VideoZoomer: Reinforcement-Learned Temporal Focusing for Long Video Reasoning

Yang Ding[1*], Xin Lai[2*], Yizhen Zhang[1*], Wei Li[3], Ruihang Chu[1], Yujiu Yang[1]

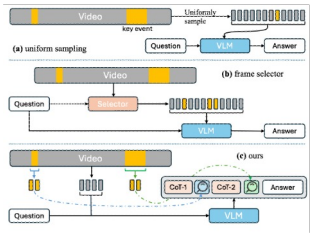
[1]Tsinghua University, [2]The Chinese University of Hong Kong, [3]ByteDance

*Equal contribution Contact:dingyang24@mails.tsinghua.edu.cn

Project Page

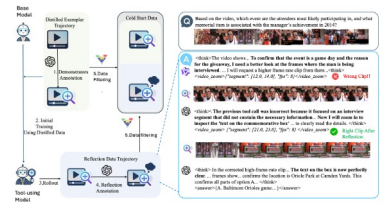
Motivaion

Multimodal Large Language Models (MLLMs) struggle with long video understanding due to limited context, which prevents processing all frames. **Existing methods** rely on **static frame selection**, which can miss key moments and cannot be corrected during reasoning. As a result, video understanding is treated as a one-time observation problem rather than an **active evidence-gathering** process.



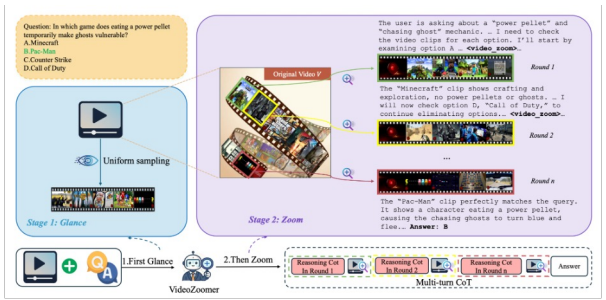
Data Construction Pipeline

We build training data by collecting **multi-step reasoning trajectories** for long video understanding. Starting from coarse video observations, we generate step-by-step interactions that include **when and where to zoom** for more detailed evidence. These trajectories are then refined with **reflection and correction signals** to improve decision quality. The final dataset contains agent-like reasoning processes, which are used for supervised fine-tuning and further optimization with reinforcement learning.

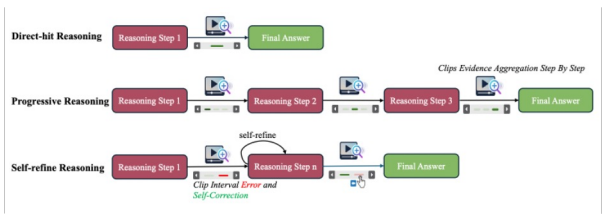


VideoZoomer Approach

We propose an agentic framework that allows MLLMs to **dynamically focus on important video segments during reasoning**. The model first processes a coarse, low-frame-rate overview of the video, then iteratively selects key moments to retrieve higher-frame-rate clips through a temporal zoom mechanism, progressively gathering more detailed evidence. This interaction forms a multi-step reasoning process of "**observe → zoom → refine**." The model is trained with a **two-stage strategy (SFT + RL)**: supervised fine-tuning on curated reasoning trajectories to learn the basic behavior, followed by reinforcement learning to optimize when and where to zoom.



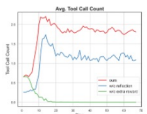
The model supports direct, step-by-step, and self-corrective reasoning by dynamically retrieving and refining video evidence.



Experiments & Result

Model	Size	Long Video Understanding						Long Video Reasoning		
		MLVU		LongVideoBench	VideoMME		LVBench	VideoMMLU	VideoMMMU	LongVideoReason
		dev	*test*	*val*	*overall*	*long*		*quiz*		*eval*
Proprietary Models										
GPT-4o	-	64.6	54.9	66.7	71.9	65.3	48.9	44.9	61.2	60.7
Gemini-1.5-Pro	-	-	-	64.0	75.0	67.4	33.1	-	53.9	67.3
Open-Source VLMs										
Video-LLaVA	7B	36.2	30.7	37.6	39.9	-	-	-	-	-
LLaVA-OneVision	7B	64.7	47.2	56.4	58.3	46.7	-	33.4	33.9	-
LLaVA-NeXT-Video	7B	-	-	49.1	-	-	-	27.6	-	-
Video-XL	7B	64.9	45.5	50.7	55.5	-	-	-	-	-
VILA-1.5	7B	56.7	-	-	-	-	-	20.5	20.9	-
Kangaroo	8B	61.0	-	54.8	56.0	-	39.4	-	-	-
LongVU	7B	65.4	-	-	60.6	-	-	-	-	-
LongVA	7B	56.3	41.1	-	52.6	-	-	-	24.0	-
LongVILA	7B	-	-	57.1	60.1	-	-	-	-	-
LongVILA-R1	7B	-	-	57.6	62.4	53.3	-	-	-	67.9
Video-R1†	7B	65.0	49.2	52.0	61.1	51.4	38.7	61.3	49.8	72.8
Qwen2.5-VL†	7B	58.3	45.5	51.0	63.5	53.9	36.9	61.0	48.1	70.8
VideoZoomer†	7B	**68.8**	**55.8**	**57.7**	**65.2**	**55.8**	**41.5**	**67.9**	**52.2**	**80.3**
Δ over base model		+10.5	+10.3	+6.7	+1.7	+1.9	+4.6	+6.9	+4.1	+9.5

Benchmarks:Evaluate on multiple long video understanding and reasoning benchmarks (e.g., LVBench and others), covering diverse and complex tasks.
Performance:Our 7B model achieves **strong accuracy across all benchmarks**, consistently outperforming existing open-source models and remaining competitive with proprietary systems.
Efficiency:Achieves **better performance under reduced frame budgets**, showing improved efficiency compared to static frame sampling methods.
Ablation Study:Both **temporal zooming** and **RL** contribute significantly to performance gains.



Model	Long Video Understanding						Long Video Reasoning		
	MLVU	LongVideoBench	VideoMME		LVBench	VideoMMLU	VideoMMMU	LongVideoReason
	dev	*test*	*val*	*overall*	*long*		*quiz*		*eval*
VideoZoomer	**68.8**	**55.8**	**57.7**	**65.2**	**55.8**	**41.5**	67.9	52.2	**80.3**
w/o RL	56.4	45.6	42.0	54.4	44.2	26.0	63.5	46.6	63.3
w/o R_{coef}	67.5	52.2	56.2	62.5	52.5	40.6	63.6	**53.8**	79.9
w/o cold-start	67.0	42.8	43.5	53.5	46.6	35.5	63.9	43.6	59.6
w/o reflection	67.0	53.2	54.8	58.7	47.4	40.9	70.1	52.2	75.1