



LongVT

Incentivizing “Thinking with Long Videos” via Native Tool Calling

Zuhao Yang* Sudong Wang* Kaichen Zhang* Keming Wu, Sicong Leng, Yifan Zhang, Bo Li,
Chengwei Qin, Shijian Lu, Xingxuan Li, Lidong Bing

MiroMind AI · NTU · HKUST(GZ) · Tsinghua University · LMMs-Lab

CVPR 2026 Main Conference · June 3-7, 2026 · Denver, CO

arxiv.org/abs/2511.20785

github.com/EvolvingLMMs-Lab/LongVT · evolvinglmm-lab.github.io/LongVT

* Equal contribution. □ Corresponding authors: Shijian.Lu@ntu.edu.sg, xingxuan.li@miromind.ai

- Long videos contain **sparse, decisive evidence** across thousands of frames.
- Existing models rely on **text-only chain-of-thought** — language-centric and prone to **hallucination** under sparse evidence.
- A human asked **“Which foot scored the equalizer?”** would skim globally, jump to crowd reactions, then zoom on the goal.
- **Our question:** can LMMs reason over long videos by performing **human-like visual operations?**

70 sec 141 sec 209 sec 347 sec 562 sec 1056 sec

Which foot did the French player use to execute the volley, equalizing the score?

Thinking with Textual CoT

[Begin of Textual CoT] The video shows a sequence of events where a French player executes a volley to equalize the score. The question asks which foot was used for this action.

However, the provided frames do not show the specific moment when the French player takes the shot. Instead, they depict various scenes from a soccer match, including a goal scored by Messi and celebrations that follow.

Without a clear depiction of the French player's volley, it is impossible to determine which foot was used based on the given information. [End of Textual CoT]

[Answer] Based on the provided frames, it is not possible to determine which foot the French player used to execute the volley, as the specific moment of the shot is not shown. Therefore, no definitive answer can be given.

Thinking with iMCoTT (Ours)

[Begin of iMCoTT] The video is a comprehensive highlight reel of the ... It begins with pre-match scenes, showing players in the tunnel and lining up... Then Argentina scores a second goal... The user is asking about the equalizing volley. Based on my skim, this happens very quickly after... I will first investigate the segment just before the first penalty, around [08:00 - 08:30], to see if the volley occurred... [Tool Calling] {"name": "crop_video", "start_time": 480.0, "end_time": 510.0}

The video segment from [480.000s - 510.000s] (08:00 - 08:30) shows various French attacking plays... However, this segment does not contain ... My initial localization was incorrect ... The equalizing goal... happened after France's first goal. Reviewing my global skim, the volley appears to happen almost immediately after the restart following that penalty. I will now examine the segment from ... to find the correct moment [Tool Calling] {"name": "crop_video", "start_time": 590.0, "end_time": 650.0}

The video segment from [590.000s - 650.000s] (09:50 - 10:50) successfully captures the equalizing goal. The play starts with... At approximately, the French player connects with the return pass in the air, striking a volley. Watching the action closely, he plants his left foot and strikes the ball with his right foot to score the goal and equalize 2-2. [End of iMCoTT]

[Answer] The French player used his right foot to execute the volley that equalized the score at 2-2, as seen around the 10:45 mark in the video.

Fig: text-only CoT (left) vs. our tool-augmented LongVT (right)

70 sec

141 sec

209 sec

347 sec

562 sec

.....

1056 sec

Which foot did the French player use to execute the volley, equalizing the score?

Thinking with Textual CoT

[Begin of Textual CoT] The video **shows** a sequence of events where **a French player executes a volley to equalize the score**. The question asks which foot was used for this action.

However, the provided frames **do not show** the specific moment when the French player takes the shot. Instead, they depict various scenes from a soccer match, including a goal scored by Messi and celebrations that follow.

Without a clear depiction of the French player's volley, it is **impossible** to determine which foot was used based on the given information. **[End of Textual CoT]**

[Answer] Based on the provided frames, it is not possible to determine which foot the French player used to execute the volley, as the specific moment of the shot is not shown. Therefore, **no definitive answer** can be given.

Thinking with iMCoTT (Ours)

[Begin of iMCoTT] The video is a comprehensive highlight reel of the ... It **begins with** pre-match scenes, showing players in the tunnel and lining up... **Then** Argentina scores a second goal... The user is asking about the equalizing volley. Based on my skim, this happens very quickly after... **I will first investigate** the segment just before the first penalty, around [08:00 - 08:30], to see if the volley occurred... **[Tool Calling]** {"name": "crop_video", "start_time": 480.0, "end_time": 510.0}

The video segment from [480.000s - 510.000s] (08:00 - 08:30) shows various French attacking plays... **However**, this segment does not contain ... My initial localization was **incorrect** ... The equalizing goal... happened after France's first goal. Reviewing my global skim, the volley **appears to happen** almost immediately after the restart following that penalty. **I will now examine** the segment from ... to find the correct moment **[Tool Calling]** {"name": "crop_video", "start_time": 590.0, "end_time": 650.0}

The video segment from [590.000s - 650.000s] (09:50 - 10:50) **successfully captures** the equalizing goal. The play starts with... At approximately, the French player connects with the return pass in the air, striking a volley. **Watching the action closely**, he plants his left foot and strikes the ball with his **right foot** to score the goal and equalize 2-2. **[End of iMCoTT]**

[Answer] The French player used his **right foot** to execute the volley that equalized the score at 2-2, as seen around the 10:45 mark in the video.

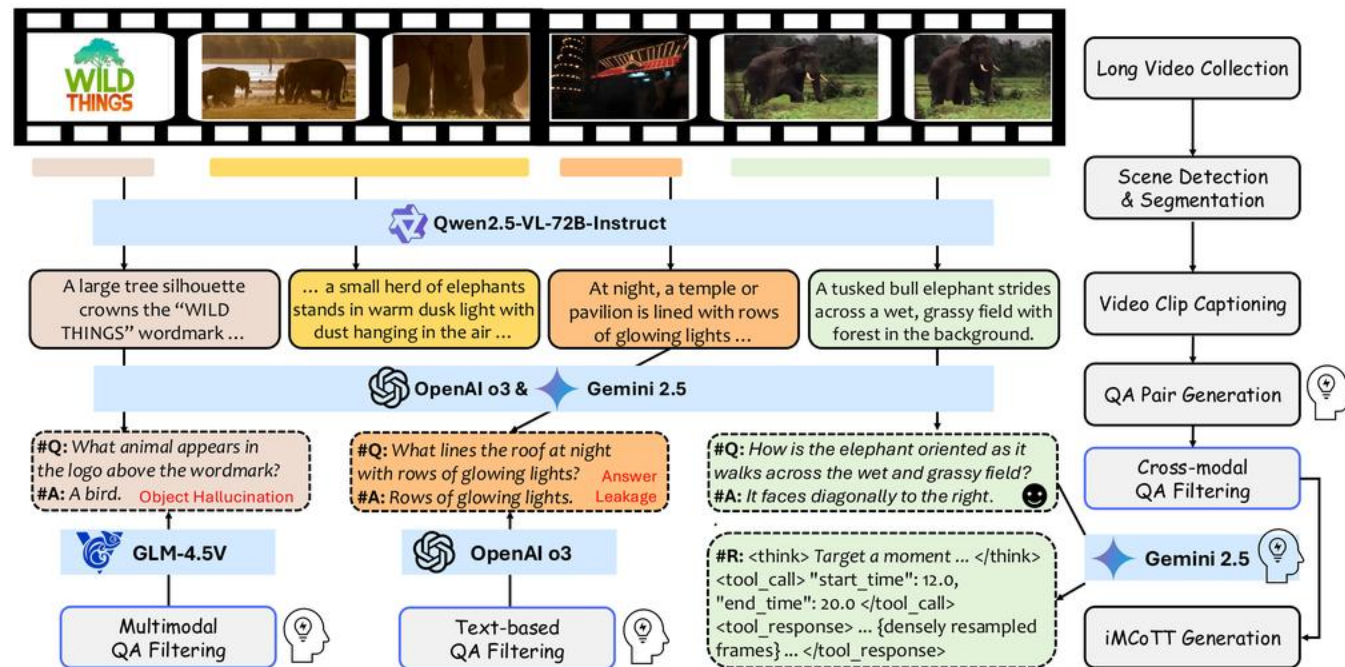
- Model performs **global skim** → **hypothesize window** → call **crop_video(t_s, t_e)** → **resample fine frames** → **re-think** → **self-correct**.
- Tool is a **latent LMM capability** — activated by tool-integrated supervision, **no external retriever** needed.

Training (265K samples):

- **228.8K** non-tool SFT samples
- **19.2K** tool-augmented iMCoTT SFT
- **1.6K** agentic-RL QAs
- **15.4K** self-distilled RFT traces

Evaluation:

- **652** QA pairs in VideoSIAH-Eval, human-validated
- Avg. duration **≈1,688 s** (71.8% in 15-30 min)



Data pipeline: scene-detect → caption → QA-gen → multimodal filter → iMCoTT trace

(1) Cold-start SFT

Tool-integrated SFT teaches the base LMM to:

- Propose precise temporal windows
- Invoke `crop_video` to resample frames
- Self-correct when the window is off

(2) Agentic RL (GRPO)

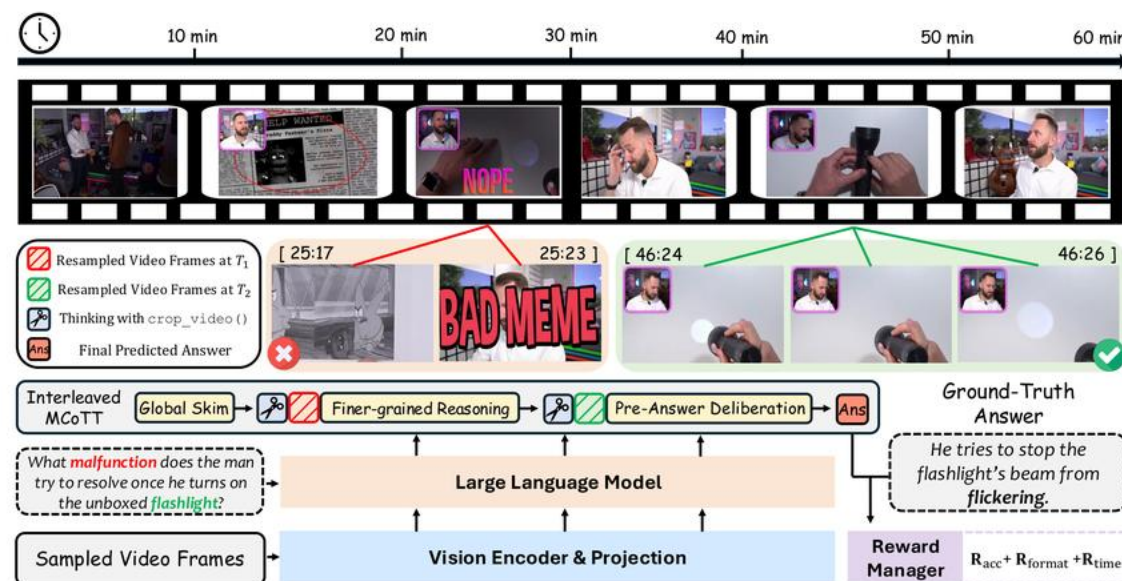
Joint answer-temporal-grounding reward refines tool-using rollouts:

- Answer correctness (LLM-as-Judge)
- IoU between predicted & GT window
- Format compliance

(3) Agentic RFT

Self-distilled rollouts (answer \square + IoU ≥ 0.3) become supervised data:

- Stabilizes learned behaviors
- High-precision in-distribution training
- Beyond the SFT-only ceiling



Framework: hypothesis → verification → self-correction loop

R_{acc} — Answer correctness (LLM-as-Judge)

Full = 1.0 · Partial = 0.5 · Inconsistent = 0

R_{format} — Schema compliance

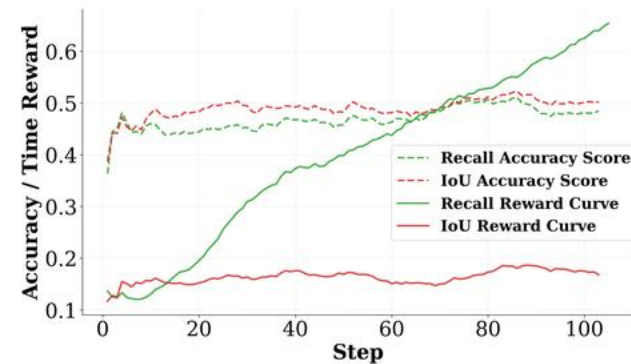
Match required output schema: 1.0 · otherwise 0

R_{time} — Temporal IoU

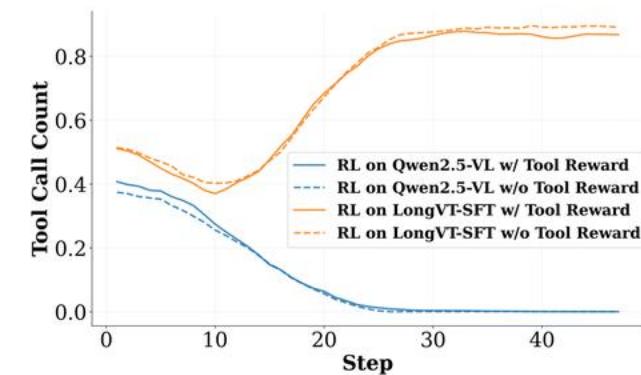
IoU between predicted and ground-truth [t_s, t_e]
→ tighter than Recall, penalizes span inflation

$$R = R_{\text{acc}} + R_{\text{format}} + R_{\text{time}}$$

Tool reward is unnecessary after cold-start:



(a) Time Reward Ablation



(b) Tool Reward Ablation

Reward ablation: tool-call bonus does not lift accuracy — we drop it.

Open-source 7B video LMMs · dense frame sampling:

Model	VideoMME	VideoMMU (avg)	LVBench	VideoSIAH-Eval	Avg.
Qwen2.5-VL-7B (base)	64.3	44.9	40.9	33.8	46.0
Video-R1-7B	60.5	40.8	40.1	33.1	42.7
VideoRFT-7B	49.2	42.4	18.7	26.9	37.0
Video-Thinker-7B	60.8	45.2	54.3	6.6	42.9
LongVT-7B-SFT (Ours)	64.9	41.3	41.1	34.8	44.1
LongVT-7B-RL (Ours)	66.1	45.4	41.4	35.9	46.6
LongVT-7B-RFT (Ours) •	67.0	45.4	41.3	42.0	47.7

• **LongVT-7B-RFT** reaches **47.7 avg** · +6 pts over second-best on VideoSIAH-Eval · **~4 pts within proprietary GPT-4o on average**

#01 Fine-grained data is essential

Removing self-curated iMCoTT crashes
VideoSIAH-Eval from

34.8 → 4.1

#02 IoU beats Recall on grounding

On Charades-STA mIoU,
IoU reward gives

+6.0 pts

#03 All three stages contribute

Avg over 4 benchmarks:
SFT only → SFT + RL + RFT

44.1 → 47.7

Take-away: each stage adds measurable gain; data quality + boundary precision matter more than reward bells & whistles.

Thank You!

Find us at our CVPR 2026 poster — June 3-7, Denver

Paper

arxiv.org/abs/2511.20785

Full paper, appendix, and citations

Code & Models

github.com/EvolvingLMMs-Lab/LongVT

Code, weights, training scripts

Project Page

evolvinglmm-lab.github.io/LongVT

Demos and dataset downloads

LongVT lets LMMs actively look again at video evidence

— rather than passively consume frames.

Corresponding authors: Shijian.Lu@ntu.edu.sg · xingxuan.li@miromind.ai

CVPR 2026 Main Conference · June 3-7, 2026 · Denver, CO