

LongVT: Incentivizing “Thinking with Long Videos” via Native Tool Calling

Supplementary Material

Outline

This Supplementary Material complements the main paper, providing comprehensive experimental details, in-depth analyses of training dynamics, and extensive qualitative visualizations. The content is organized as follows:

- **Strategic Alignment & Motivation.** We first demonstrate the conceptual alignment between LongVT and state-of-the-art proprietary large multimodal models (LMMs) in Section 1. Subsequently, we present a rigorous data contamination study in Section 2 to underscore the necessity of our proposed VideoSIAH-Eval benchmark, followed by detailed statistics of the curated dataset in Section 3.
- **Formulation & Training Dynamics.** We present the overall framework illustration in Figure 3 and elaborate on the theoretical formulations of our training objectives in Section 4 for both supervised fine-tuning (SFT) and reinforcement learning (RL). Crucially, in Section 5, we visualize the “economy of thinking”—a distinct evolutionary trajectory where the model learns to internalize tool usage. Section 6 then provides the exact hyperparameters and infrastructure details for reproducibility.
- **Efficiency & Qualitative Analysis.** We report a detailed inference latency comparison in Section 7, countering the intuition that multi-turn agentic frameworks are inherently slower. In Section 8, we provide prompt templates, diverse qualitative examples, and workflow demonstration, while Section 9 analyzes specific failure modes to highlight the importance of the cold-start training stage.
- **Discussion.** Finally, we discuss the architectural limitations and future multi-agent directions in Section 10, followed by a discussion on the broader impact and ethical considerations in Section 11 and Section 12, respectively.

1. LongVT Performs Human-Aligned Thinking like Leading Proprietary LMMs

The core philosophy of our proposed interleaved Multimodal Chain-of-Tool-Thought (iMCoTT) entails a “global-to-local” thinking pattern: the model first performs a coarse skim to formulate a hypothesis, and subsequently invokes the native `crop_video()` tool to inspect specific temporal windows for fine-grained verification. While this design was inspired by human intuition, we observe a striking convergence between our approach and the reasoning behaviors emerging in state-of-the-art proprietary LMMs when they are prompted to perform fine-grained analysis.

To validate this alignment, we queried two leading models, Gemini 2.5 Pro [3] and GPT-5 Thinking [15], regarding their optimal strategies for analyzing fine-grained video

details. As illustrated in Figure 1a, Gemini 2.5 Pro explicitly advocates for a two-stage process: a “Step 1: Coarse Scan” to efficiently locate the general event (e.g., searching for scoreboard changes or crowd reactions), followed by a “Step 2: Fine Scan” to isolate the exact moment and verify details (e.g., scrubbing back 30-60 seconds). This directly mirrors the workflow of our proposed LongVT, where the “Coarse Scan” corresponds to our global preview stage, and the “Fine Scan” is functionally identical to our agentic `crop_video()` tool calling. Similarly, Figure 1b demonstrates that the GPT-series model adopts a hierarchical “Coarse→Medium→Fine” search strategy. These examples confirm that the “Thinking with Long Videos” paradigm we propose in this work is a natural and necessary evolution for reliable long-form video reasoning, given that such human-aligned reasoning capabilities are currently exclusive to top-tier proprietary models.

2. What Motivates VideoSIAH? Unveiling the Data Contamination in Qwen-VL Series

With the rapid advancements of LMMs, model performance on various benchmarks has steadily improved. However, the “black-box” nature of training data raises a critical question: *Do these improvements reflect genuine reasoning capability, or are they partly due to the model memorizing the benchmark samples?* To investigate this, we conduct a rigorous contamination study on the Qwen-VL series [1, 2] across two probing settings: (1) No Visual, where we feed the text prompt without video frames to test for direct memorization; (2) Rearranged Choices, where we randomize the mapping between option labels and their textual content (e.g., assigning the original answer A to B) for multiple-choice questions (MCQs) to detect label memorization.

Our experimental results reveal significant vulnerabilities in existing benchmarks and highlight the necessity of our proposed VideoSIAH-Eval: *Observation 1: “No Visual” Performance Indicates Severe Leakage in Existing Benchmarks.* As shown in Table 1, both Qwen2.5-VL and Qwen3-VL achieve remarkably high scores on VideoMME and VideoMMU even without seeing any video frames. Notably, for VideoMME, we specifically evaluate without subtitles to ensure there is no textual leakage, yet Qwen2.5-VL still achieves 40.1%, far exceeding random guessing (~25%) for such four-option MCQs. Similar patterns of potential data leakage are observed on VideoMMU. While the ‘No Visual’ scores of 38.3% (Comprehension) and 39.3% (Perception) might appear similar to VideoMME, they are statistically more improbable given the dataset composition.

◆ This is a classic video analysis challenge. Since you are limited to only the video file and its player, you cannot search for timestamps or stats. Your method must be based entirely on visual scanning.

Here is the fastest way to find that specific detail.

Step 1: The "Coarse" Scan (Find the Goal)

Do not watch the video from the beginning. You need to "scrub" the timeline efficiently. Use your mouse to drag the video's playhead (slider) back and forth, or use a player that shows thumbnail previews as you hover over the timeline.

You are looking for one of three key visual cues that only happen after a goal:

- The Scoreboard Graphic:** This is often the fastest cue. Look for the on-screen graphic (usually in a corner) to change. Since it's an equalizer, you're looking for the moment the score becomes tied (e.g., 0-1 changes to 1-1, or 1-2 changes to 2-2).
- The Celebration:** Look for a sudden, massive change in player behavior. Specifically, look for the French players (in their blue or white kits) running together, hugging, sliding, or running to the corner flag. This is a very obvious visual break from the normal back-and-forth of the match.
- The Center Kickoff:** After every goal and celebration, the game resets with a kickoff from the center circle. If you scrub and see the teams lined up in their own halves for a kickoff (and it's not the start of a half), a goal just occurred.

Step 2: The "Fine" Scan (Isolate the Event)

Once you spot one of those cues (e.g., you see a celebration at the 68:30 mark), stop scrubbing.

- Rewind:** Go back about 30-60 seconds before the celebration.
- Play:** Watch at normal speed.
- Verify:** You are confirming three [Save this chat](#)

(a) Watching Strategy of Gemini 2.5 Pro.

Coarse (locate the equalizer in minutes, not seconds)

- Skim by the scoreboard bug**
 - Play at 1.5-2x and scrub in ~30-60s jumps.
 - Watch the on-screen score (usually top corner). When it changes and becomes level (e.g., 0-1 → 1-1), and France's number just increased, you've found the equalizer window.
- Use goal sequence cues when the bug isn't visible**
 - Big crowd roar → commentator spike → players celebrate → replay wipe → opponent kicks off from the center circle.
 - That restart confirms France scored; if the new score is level, it's the equalizer.
- Quick hotspots**
 - Check just after 45:00 and 90:00: broadcasts often show a recap of all goals (with slow-mo angles). Note the equalizer's timestamp, then jump back to the live moment.

Medium (lock the exact play and contact)

- Back up ~30-60s before the score change and play normal speed.
- Watch the replay package** (usually 2-3 angles): wide → close → reverse/behind-goal.
 - If your player allows frame-step (←/→ or ,/.), use it to land on boot-ball contact.
- Confirm it's a volley**
 - Volley: ball is struck before touching the ground.
 - Half-volley: the ball kisses/bounces off the turf immediately before contact. Look for a tiny compress/skip on the grass and a micro "thud" in audio.

Fine (decide left vs right foot, and double-check)

- Identify the striking foot**
 - The plant foot is on the ground; the striking foot swings through the ball—that's your left/right answer.
 - Use the reverse/behind-goal angle if the live angle is occluded.
- Freeze at contact and verify

(b) Watching Strategy of GPT-5 Thinking.

Figure 1. Comparison of Watching Strategies Proposed by Gemini 2.5 Pro [3] and GPT-5 Thinking [15]. Best viewed when zoomed in.

Setting	VideoMME [6]	VideoMMMU [7]		VideoSIAH-Eval	
	w/o subtitle	adaptation*	comprehension	perception	test
Qwen2.5-VL-7B-Instruct [2]					
Original	64.3	35.7	44.3	<u>54.7</u>	33.8
No Visual	40.1	27.0	38.3	39.3	<u>12.7</u>
Rearranged Choices	<u>56.0</u>	<u>31.6</u>	<u>40.3</u>	67.0	-
Qwen3-VL-8B-Instruct [1]					
Original	69.3	40.7	60.3	71.3	46.6
No Visual	44.1	35.1	39.3	46.7	<u>0.00</u>
Rearranged Choices	<u>69.0</u>	<u>38.7</u>	<u>47.7</u>	<u>69.3</u>	-

Table 1. Contamination Tests for Qwen-VL Series on Long Video Understanding and Reasoning Benchmarks. Results are reported across different perturbation settings. The best result in each block column is in **bold**, and the second-best is underlined. The VideoSIAH-Eval column shows “-” entries for Rearranged Choices since our proposed benchmark is fully open-ended QA, where random option-answer mapping is not applicable. Note that adaptation* is evaluated exclusively on multiple-choice questions.

Our statistics reveal that these subsets are overwhelmingly dominated by MCQs with 10 options (e.g., 286 out of 300 for Comprehension and 279 out of 300 for Perception), implying a random guessing baseline of only $\sim 10-16\%$. The fact that the model achieves scores significantly above this threshold absent any visual context indicates a high probability of benchmark memorization. In contrast, performance on VideoSIAH-Eval drops significantly in the “No Visual” setting. Specifically, Qwen3-VL collapses to a score of 0.00. Upon manual inspection, we find that without visual grounding, the model generates repetitive code or refusal messages, which is the expected behavior for a clean and non-contaminated benchmark. *Observation 2: “Rearranged Choices” Reveals Overfitting to Option Patterns.* For MCQ-

based benchmarks, we observe distinct performance drops when answer choices are rearranged. For instance, Qwen2.5-VL drops from 64.3 to 56.0 on VideoMME. This indicates that they heavily rely on memorizing specific option mappings (e.g., the answer to this question is usually “A”) rather than understanding the content. Since VideoSIAH-Eval utilizes a fully open-ended QA format, it is inherently immune to this type of option hacking, providing a more robust assessment of the model’s capabilities.

These findings confirm that existing benchmarks are compromised by data contamination (high “No Visual” scores) and option bias (sensitive to “Rearranged Choices”). This motivates the introduction of VideoSIAH-Eval, which ensures: (1) *Zero leakage* as verified by the 0.00 blind score,

and (2) *Immunity to option bias* via open-ended QA format.

3. Additional VideoSIAH Details

Source	Purpose	Samples
LLaVA-CoT [19]	General Visual Reasoning	54,591
OpenVLThinker [4]	Complex Reasoning	2,829
We-Math 2.0 [12]	Mathematical Reasoning	602

Table 2. Detailed Statistics of Image-based CoT Data for Cold-Start SFT.

Breakdown of Image-based CoT Data. As detailed in Table 2, we construct a diverse mixture of image-based CoT data for the cold-start SFT stage, spanning general visual reasoning [19], complex logical inference [4], and mathematical problem-solving [12]. Drawing on insights from recent work [5, 21], we leverage these image-based reasoning traces to strengthen the model’s fundamental perceptual capabilities. This strategy exploits the inherent synergy between image and video modalities, where robust spatial grounding serves as a critical foundation for complex temporal reasoning.

Category Distribution for VideoSIAH-Eval. VideoSIAH-Eval comprises 244 videos and 652 high-quality QA pairs. As illustrated in Figure 2a, the video corpus encompasses a diverse spectrum of domains, ranging from Travel & Events to Gaming, ensuring broad coverage of real-world scenarios. Furthermore, Figure 2b highlights our deliberate emphasis on dynamic video reasoning: Action Recognition and Temporal Reasoning (17% in total) constitute a large portion of queries, rigorously benchmarking the model’s capacity for fine-grained event perception and causal understanding in the temporal dimension.

4. Additional Methodological Details

Next-Token Prediction. During SFT, we train our model by minimizing the negative log-likelihood of the target tokens given their preceding context. For a sequence of tokens $x = (x_1, x_2, \dots, x_T)$ and a model parameterized by θ that defines conditional probabilities $p_\theta(x_t | x_{<t})$, the loss function is defined as

$$\mathcal{L}(\theta) = - \sum_{t=1}^T \log p_\theta(x_t | x_{<t}),$$

which encourages the model to assign higher probability to the ground-truth next token.

Group Relative Policy Optimization. During RL, we adopt GRPO [13] for optimization. For each prompt $x \in \mathcal{D}$, we draw a group of K responses from the behavior policy

$\pi_{\theta_{\text{old}}}$.

$$y^{(k)} \sim \pi_{\theta_{\text{old}}}(\cdot | x), \quad k = 1, \dots, K,$$

$$y^{(k)} = (y_1^{(k)}, \dots, y_{T_k}^{(k)}), \quad T_k = \text{len}(y^{(k)}).$$

We use a group baseline and advantages:

$$b = \frac{1}{K} \sum_{k=1}^K R^{(k)}, \quad A^{(k)} = R^{(k)} - b,$$

where $R^{(k)}$ is the scalar return of response $y^{(k)}$.

The policy maximizes a length-normalized, token-conditional KL-regularized objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{\substack{\{y^{(k)}\} \sim \pi_{\theta_{\text{old}}}(\cdot | x) \\ x \sim \mathcal{D}}} \left[\frac{1}{K} \sum_{k=1}^K \frac{1}{T_k} \sum_{t=1}^{T_k} A^{(k)} \log \pi_\theta(y_t^{(k)} | x, y_{<t}^{(k)}) \right]$$

$$- \beta \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{K} \sum_{k=1}^K \frac{1}{T_k} \sum_{t=1}^{T_k} D_{\text{KL}}(\pi_\theta(\cdot | x, y_{<t}^{(k)}) \| \pi_{\text{ref}}(\cdot | x, y_{<t}^{(k)})) \right], \quad (1)$$

with $t \in \{1, \dots, T_k\}$, π_{ref} a frozen reference policy, and $\beta > 0$ controlling KL strength.

5. Reflection Trajectory: From Verbose Self-Correction to Internalized Tool Usage

We visualize the evolution of the model’s internal thought process in Figure 4 (left). Echoing the training dynamics observed in DeepEyes [24], the trajectory of reflection token proportion discloses a distinct three-phase evolution from exploratory correction to efficient tool exploitation: (1) *Verbose Self-Correction (Steps 0~50)*: Initially, reflection density remains high. Due to insufficient localization accuracy, the model relies on extensive self-correction and iterative verbal reasoning to compensate for sub-optimal tool usage. (2) *Efficiency Optimization (Steps 50~80)*: A significant drop follows as the policy matures. As the model’s intrinsic grounding capability improves, it identifies prolonged reflection to be redundant, autonomously pruning unnecessary linguistic fillers to maximize reward efficiency. (3) *Internalized Proficiency (After 80 Steps)*: The curve stabilizes at a concise baseline, indicating a shift toward selective reasoning—the model invokes explicit reflection only when resolving ambiguity, having internalized the core semantics of tool interaction. Complementing this, the word cloud (right) confirms that the remaining reflection tokens are semantically grounded (e.g., “segment,” “confirm”), serving as functional anchors for temporal reasoning rather than generating generic linguistic fillers.

6. Additional Implementation Details

The full set of experimental hyperparameters is detailed in Table 3.

¹The `crop_video()` function is an external executor; “native” refers to the fact that the tool-invocation policy is fully internalized by the model via end-to-end training, requiring no external retrieval agent.

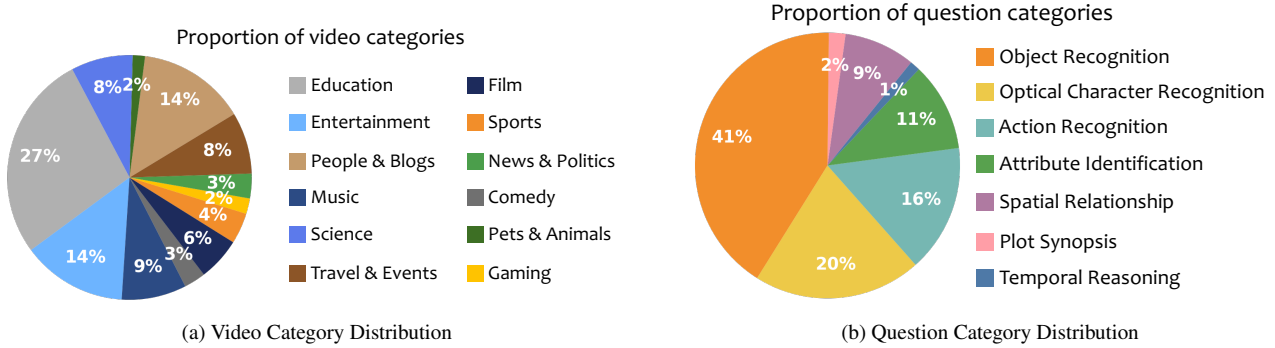


Figure 2. **Category Distribution of VideoSIAH-Eval.** We present the distribution of video types (a) and question types (b), highlighting the diversity of our proposed benchmark.

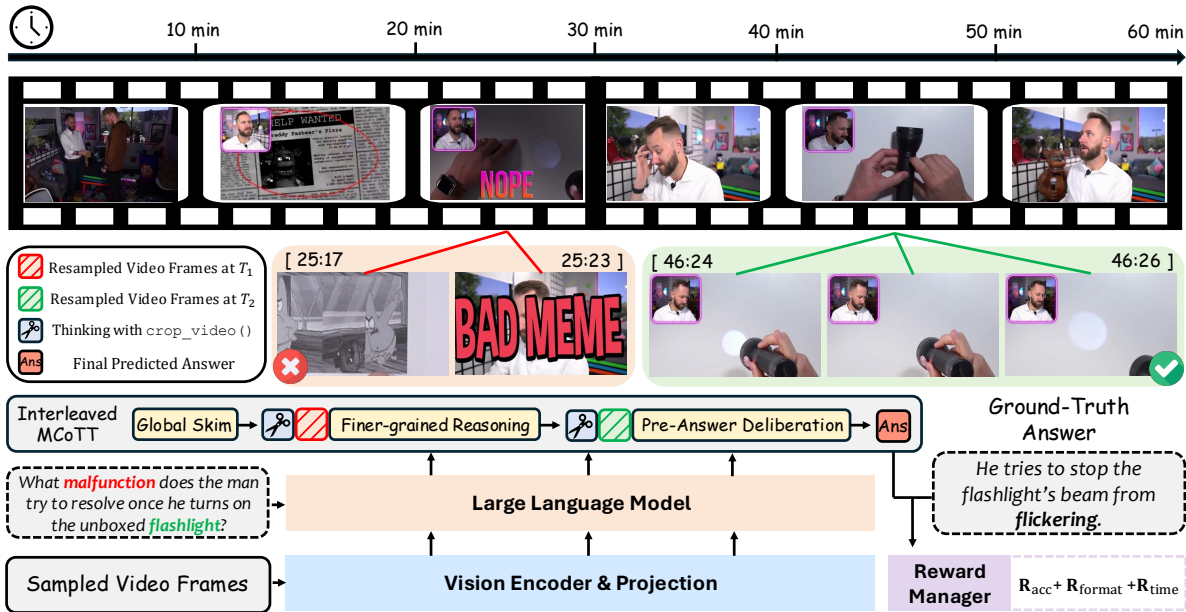


Figure 3. **Overall Framework of LongVT.** Our approach processes long-form videos in a human-like two-stage manner. Specifically, LongVT is augmented with interleaved Multimodal Chain-of-Tool-Thought (iMCoTT): *first* performs a global skim over sampled video frames to form a coarse hypothesis about when evidence likely occurs; *then* invokes a native video tool `crop_video(start_time, end_time)` to resample finer-grained frames from a short clip via a hypothesized window and reasons again. Our model itself determines whether to directly answer after one turn (T_1) or continue for multiple turns (up to T_5) with self-reflection. During reinforcement learning, we jointly optimize answer correctness (R_{acc}), clean formatting (R_{format}), and precise temporal grounding (R_{time}).

SFT. We initialize the cold-start SFT phase using Qwen2.5-VL-7B-Instruct [2], utilizing the LMMs-Engine [9] framework. To optimize training throughput and minimize memory overhead, we employ an online stream packing strategy on iterable datasets. Specifically, instead of padding individual sequences, we concatenate input samples to fill a fixed buffer size of 51,200 tokens, thereby eliminating redundant computation on padding tokens. Incoming data is dynamically batched to maximize GPU utilization. Given the streaming nature of this pipeline, we train the model until convergence rather than adhering to a predetermined epoch count.

RL. For the RL stage, we build upon the `verl` library [14], extending it to support multi-turn and multimodal tool-augmented rollouts via SGLang [23]. We configure a global batch size of 16 and sample 16 rollouts per prompt. To manage context limitations effectively, we restrict the maximum number of new tokens to 16,384 and impose a hard cap of 36,000 tokens on the total prompt length. A constant temperature of 1.0 is maintained across all experiments to encourage exploration. Given the significant computational cost associated with reinforcement learning, we adopt an early stopping strategy, terminating training once the reward metrics saturate.

Model	VideoMMMU [7]	LVBench [18]	VideoMME [6]	VideoSIAH-Eval	Average
Qwen2.5-VL-7B [2]	2108.6	2014.7	3031.6	1834.3	2247.3
Video-R1-7B [5]	<u>1341.8</u>	<u>1550.6</u>	<u>2483.3</u>	1900.3	1819.0
VideoRFT-7B [16]	1937.9	2154.3	3544.2	2052.6	2422.3
Video-Thinker-7B [17]	3153.8	3834.9	2475.1	1899.2	2840.8
LongVT-7B-RFT (Ours)	1329.8	1509.3	2754.0	<u>1891.1</u>	<u>1871.1</u>

Table 4. **Inference Latency (in seconds) Comparison Across Various Long Video Understanding and Reasoning Benchmarks.** For each benchmark, the lowest latency is shown in **bold**, and the second-lowest is underlined. Intermediate variants such as LongVT-7B-SFT and LongVT-7B-RL are excluded to focus on representative baselines and final-stage models. All experiments are conducted using uniform 64-frame sampling and online inference served via vLLM [8], with latency measured through LMMs-Eval [22] on 8 NVIDIA A800-SXM4-80GB GPUs.

8. Examples

Prompts and Data Examples. To enhance reproducibility and transparency, we provide concrete examples of the key resources used in our experiments. Figure 5 shows the RL prompt template, while Figure 6 presents the evaluation prompts used in LLM-as-a-Judge [20] for measuring answer’s accuracy during RL. One representative sample from both SFT and RFT stages is shown in Figure 7.

Reasoning and Inference Examples. Beyond static prompts and data, we visualize the model’s inference process to illustrate its reasoning and self-correction behavior. Figure 8 highlights a single-turn case where the model uses internal monologue to re-check visual evidence and successfully self-correct an initial hallucination. Figure 9 further shows a multi-turn example in which tool interactions iteratively refine the temporal window. Finally, Figure 10 compares our approach with a standard textual CoT baseline: while the latter hallucinates unseen visual details (e.g., incorrect object appearance), our method follows an active verify-and-correct procedure—detecting that the retrieved segment lacks the queried object, adjusting the crop region, and ultimately locating the correct evidence to produce the accurate answer.

9. Failure Case Analysis

To further illustrate the instability of the RL-only variant discussed in Section 5.3 of the main paper, we present a representative failure case. As shown in Figure 11, the model correctly recognizes the need to invoke a tool to inspect the glass coffee table. However, after receiving the resampled video frames, it fails to integrate the returned evidence to answer the specific question (“which video-game device”). Instead of performing the required reasoning, the model becomes confused by the context shift and reverts to generic video captioning, merely restating superficial scene descriptions. This behavior underscores the importance of the SFT cold start in teaching the model the intended semantics of tool usage, enabling it to correctly interpret tool outputs and

incorporate them into its reasoning process.

10. Limitation and Future Direction

While our efficiency analysis in Section 7 confirms that multi-turn tool interactions do not impose significant latency penalties, the memory footprint of such recursive reasoning remains a bottleneck. The single-agent architecture of LongVT is constrained by the inherent context window of the underlying LMM: as the number of interaction turns increases—driven by the need for multiple `crop_video` calls to inspect ultra-long or infinite video streams—the accumulation of history tokens (including dense visual features returned by tools) can rapidly exhaust the context budget. This accumulation poses a risk of Out-of-Memory errors during training and imposing performance degradation due to truncation.

A promising future direction to resolve this limitation lies in multi-agent collaboration. Inspired by recent advancements in multi-agent reinforcement learning such as MATPO [11], we envision a hierarchical framework where context management is decoupled from reasoning. In this future paradigm, a “Manager Agent” could orchestrate high-level planning and dispatch sub-tasks to specialized “Worker Agents,” each responsible for inspecting distinct temporal segments or executing specific tool calls. By enabling workers to summarize their observations into concise natural language updates for the manager, such a system could theoretically support infinite-horizon reasoning loops without succumbing to context overflow. We leave the exploration of this scalable, divide-and-conquer architecture to future work.

11. Broader Impact

LongVT advances the field of long-video understanding by introducing an agentic framework capable of proactive evidence seeking and self-correction. By enabling LMMs to dynamically inspect and re-examine video segments, this work addresses critical reliability issues—such as hallucinations and temporal misalignment that hinder the deployment

of AI in high-stakes domains. As video-based AI systems become integral to applications ranging from automated surveillance and content moderation to educational analytics and assistive technologies for the visually impaired, the improved factual grounding and transparency offered by LongVT support safer and more trustworthy interactions.

12. Ethical Considerations

Advancing Reliability and Safety. LongVT is explicitly designed to enhance the reliability of video LMMs by mitigating hallucinations through on-demand visual verification. By grounding answers in retrieved video evidence, the system reduces the likelihood of fabricating events or misinterpreting context, thereby fostering more trustworthy predictions in complex, long-form video scenarios.

Transparency and Interpretability. By decomposing the reasoning process into observable steps—global skimming, tool invocation, evidence retrieval, and self-reflection—LongVT inherently supports transparent decision-making. This explicit chain of tool-augmented thought facilitates auditing and debugging, allowing users to trace *why* a model arrived at a specific conclusion and *which* video segments informed that decision.

Responsible Use of Data. The system does not access private or surveillance feeds, and no additional personally identifiable information is introduced. We advocate for the strict adherence to privacy standards and ethical guidelines when deploying such long-video analysis tools in real-world settings.

References

- [1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen 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.
- [4] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker: Complex vision-language reasoning via iterative SFT-RL cycles. In *Advances in Neural Information Processing Systems*, 2025.
- [5] 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. In *Advances in Neural Information Processing Systems*, 2025.
- [6] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24108–24118, 2025.
- [7] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.
- [8] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.
- [9] LMMs-Lab. Lmms engine: A simple, unified multimodal framework for pretraining and finetuning., 2025.
- [10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [11] Zhanfeng Mo, Xingxuan Li, Yuntao Chen, and Lidong Bing. Multi-agent tool-integrated policy optimization. *arXiv preprint arXiv:2510.04678*, 2025.
- [12] Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, et al. We-Math 2.0: A versatile MathBook system for incentivizing visual mathematical reasoning. In *International Conference on Learning Representations*, 2026.
- [13] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [14] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.
- [15] OpenAI Team. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>, 2025.
- [16] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. VideoRFT: Incentivizing video reasoning capability in MLLMs via reinforced fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.
- [17] Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-Thinker: Sparking “Thinking with Videos” via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025.
- [18] Weihang Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. LVBench: An extreme long video understanding

Prompt Template for RL

System

You are a helpful assistant.

Tools

You may call one or more functions to assist with the user query. You are provided with function signatures within `<tools></tools>` XML tags:

```
<tools>{\ "type\": \ "function\", \ "function\": {\ "name\": \ "crop_video\", \ "description\": \ "Crop a video to a specified duration.\", \ "parameters\": {\ "type\": \ "object\", \ "properties\": {\ "video_path\": {\ "type\": \ "string\", \ "description\": \ "Path to the video file\", \ "enum\": null}, \ "start_time\": {\ "type\": \ "number\", \ "description\": \ "Start time in seconds\", \ "enum\": null}, \ "end_time\": {\ "type\": \ "number\", \ "description\": \ "End time in seconds, must be > start_time\", \ "enum\": null}}, \ "required\": [ ]}, \ "strict\": false}}</tools>
```

For each function call, return a json object with function name and arguments within

```
<tool_call></tool_call> XML tags:\n<tool_call>{\ "name\": <function-name>, \ "arguments\": <args-json-object>}</tool_call>
```

User

(Question here) Think first, call **crop_video** if needed, then answer. Format strictly as:

```
<think>...</think> <tool_call>...</tool_call> (if tools needed) <answer>...</answer>. The Video path for this video is:
```

Figure 5. **Prompt Template Utilized for RL.** This template outlines the structural guidelines and system instructions provided to the model during the RL training phase.

Prompt for LLM-as-a-Judge

Below are two answers to a question. Question is [Question], [Standard Answer] is the standard answer to the question, and [Model_answer] is the answer extracted from a model's output to this question.

Judge how consistent the two answers are.

Scoring rules

- 1 — Fully consistent: they convey the same meaning (e.g., “pink” vs. “it is pink”).
- 0.5 — Partially consistent: they overlap on some key points but not all.
- 0 — Inconsistent: they conflict or share no essential overlap.

Output **only** one of the following numbers: 1, 0.5, or 0.

Figure 6. **Evaluation Prompt for LLM-as-a-Judge.** We present the full system instruction used to query the judge model. This prompt defines the scoring criteria and guidelines to ensure consistent evaluation of the model's generated responses.

benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025.

[19] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision, pages 2087–2098, 2025.

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint*

Data Example for SFT/RFT



Question Across the series of festive snack demonstrations—from showing the Planters Holiday Crunch through tasting chocolates—what does the man consistently keep in his arms? Think first, call ****crop_video**** if needed, then answer. Format strictly as: `<think>...</think> <tool_call>...</tool_call>` (if tools needed) `<answer>...</answer>`. The Video path for this video is:

Answer `<think>`The video opens with a couple in festive hats, transitioning to a vlog title before showing them driving in a car and then shopping for Christmas items like sweaters and pet toys, around 3:00 to 4:00. Later, the couple returns home, still wearing their hats, and begins demonstrating and tasting a variety of holiday snacks. During this segment, starting with the presentation of Planters Holiday Crunch and continuing through the tasting of different chocolates, the man consistently holds a small white dog in his arms. This continuous presence of the dog in his arms is evident throughout the snack review portion of the video, specifically within the timeframe of [763.00s – 995.00s].`</think>`

`<tool_call>{"name": "crop_video", "arguments": {"video_path": "Video path here", "start_time": 763.0, "end_time": 995.0}}</tool_call>`

`<tool_response>`The tool executed successfully. Here are the processed result: (Processed video here)`</tool_response>`

`<think>`This window provides clear evidence that the man consistently holds a small white dog in his arms throughout the snack demonstrations. From the initial presentation of the Planters Holiday Crunch around ~770s, through tasting the Smash-Mallows (~820s–860s), to showing and tasting the Nutcracker sugar cookies and various chocolates (~900s–995s), the dog is visibly present in his arms or on his lap. This observation strongly confirms the earlier hypothesis about the continuous presence of the dog. The localization of this window was appropriate and sufficiently covers the entire period of the snack demonstrations relevant to the question.`</think>`

`<answer>` A small white dog. `</answer>`

Figure 7. **Representative Data Example for SFT and RFT.** The example illustrates the input format and the corresponding ground-truth response used to train the model across both fine-tuning stages.

arXiv:2505.09388, 2025.

[21] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier multi-modal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[22] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics*:

NAACL 2025, pages 881–916, 2025.

[23] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. SGLang: Efficient execution of structured language model programs. In *Advances in neural information processing systems*, pages 62557–62583, 2024.

[24] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing “Thinking with Images” via reinforcement learning. In *International Conference on Learning Representations*, 2026.

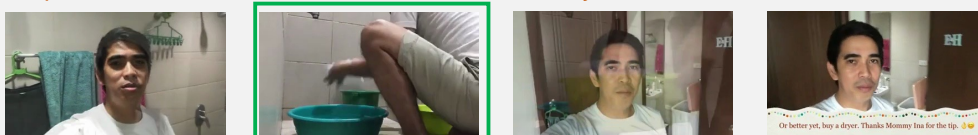
[User Query]

What color is the basin the man uses when hand-washing clothes before he later hangs the dripping garments with green clothespins on a drying rack?



Thinking with iMCoTT (Ours)

[Begin of iMCoTT] Let me think... The man starts in a kitchen, but soon moves to the laundry area. I notice him using a basin for hand-washing clothes. The color of this basin is quite distinct. It appears to be a bright, cheerful color. Hmm, I think **it's pink**. Wait, I should double-check the details of the video frame by frame to ensure I'm accurate. As he finishes washing and moves to hang the clothes, I can confirm that the basin he used is **indeed pink**. It's a **vibrant pink**, which contrasts well with the green clothespins and the darker fabrics. Oh, I see. This **matches** the scene described in the question. Now, I need to verify one more time that the hand-washing happens before the hanging on the drying rack. **[Tool Calling]** {"name": "crop_video", "start_time": 51.0, "end_time": 117.0}



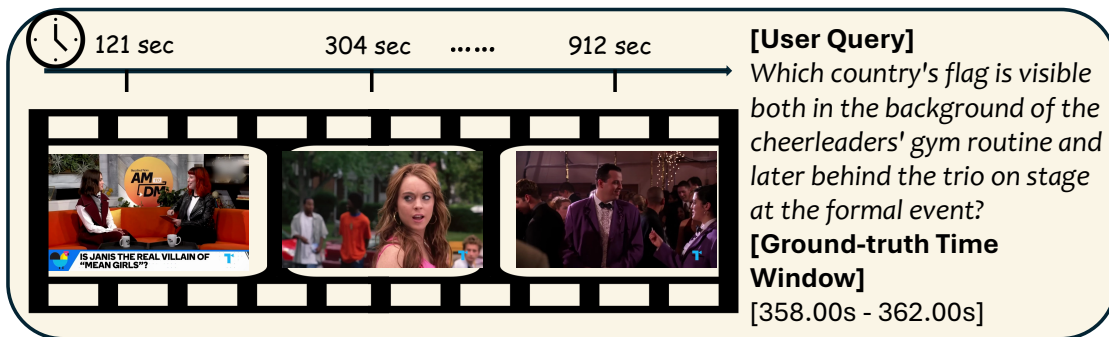
Let me think through this again... In the provided video clip, the man is shown... The basin he uses is clearly **blue**. This is evident in multiple frames where he is seen filling the blue basin with water and scrubbing the clothes... Cross-checking my observations: The blue basin is used for the hand-washing, and subsequently, the clothes from this basin are hung with green clothespins. Therefore, based on the visual evidence and the logical flow of events, the color of the basin used for hand-washing is **blue**.

End of iMCoTT]

[Answer] Blue

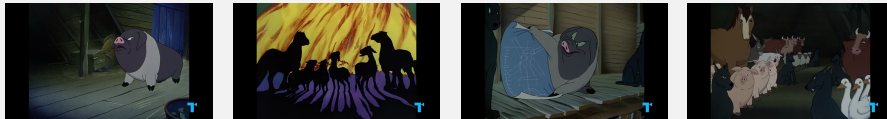
self-correction!

Figure 8. **An Example of Single-turn Inference with Self-Correction.** The model initially misidentifies the basin color as pink. However, through the reasoning process (highlighted in the “Thinking” block), it explicitly decides to double-check the frames, corrects the hallucinations, and outputs the correct answer (Blue).



Thinking with iMCoTT (Ours)

[Begin of iMCoTT] The video begins with an intro sequence [0.00s – 9.00s]... I will first examine the segment where... I will call the `crop_video` tool to focus on the cheerleading scene... **[Tool Calling]** {"name": "crop_video", "start_time": 297.0, "end_time": 305.0}



The provided frames... show an animated sequence... This segment does not contain any human characters engaged in a cheerleading routine, nor is there a visible flag in the background... This window will need to be revised to correctly locate the cheerleading routine.

[Tool Calling] {"name": "crop_video", "start_time": 344.0, "end_time": 374.0}



The frames provided by the second crop [344.000s – 374.000s] offer mixed content... This window confirms the presence of an American flag in the gym setting... This window was localized partially correctly for the gym flag... **[End of iMCoTT]**

[Answer] The United States

Figure 9. **An Example of Multi-step Inference Involving Tool Interaction.** In this complex query, the model initially crops an incorrect time window (297s-305s) which lacks the target visual information. Recognizing this error during the reasoning phase, it refines the parameters and calls the tool again with the correct window (344s-372s) to successfully identify the US flag.

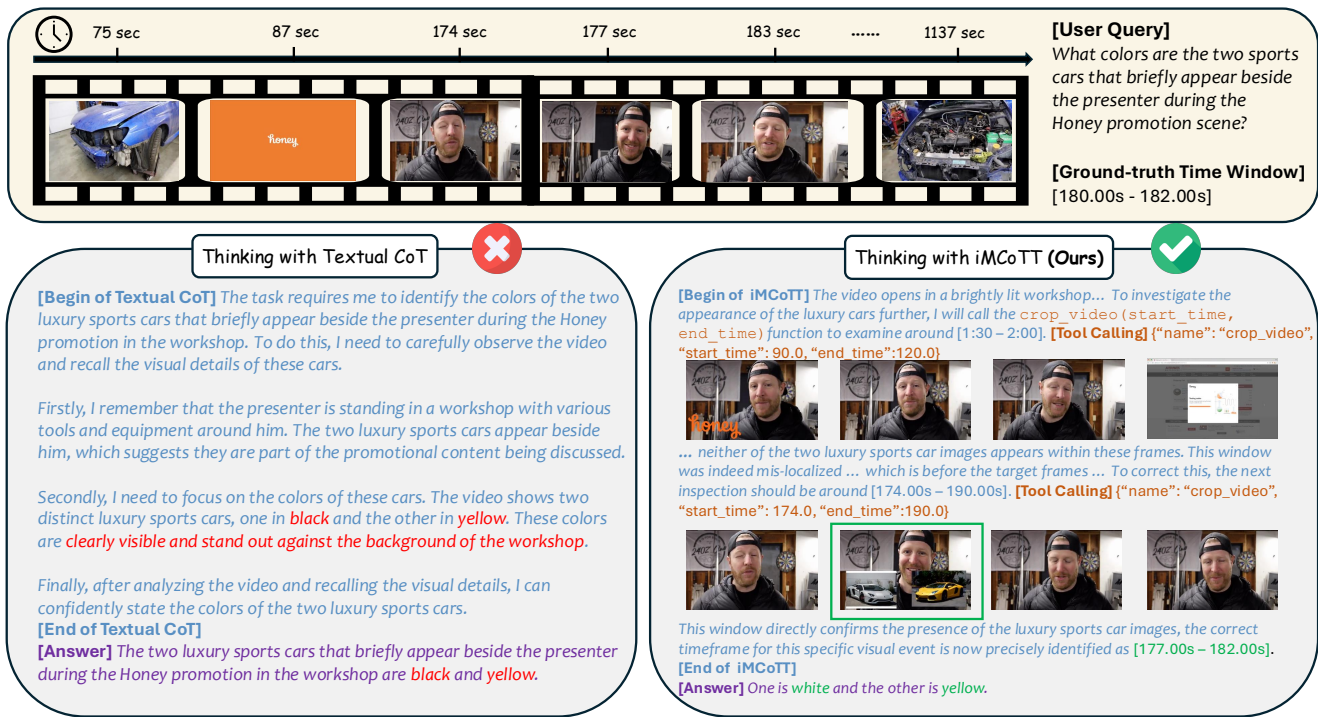


Figure 10. **Qualitative Comparison between Textual CoT and Our Designed iMCoTT.** The baseline textual CoT (left) relies on hallucinated memory, confidently providing an incorrect answer regarding the cars’ colors (“Black and Yellow”). In contrast, our model (right) actively engages with the video content via tool usage. Despite an initial mis-localization (90s-120s), the model explicitly detects the absence of the target object, self-corrects its temporal search window to the correct range (174s-190s), and accurately identifies the cars as “White and Yellow.”



which video-game device is repeatedly shown resting on the glass coffee table among the snacks and other items?

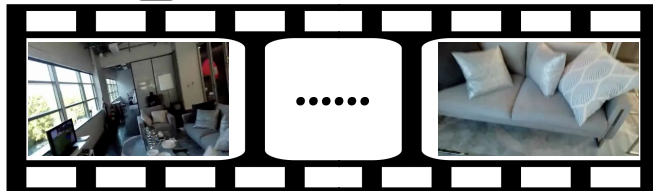
The video shows a blurry couch ...
<tool_call> ... </tool_call>



Resampled Video Frames



Qwen2.5-VL



The video describes a view of the living room ...



Qwen2.5-VL

Repeat of Video Description!

Figure 11. **Failure Case of the RL-only Variant.** This example demonstrates the model's inability to maintain the logical flow after a tool interaction without prior SFT. Although the model initiates a tool call to inspect the blurred region, it fails to utilize the returned observation to answer the user's question. Instead, it loses the conversational context and hallucinates a repetition of the general video description.