter. 1	14.8	27.6	57.6
Iter. 2	21.7	26.9	51.4
Iter. 3	**28.4**	**24.1**	**47.5**
Iter. 4	27.6	25.8	48.6

Table A3: **Category-wise performance** on Omni-BrowseComp. We report Overall Accuracy (OA) and Strict Accuracy (SA), together with category-level results. For subtopics, Med., Geo., Tech., Acad., Ent., and Gen. denote the categories of "Media", "Geography", "Technology", "Academics", "Entertainment", and "General", respectively. **Bold** and underlined values denote the best and second-best performance in each column, respectively.

Model	OA (%)							SA (%)						
	Overall	Med.	Geo.	Tech.	Acad.	Ent.	Gen.	Overall	Med.	Geo.	Tech.	Acad.	Ent.	Gen.
Tool-Free MLLMs														
o4-mini	18.35	12.0	15.2	25.3	17.6	14.5	39.3	0.26	0.0	0.0	0.0	1.4	0.0	0.0
GPT-5.1	10.57	6.8	0.0	0.0	32.5	0.0	**61.7**	0.77	0.0	0.0	0.0	4.3	0.0	0.0
GPT-4o-2024-11-20	10.05	5.7	10.0	20.0	13.8	4.6	0.0	1.55	0.0	0.0	0.0	0.0	6.5	0.0
GPT-4o-mini	7.13	4.0	7.1	14.2	9.8	3.3	0.0	0.98	0.0	0.0	0.0	0.0	4.1	0.0
Claude Sonnet 4.5	8.53	3.0	0.0	0.0	19.3	9.6	38.2	1.81	0.0	0.0	0.0	7.6	1.9	0.0
Qwen2.5-VL-72B-Instruct	10.03	0.0	0.0	32.1	10.7	8.4	0.0	1.85	0.0	0.0	0.0	0.0	7.7	0.0
Qwen2.5-VL-32B-Instruct	9.18	0.0	0.0	29.3	9.8	7.7	0.0	1.54	0.0	0.0	0.0	0.0	6.4	0.0
Qwen2.5-VL-7B-Instruct	6.34	0.0	0.0	20.3	6.8	5.3	0.0	0.71	0.0	0.0	0.0	0.0	3.0	0.0
Tool-Augmented MLLMs														
GPT-5.5	24.33	20.4	17.9	0.0	24.8	44.8	47.0	1.10	1.7	0.0	0.0	2.9	1.3	0.0
o3	23.45	19.7	17.3	0.0	23.9	43.2	45.3	0.77	0.0	1.0	0.0	3.3	0.0	0.0
Gemini-2.5-Pro-Preview	17.48	**21.5**	0.0	0.0	44.5	26.0	0.0	5.14	**7.9**	0.0	0.0	**13.6**	6.2	0.0
Gemini-3-Flash-Preview	22.74	11.1	0.0	0.0	44.9	44.9	36.7	7.24	0.0	0.0	0.0	7.1	24.8	0.0
Open-source Agents														
Agent-R1														
Gemini-2.5-Flash-Preview	23.25	11.3	0.0	0.0	**46.0**	**46.0**	37.5	10.68	0.0	0.0	0.0	10.5	**36.6**	0.0
GPT-4o-2024-11-20	16.91	9.6	16.8	33.6	23.3	7.8	0.0	4.37	0.0	0.0	0.0	0.0	18.2	0.0
Qwen2.5-VL-72B-Instruct	14.39	0.0	0.0	**46.1**	15.3	12.0	0.0	4.15	0.0	0.0	0.0	0.0	17.3	0.0
OWL														
Gemini-2.5-Flash-Preview	20.14	9.8	0.0	0.0	39.8	39.8	32.5	8.73	0.0	0.0	0.0	8.6	29.9	0.0
GPT-4o-2024-11-20	12.37	7.0	12.3	24.6	17.0	5.7	0.0	3.18	0.0	0.0	0.0	0.0	13.2	0.0
Qwen2.5-VL-72B-Instruct	13.42	0.0	0.0	42.9	14.3	11.2	0.0	3.04	0.0	0.0	0.0	0.0	12.7	0.0
DeerFlow														
Gemini-2.5-Flash-Preview	18.37	9.0	0.0	0.0	36.3	36.3	29.6	6.14	0.0	0.0	0.0	6.0	21.1	0.0
GPT-4o-2024-11-20	11.45	6.5	11.4	22.7	15.8	5.3	0.0	2.53	0.0	0.0	0.0	0.0	10.5	0.0
Qwen2.5-VL-72B-Instruct	11.07	0.0	0.0	35.4	11.8	9.2	0.0	1.93	0.0	0.0	0.0	0.0	8.0	0.0
WebDancer														
Gemini-2.5-Flash-Preview	18.14	8.8	0.0	0.0	35.9	35.9	29.2	5.73	0.0	0.0	0.0	5.6	19.7	0.0
GPT-4o-2024-11-20	11.37	6.4	11.3	22.6	15.7	5.2	0.0	2.08	0.0	0.0	0.0	0.0	8.7	0.0
Qwen2.5-VL-72B-Instruct	11.08	0.0	0.0	35.4	11.8	9.3	0.0	1.97	0.0	0.0	0.0	0.0	8.2	0.0
WebDancer-32B	17.37	8.5	0.0	0.0	34.3	34.3	28.0	5.41	0.0	0.0	0.0	5.3	18.6	0.0
VideoSailor (Ours)	**25.82**	20.8	**18.5**	34.5	29.2	27.0	18.0	**11.96**	3.5	**5.0**	**4.8**	12.0	27.6	**14.0**

Category-wise performance on Omni-BrowseComp. Table A3 presents category-wise performance on the curated Omni-BrowseComp benchmark. We observe that most models exhibit highly uneven performance across categories, often achieving non-trivial accuracy only on a subset of domains while failing on others, as indicated by near-zero scores. This reflects their limited ability to generalize across diverse reasoning scenarios and their difficulty in handling tasks requiring multi-hop, cross-modal evidence aggregation. In contrast, VideoSailor achieves more balanced performance across categories and obtains the best overall OA and SA. In particular, VideoSailor achieves the best OA on Geography and the best SA on Geography, Technology, and General, while remaining competitive on Media and Academics. The overall superior results demonstrate that the proposed trajectory-to-policy flywheel improves not only final-answer accuracy but also strict end-to-end reasoning robustness across diverse open-world video research scenarios.

Training Stability Diagnostics. To incorporate the off-policy guidance, a native mixing approach is to treat off-policy trajectories equally with on-policy samples when computing advantages (*c.f*. Eq (A3)). Figure A2 compares the training dynamics of native mixing (*c.f*. Eq A3) and our calibrated advantage strategy. As shown in Figure A2 (a), although native mixing improves accuracy in the early stage, its performance quickly saturates, while our proposed calibrated advantage achieves consistently higher and more stable accuracy.

In addition, we visualize the *skewness* of advantage distributions, which is defined as the third standardized moment:

$$\text{Skewness}(A) = \mathbb{E}\left[\left(\frac{A-\mu}{\sigma}\right)^3\right], \tag{A1}$$

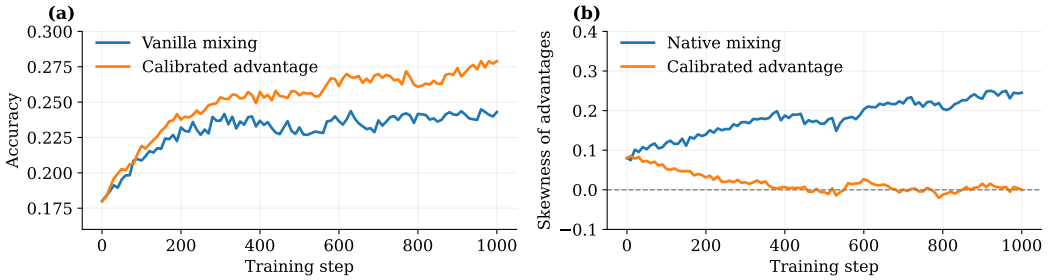


Figure A2: **Training stability comparison** between native mixing and calibrated advantage in terms of accuracy and advantage skewness.

where μ and σ are the mean and standard deviation of the advantage values. Skewness characterizes the asymmetry of the distribution: a positive value indicates that large positive advantages occur more frequently (*i.e.*, a heavy right tail), while a negative value suggests the opposite. This provides additional insight beyond variance into how learning signals are distributed across samples.

As shown in Figure A2 (b), native mixing exhibits persistent positive skewness, indicating that high-reward off-policy trajectories dominate the shared normalization statistics and bias policy updates. In contrast, calibrated advantage keeps the skewness close to zero by anchoring off-policy trajectories to the on-policy advantage scale, thereby preserving useful off-policy guidance while maintaining balanced and stable optimization.

A.2 Experimental Setups

Training Configurations. We summarize the full set of training configurations in Table A4.

Tool Configurations. To improve reproducibility, we use a fixed tool repository for all training and evaluation experiments, including text search, image search, video search, and ASR. Text search is implemented with Google Search via SerpAPI [52] and returns the top-5 organic results with titles, URLs, and snippets. Image search uses the same backend in image-search mode and returns the top-5 visually relevant results with thumbnails and source URLs. Video search queries YouTube through SerpAPI and retrieves the top-5 candidate videos with titles, URLs, channel names, publication dates, and descriptions. Candidate videos are accessed only when selected by the policy for further evidence extraction. ASR is performed on the selected video segment using Whisper-large-v3 [48], with transcripts restricted to the timestamp range specified by the tool call. To reduce open-web stochasticity, we cache all search results and ASR outputs after their first execution and reuse them in subsequent SFT, RL, and evaluation runs. We disable personalized search, use a fixed locale and language setting, and exclude login-gated, paid, or private content. Each trajectory can invoke at most five tools, and each search call returns at most five results. Transient API failures are retried automatically while persistent failures are treated as empty observations.

Evaluation Metrics. We follow MM-BrowseComp [33] to adopt three evaluation metrics: *1) Overall Accuracy (OA)*, which measures the correctness of final answers; *2) Strict Accuracy (SA)*, which requires both correct answers and complete reasoning according to the checklist; and *3) Average Checklist Score (AVG CS)*, which evaluates the average completion rate of checklist steps as a fine-grained measure of reasoning quality.

Prompts. For the seed dataset generation (*c.f.* Sec 3.1), the prompts for unified captioning, multi-hop query generation, and quality control are presented in Figure A6, Figure A5, and Figure A7, respectively. For the evaluation metric computation on Omni-BrowseComp, the judge prompt is provided in Figure A8. For Video-BrowseComp [38], we follow the official prompts.

Illustrations of Naive Mixing. A straightforward strategy for mixed-policy optimization is to treat on-policy and off-policy trajectories equally when computing advantages. Specifically, given the union of on-policy trajectories $\mathcal{T}^{\text{on}}$ and off-policy trajectories $\mathcal{D}_i$, the advantage for each trajectory $\tau \in \mathcal{T}^{\text{on}} \cup \mathcal{D}_i$ is computed using shared statistics:

$$A_{\text{naive}}(\tau) = \frac{r(\tau) - \mu_{\text{mix}}}{\sigma_{\text{mix}}}, \quad \tau \in \mathcal{T}^{\text{on}} \cup \mathcal{D}_i, \tag{A2}$$

Table A4: **Training configurations** for SFT and RL stages.

Configuration	SFT	RL
Trainable	Full Model	Full Model
Base Model	Qwen2.5-VL-7B-Instruct	SFT-initialized Model
Training Samples	3,129	7,032
Per-device Batch Size	1	1
Gradient Accumulation	1	1
Global Batch Size	8	8
Epoch	2	1
Optimizer	AdamW	AdamW
DeepSpeed	ZeRO-3 Offload	ZeRO-3 Offload
Learning Rate	5×10^{-6}	1×10^{-6}
LR Schedule	Cosine	Cosine
Warmup Ratio	0.03	0.03
Weight Decay	0.01	0.01
Max Sequence Length	8192	8192
Rollout Group Size	–	8
GPU Nums	8 A100 (80 GB)	8 A100 (80 GB)

where μ_{mix} and σ_{mix} denote the mean and standard deviation of rewards computed over the combined set $\mathcal{T}^{\text{on}} \cup \mathcal{D}_i$.

The policy is then optimized using the standard GRPO objective:

$$\mathcal{L}_{\text{naive}}(\theta) = \frac{1}{|\mathcal{T}^{\text{on}}| + |\mathcal{D}_i|} \sum_{\tau \in \mathcal{T}^{\text{on}} \cup \mathcal{D}_i} \min \Bigg(\frac{\pi_\theta(\tau \mid x)}{\pi_{\theta_{\text{old}}}(\tau \mid x)} A_{\text{naive}}(\tau), \\ \text{clip}\Big(\frac{\pi_\theta(\tau \mid x)}{\pi_{\theta_{\text{old}}}(\tau \mid x)}, 1-\epsilon, 1+\epsilon \Big) A_{\text{naive}}(\tau) \Bigg) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}). \tag{A3}$$

Limitations. Despite the promising performance, VideoSailor has several limitations. First, the trajectory-to-policy flywheel relies on an external LLM-based judge for trajectory refinement, whose quality directly affects the constructed off-policy supervision and may introduce additional computational overhead. Second, the effectiveness of the framework depends on the quality of on-policy rollouts in early iterations; when the initial policy is weak, the generated trajectories may be noisy, limiting the benefits of subsequent refinement. Finally, our method focuses on improving reasoning through data refinement and policy optimization, while the integration of more advanced retrieval systems and long-context modeling remains an important direction for future work.

Broader Impacts. This work has positive societal impacts by improving open-world video reasoning, which can benefit applications such as education, research, and assistive technologies. The proposed framework and benchmark also promote more reliable and standardized evaluation of multimodal reasoning systems.

Potential risks include misuse of cross-video and multimodal reasoning for large-scale information aggregation or surveillance, as well as the amplification of misinformation or biases from external sources. We mitigate these risks by focusing on publicly available data and research settings, and we encourage future work on robustness, factual grounding, and responsible deployment.

Safeguard Discussion. The Omni-BrowseComp benchmark is constructed using publicly available data sources, including videos, images, audio, and text, without collecting or introducing private or sensitive personal information. During dataset curation, we carefully filter out inappropriate, harmful, or unsafe content to reduce potential risks. Potential misuse may arise from large-scale cross-video retrieval and multimodal reasoning, which could be applied to unintended information aggregation or analysis. To mitigate such risks, we position Omni-BrowseComp as a research-only benchmark and do not release any raw scraped data that could raise safety concerns. We also encourage responsible use of the dataset and emphasize that it should not be deployed in high-risk or sensitive applications without additional safeguards.

License Discussion. Our training and evaluation datasets are constructed from publicly available sources, including existing datasets (*e.g.*, ActivityNet-QA [71], How2QA [32], Video SimpleQA [6]), as well as publicly accessible web content. We ensure that all original datasets and sources are properly cited, and we follow their respective licenses and terms of use.

For the constructed dataset (Omni-BrowseComp and the seed training data), we do not redistribute raw media content. Instead, we release only processed annotations, metadata, and task formulations (*e.g.*, questions, answers, reasoning checklists, and temporal segments), which reference the original sources. Users are required to obtain access to the underlying data from the original providers in accordance with their licenses. Since the dataset is derived from multiple sources with potentially different licenses, we provide references and links to the original datasets where applicable. Our released annotations are intended for research use only and comply with the terms of the original data sources to the best of our knowledge.

A.3 Additional Related Work

On-Policy and Off-Policy RL. RL methods are typically distinguished by how training data is generated and reused during policy optimization. On-policy methods (*e.g.*, GRPO [21], REINFORCE [55], PPO [51]) update the policy using trajectories sampled from the current policy, ensuring distributional consistency, but often at the cost of low sample efficiency. In contrast, off-policy methods (*e.g.*, DQN [42], SAC [22], DPO [49]) decouple data collection from policy learning, enabling reuse of historical data and thus improving sample efficiency, while introducing optimization challenges due to distribution mismatch between behavior and target policies. The recent works [34, 69, 27] have explored hybrid policy optimization to combine the strengths of both paradigms. RePO [34] retrieves off-policy samples from a replay buffer, allowing policy optimization based on more diverse samples. LUFFY [69] augments on-policy RL with off-policy reasoning traces to dynamically balance imitation and exploration. For the first time, our proposed trajectory-to-policy flywheel extends this mix-policy optimization into the data flywheel paradigm, which provides a scalable pathway for consistent improvement for video deep research systems.

A.4 Visualization

Visualization of Omni-BrowseComp Benchmark. Figure A4 presents representative examples from our constructed Omni-BrowseComp benchmark. Each instance requires multi-hop reasoning across heterogeneous evidence sources, including images, videos, audio, and texts. The provided checklist decomposes the reasoning process into sequential steps, where the model must iteratively retrieve, verify, and integrate evidence from multiple modalities and, in some cases, across different videos. These examples highlight the challenges of omni-modal evidence aggregation, cross-video interaction, and fine-grained temporal grounding.

Visualization of Seed Training Dataset. Figure A3 illustrates representative examples from the constructed seed training dataset (*c.f.* Sec 3.1). Each instance is formulated as a multi-step reasoning trajectory that integrates heterogeneous evidence sources. Starting from a user query, the model progressively retrieves and verifies relevant information through a sequence of reasoning steps, where each step is grounded in a specific modality via tool usage (*e.g.*, image search, text search, video search, and ASR). The trajectory explicitly captures the reasoning process in a structured `<think>`–`<tool>` format, culminating in a final answer derived from compositional evidence aggregation.

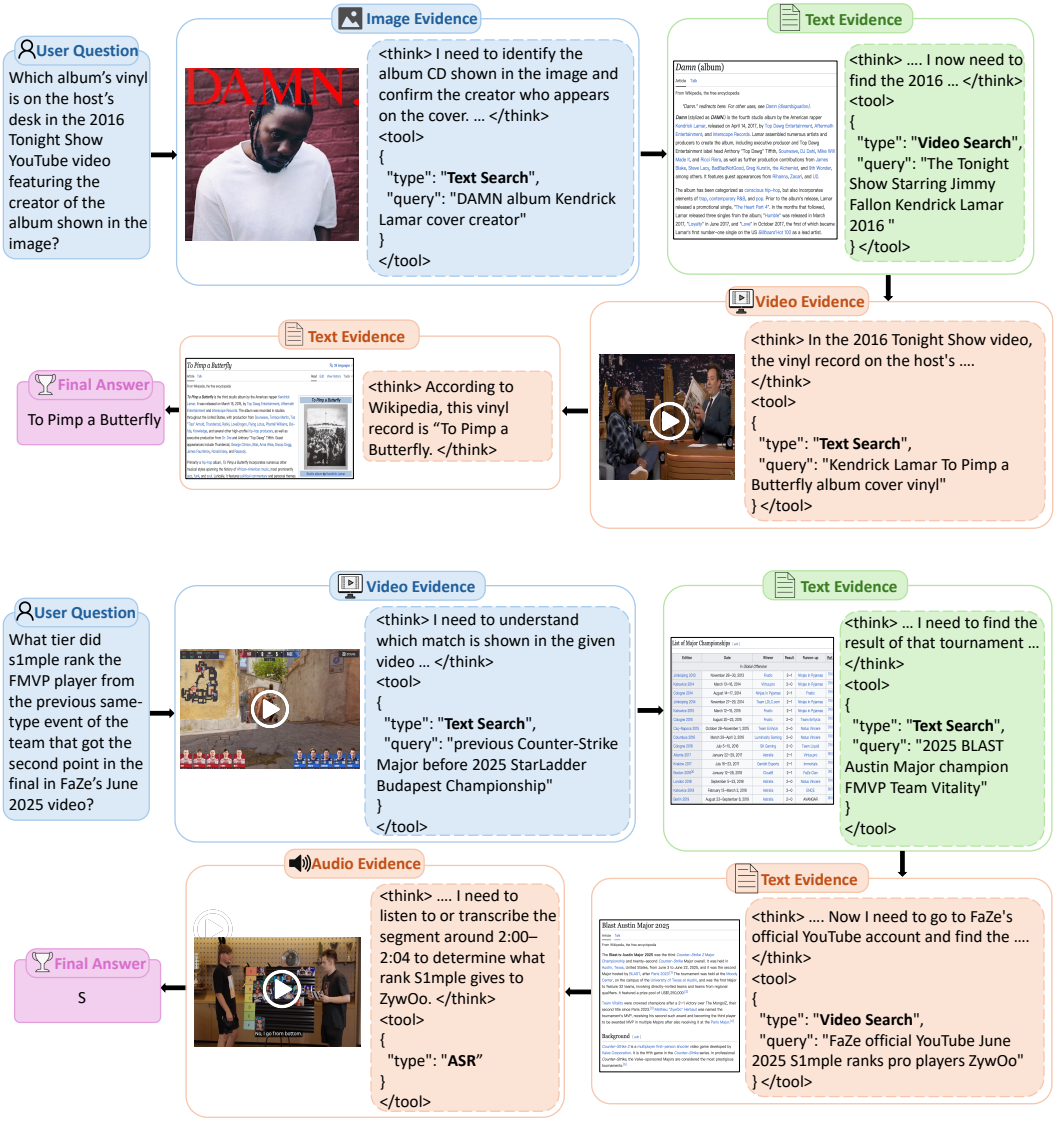


Figure A3: **Representative cases** from the seed training dataset.

Input with Image



Question: The actress in the image starred in a Greta Gerwig film. Using the official trailer and Oscar evidence, what was the film's only win at the 92nd Academy Awards?

Checklist:

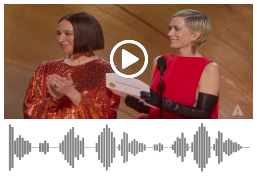
1. By inspecting the image, the actress is identified as Saoirse Ronan.
2. By searching, confirm that Saoirse Ronan has starred in a film directed by Greta Gerwig.
3. By watching the official trailer, confirm that the film is Little Women.
4. By listening to the acceptance speech, the award is identified as Best Costume Design.

Ground Truth: Best Costume Design





Time: [03:34, 03:45]



Time: [11:45, 12:20]

Question: In the official highlights of the Rugby World Cup 2023 final, identify the winning side. Then find the World Rugby video highlighting a player from that side who made 28 tackles in the final. What award did that player receive after the match?

Checklist:

1. Watch the official Rugby World Cup 2023 final highlights and identify the match as New Zealand vs South Africa.
2. Determine that South Africa won the final.
3. Use the winning side clue to infer that the target player should be a Springbok/South Africa player.
4. Find World Rugby's 28-tackle video and identify the highlighted player as Pieter-Steph du Toit.
5. Verify that the award received by that player after the match was Mastercard Player of the Match.

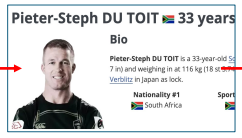
Ground Truth: Mastercard Player of the Match



Time: [51:22, 52:43]



Time: [01:40, 01:55]





Time: [07:34, 07:42]

Question: In NASA's video introducing the commander of Artemis II, identify the astronaut. According to NASA's official biography and ISS mission evidence, which Expedition did he previously serve on as a flight engineer?

Checklist:

1. Watch NASA's Artemis II commander introduction video and identify the astronaut as Reid Wiseman.
2. Confirm that Reid Wiseman served as the commander of NASA's Artemis II mission.
3. Search NASA's official biography for Reid Wiseman's previous International Space Station experience.
4. Find that he previously served as a flight engineer aboard the International Space Station.
5. Verify that the Expedition associated with this flight engineer role was Expedition 41.

Ground Truth: Expedition 41



Time: [01:14, 04:39]





Figure A4: **Representative cases** from the curated Omni-BrowseComp benchmark.

```
prompt_multihop_generation = """
You are given several multimodal content summaries from the same domain. Each item may be an
image, a video, or an audio clip.

Your task is to generate one multimodal multi-hop QA sample.

Requirements:
1. The final question must include one image or video as the question-side modality input.
2. The question-side modality should provide only a starting clue, not the final answer.
3. The final answer must require reasoning through the intermediate hops.
4. The answer must NOT be directly obtainable from the question-side modality alone.
5. Each hop should contribute necessary new information.
6. All items must stay within the same domain and form a natural reasoning path.
7. The question text should be natural and should not reveal the intermediate findings explicitly.
8. Do not explicitly explain how the chain was designed. Just produce the final QA sample with the
intermediate hops.

Input:
{summaries_block}

Output JSON only:
{
  "question_text": "...",
  "question_modality": {
    "type": "image or video",
    "url": "..."
  },
  "answer": "...",
  "hops": [
    {
      "step": 1,
      "query": "...",
      "source_url": "...",
      "finding": "...",
      "next_trigger": "..."
    }
  ]
}
"""
```

Figure A5: **The prompt for multi-hop query generation** in the seed dataset.

```
prompt_caption = """
You are given one piece of multimodal content, which may be an image, a video, or an audio clip.

Your task is to produce a concise semantic description that will later be used for multi-hop question
construction.

Requirements:
1. Identify the main topic of the content.
2. Write a concise summary of the most important information in the content.
3. The summary should be informative enough for downstream multi-hop data construction.
4. Keep the output compact and factual.

Output JSON only:
{
  "topic": "...",
  "summary": "..."
}
"""
```

Figure A6: **The prompt for unified captioning** in the seed dataset generation.

```
prompt_quality_control = """
You are given one multimodal multi-hop QA sample.

Decide whether it meets the required standard.

The sample is valid only if ALL of the following are true:
1. The question contains one image or video as the input modality.
2. That input modality serves only as a starting clue.
3. The final answer is not directly visible or inferable from the input modality alone.
4. Removing the intermediate hops would make the question unanswerable.
5. Each hop contributes essential new evidence.
6. The hops form a coherent multi-hop reasoning process.
7. All sources are from the same domain.
8. The question text does not reveal the answer or the hidden intermediate findings.

Output only:
Yes
or
No
"""
```

Figure A7: **The prompt for quality control** in the seed dataset generation.

```
prompt_evaluation = """
You are an AI evaluator. Your task is to evaluate the quality of an answer. I will provide you with the
user's question, the reference answer, a checklist, and the answer to be evaluated.
— USER QUESTION —
{question}
— REFERENCE ANSWER —
{reference_answer}
This reference answer is considered the correct and ideal response content-wise.
— REFERENCE CHECKLIST —
{checklist_items_formatted}
— MODEL'S GENERATED ANSWER TO EVALUATE —
{generated_answer_to_eval}
— EVALUATION INSTRUCTIONS —
Please provide your evaluation strictly in the following format on separate lines:
1. Checklist Score: First, determine how many of the {total_checklist_items} items in the 'REFERENCE
CHECKLIST' have been correctly and completely addressed by the 'MODEL'S GENERATED ANSWER TO
EVALUATE'. Please remember that for any item in the checklist, the model's generated answer to
evaluate must fully comply in order for that item to be considered complete.
State this as 'CHECKLIST_SCORE: [correct_items]/{total_checklist_items}' (e.g., CHECKLIST_- SCORE: 2/3).
2. Checklist Result Vector: Next, please provide a 0-1 vector to indicate whether each checklist item
passed. Output the vector in the order of the items in the checklist, for example, [1,0,1]. '1' means the
item is 'fully satisfied,' and '0' means 'not fully satisfied.' If there is no checklist for this question, please
return N/A. Output in the format 'CHECKLIST_RESULT: ...' (e.g., CHECKLIST_RESULT: [1,0,1]).
3. Overall Correctness: Next, you need to judge whether the 'MODEL'S GENERATED ANSWER TO
EVALUATE' is consistent with the 'REFERENCE ANSWER (Ground Truth)' in terms of its core content and
information.
- Content consistency is key. Differences in formatting or minor wording variations are acceptable as long
as the essential information and meaning conveyed by the generated answer align with the reference
answer.
- If the generated answer accurately reflects the information in the reference answer, it should be
considered correct. State your judgment as 'OVERALL_CORRECTNESS: [YES/NO]' (e.g.,
OVERALL_CORRECTNESS: YES).
Example 1 (Checklist provided, generated answer consistent with reference, some checklist items
missed):
CHECKLIST_SCORE: 1/3
CHECKLIST_RESULT: [1,0,1]
OVERALL_CORRECTNESS: YES
Example 2 (Checklist provided, generated answer NOT consistent with reference, even if the checklist is
met):
CHECKLIST_SCORE: 4/4
CHECKLIST_RESULT: [1,1,1,1]
OVERALL_CORRECTNESS: NO
Provide only these formatted lines (CHECKLIST_SCORE, CHECKLIST_RESULT, OVERALL_CORRECTNESS) as
your response.
"""
```

Figure A8: **The judge prompt** for our Omni-BrowseComp benchmark.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are supported by the methodological design in Sec 3, the benchmark design in Sec 4 and extensive experimental results in Sec 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The discussion of the limitations is presented Sec A.2.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details are provided in Sec 5.1 and Table A4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code is provided in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details are provided in Sec 5.1, Figure 5 and Table A4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the high computational cost and long runtime of the experiments, we do not report statistical significance such as error bars. Each result is obtained from a full training and evaluation run under fixed settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute resources in Table A4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work complies with the NeurIPS Code of Ethics. It focuses on algorithmic design and benchmark construction for video reasoning, and does not involve human subjects, sensitive personal data, or high-risk applications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts are discussed in Sec A.2.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The discussion of safeguards is provided in Sec A.2.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The licenses discussion is provided in Sec A.2.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The documentation is provided in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Large language models are used as core components of the proposed method. Specifically, LLMs are employed for trajectory generation, LLM-based judging and refinement, and data construction in the trajectory-to-policy flywheel.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.