Think with Grounding: Curriculum Reinforced Reasoning with Video Grounding for Long Video Understanding

Houlun Chen
chenhl23@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing, Beijing, China

Xin Wang*
xin_wang@tsinghua.edu.cn
DCST, BNRist, Tsinghua University
Beijing, Beijing, China

Guangyao Li
guangyaoli@tsinghua.edu.cn
DCST, Tsinghua University
Beijing, Beijing, China

Yuwei Zhou
zhou-yw21@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing, Beijing, China

Yihan Chen
chenyiha24@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing, Beijing, China

Jia Jia*
Wenwu Zhu*
jjia@tsinghua.edu.cn
wwzhu@tsinghua.edu.cn
DCST, BNRist, Tsinghua University
Beijing, Beijing, China

Abstract

Long video understanding (LVU) is challenging due to rich and complicated multimodal clues in long temporal range. Current methods adopt reasoning to improve the model's ability to analyze complex video clues in long videos via text-form reasoning. However, the existing literature suffers from the fact that the text-only reasoning under fixed video context may exacerbate hallucinations since detailed crucial clues are often ignored under limited video context length due to the temporal redundancy of long videos. To address this gap, we propose **Video-TwG**, a curriculum reinforced framework that employs a novel ***T**hink-**w**ith-**G**rounding* paradigm, enabling video LLMs to actively decide when to perform on-demand grounding during interleaved text–video reasoning, selectively zooming into question-relevant clips only when necessary. Video-TwG can be trained end-to-end in a straightforward manner, without relying on complex auxiliary modules or heavily annotated reasoning traces. In detail, we design a Two-stage Reinforced Curriculum Strategy, where the model first learns *think-with-grounding* behavior on a small short-video GQA dataset with grounding labels, and then scales to diverse general QA data with videos of diverse domains to encourage generalization. Further, to handle complex *think-with-grounding* reasoning for various kinds of data, we propose the TwG-GRPO algorithm, which features the fine-grained grounding reward, self-confirmed pseudo reward, and accuracy-gated mechanism. Finally, we propose to construct a new TwG-51K dataset that facilitates training. Experiments on Video-MME, LongVideoBench, and MLVU show that Video-TwG consistently outperforms strong LVU baselines. Further ablation validates the necessity of our Two-stage Reinforced Curriculum Strategy and shows our TwG-GRPO better leverages diverse unlabeled data to improve grounding quality and reduce redundant groundings without sacrificing QA performance.[1]

CCS Concepts

• **Computing methodologies** → **Temporal reasoning**; **Computer vision tasks**.

Keywords

Video Reasoning, Video Grounding, Reinforcement Learning

ACM Reference Format:
Houlun Chen, Xin Wang, Guangyao Li, Yuwei Zhou, Yihan Chen, Jia Jia, and Wenwu Zhu. 2026. Think with Grounding: Curriculum Reinforced Reasoning with Video Grounding for Long Video Understanding. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805712.3809757

1 Introduction

Long video understanding (LVU) [12, 45, 52, 53, 70] has become a cornerstone for real-world multimodal information access. Users increasingly pose natural-language questions about movies, vlogs, news, and instructional videos, expecting accurate answers grounded in long-form video content. Recent benchmarks further push LVU toward realistic, hour-scale videos and compositional reasoning, exposing substantial gaps in current video large language models (LLMs) with respect to long-horizon perception and reasoning [12, 53, 70].

Many works address LVU from multiple perspectives and achieve good performance, such as long context modeling [6, 32, 37, 38], memory-centric designs [8, 34, 39], token compression [28, 36, 42, 43, 66], adaptive sampling [25, 41], video agent pipelines [27, 33, 50], *etc.* Since long videos contain rich and complicated multimodal clues in long temporal range, recent video reasoning LLMs [5, 10, 22] benefit from test-time-scaling via conducting text-form reasoning to better analyze the complex video clues, which contributes to LVU significantly. However, they conduct text-only reasoning under

*Corresponding authors.

This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '26, Melbourne, VIC, Australia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809757

[1] https://github.com/hlchen23/Video-TwG

fixed video context, suffering from hallucinations since the important detailed information might be ignored under limited video context length due to the temporal redundancy of long videos, where only a few sparsely distributed moments are relevant for answering a given question, and thus, achieve sub-optimal results.

To address this problem, we propose **Video-TwG**, a curriculum reinforced framework where we employ a novel ***Think-with-Grounding*** paradigm, which equips video LLMs with explicit on-demand video grounding behavior during reasoning, without relying on complex auxiliary modules or costly supervised fine-tuning reasoning traces. As shown in Fig. 1, instead of passively reasoning from a fixed, statically sampled video context, our method follows the insight of Retrieval-Augmented-Generation (RAG) and iteratively decides what evidence it needs next and issues grounding actions to zoom into query-relevant video clips with fine-grained representation.

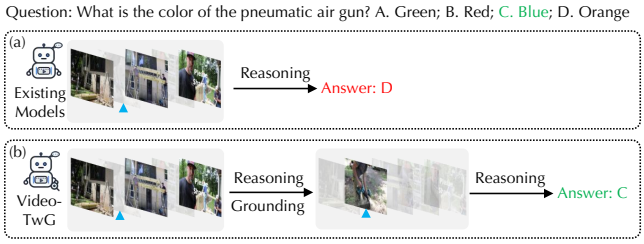


Figure 1: Conceptual level comparison between our Video-TwG and existing video reasoning models. The frame with blue triangle means the important detail for this question. Existing reasoning models reach the wrong answer via reasoning on missed video context, while we address it via leveraging grounding during reasoning to dynamically perceive important clues.

In detail, our Video-TwG mainly contains the following parts. To reduce the training difficulty, we propose a Two-stage Reinforced Curriculum Strategy. In the first stage, the model learns *think-with-grounding* paradigm on easier, short-video GQA data. In the second stage, the training scales to general QA data on videos of various domains and duration, where grounding labels are largely unavailable. In this stage, the model is encouraged to generalize its *think-with-grounding* ability to richer scenarios. Furthermore, to handle the complicated reasoning trajectories for various kinds of data, we propose the TwG-GRPO algorithm, where we assign the fine-grained grounding reward for GQA data, and specifically design the self-confirmed pseudo reward for large amounts of general QA data to encourage high-quality grounding actions, and we apply the accuracy-gated mechanism to avoid the optimization conflicts between accuracy and grounding targets. To support our training, we propose to construct TwG-51K, a novel dataset which combines (a) a small subset of video GQA data with grounding annotations and (b) a large collection of unlabeled general video QA data for generalization purposes.

Empirically, without relying on heavily annotated reasoning traces for supervised fine-tuning, Video-TwG consistently outperforms strong baselines across three mainstream benchmarks after two-stage curriculum training: Video-MME [12], LongVideoBench [53], and MLVU [70]. Compared to Qwen2.5-VL-7B [2], our model achieves improvements of 7.0, 5.3, and 7.1 under low-resolution inputs, and 2.5, 3.9, and 5.0 under high-resolution inputs on these benchmarks, respectively. Ablation study validates the necessity of our Two-stage Reinforced Curriculum Strategy and shows that our TwG-GRPO effectively helps the model to learn *think-with-grounding* from various data, especially in unlabeled data. It's worth noting that with the self-confirmed pseudo reward, the model learns to conduct grounding actions more adaptively, improving the grounding quality and reducing grounding actions by 15.7% without sacrificing QA performance.

In summary, our contributions are fourfold:

- We propose Video-TwG, a curriculum reinforced framework that enables video LLMs to iteratively think with grounding for long video understanding, without heavy external modules or costly supervised reasoning traces.
- We introduce a Two-stage Reinforced Curriculum Strategy to reduce the training difficulty and TwG-GRPO, a novel algorithm which combines fine-grained grounding rewards, self-confirmed pseudo rewards, and an accuracy-gated mechanism to jointly improve answer correctness and grounding quality.
- We construct a new dataset, TwG-51K, which can facilitate the training by providing both labeled and unlabeled data.
- Extensive experiments on Video-MME, LongVideoBench, and MLVU show consistent advantages over strong LVU baselines.

2 Related Works

2.1 Long Video Understanding

Long video understanding (LVU) is the fundamental cornerstone for real-world applications and is quite challenging for current video LLMs. Recent LVU methods mainly fall into the following directions. (i) Long-context modeling scales video LLMs to longer inputs via architectural/efficiency improvements [6, 32, 37, 38] with hour-scale attempts. (ii) Memory-centric designs maintain learnable or external memory to accumulate salient cues over time [8, 34, 39]. (iii) Token compression reduces the cost of long sequences [28, 36, 42, 43, 66]. (iv) Adaptive sampling selects informative frames/clips under budget [25, 41] and highlights "needle-in-a-haystack" temporal localization [60]. Meanwhile, realistic benchmarks like MLVU expose persistent length-induced degradation [70]. (v) video agent systems decompose the long video question into several intermediate steps and design cascaded pipelines [27, 33, 50, 57] Above all, the insight is to model long-range relations among multimodal cues and allocate computation to the most informative parts of long videos. This motivates our Video-TwG framework that enables video LLMs to invoke on-demand grounding during reasoning: reasoning interprets complex cues, while grounding pinpoints the most relevant video clips.

2.2 Video Reasoning

Recent "test-time scaling" suggest that allocating more computation at inference can further unlock LLM capabilities. Among this, LLM reasoning (e.g., CoT [51]/ToT [59]) has shown clear scaling

benefits for solving complex problems, and recent work further demonstrates that post-training with reinforcement learning can substantially strengthen reasoning behaviors [14]. Building on these advances, video reasoning emphasizes producing process-like intermediate traces, like stepwise rationales, to improve output quality and interpretability, rather than only emitting final answers, to better grasp multimodal cues and analyze their relations. Current post-training reasoning methods, often RL-flavored, encourage models to produce more faithful, verifiable reasoning traces, especially for temporal understanding cases [5, 10, 10, 17, 22, 26, 49]. Besides, new datasets/benchmarks provide reasoning-trace supervision for video understanding and analysis, showing that temporal localization and evidence alignment remain key bottlenecks for current video LLMs [29, 30]. DeepEyes [69] enhances the image reasoning with cropping to better align with picture details. However, most works in videos optimize text-form reasoning under fixed video context [10, 30], which doesn't work well in LVU due to its high temporal redundancy, whereas our Video-TwG triggers dynamic grounding actions during reasoning to retrieve and zoom into relevant clips, letting the reasoning enhance perception. During our work, concurrent works, such as VideoChat-R1 [22], Video-MTR [56], introduce a similar tool calling fashion in the reasoning process. Different from them, our method can directly learn from annotation-free data, *i.e., data without reasoning and grounding labels*, making our method more scalable in practice.

2.3 Video Grounding

Video grounding localizes the start and end timestamps of a clip that semantically matches a natural-language query [1, 13, 20]. With the rise of Video LLMs, grounding is increasingly used to test whether models can consistently point to temporal evidence rather than only generate plausible descriptions. Recent works [7, 11, 16, 18, 23, 31, 40, 48, 54, 58, 63, 64] advance Video-LLM grounding by improving temporal referencing and grounded training/evaluation, yielding better accuracy on standard benchmarks. Beyond the simple grounding task with a descriptive text query, we treat video grounding as an RAG-style retrieval step conditioned on the Video-LLM's accumulated video-text reasoning context. At each step, the model first retrieves the most relevant clip given the video and prior reasoning, then reloads the grounded clip with a fine-grained representation into the context to support subsequent reasoning.

3 Video-TwG Framework

3.1 Preliminary

Our framework formulates *think-with-grounding* to multi-turn dialog. Given a question Q and a video V, the video LLM $\mathcal{M}$ is required to give its answer a in a reasoning chain consisting of grounding(g) and answering(a) actions along with thinking process(t) in a maximum of K turns. The k-th turn is formulated to:

$$(t_k, g_k) \text{ or } (t_k, a) = \mathcal{M}(V, p, t_1, g_1, v_1, p_1, ..., t_{k-1}, g_{k-1}, v_{k-1}, p_{k-1}), \tag{1}$$

where p and $p_1, ..., p_{k-1}$ are prompts for initial and intermediate turns containing the question Q and the task instruction. In each turn, video LLM $\mathcal{M}$ takes the interaction history as input and outputs the next thinking t_k and action, *i. e.* grounding action g_k or answering action a. The thinking t, grounding action g and answering action a are formatted as

$$\begin{aligned} t &\leftarrow \text{<think>thinking process </think>;} \\ g &\leftarrow \text{<ground>start frame, end frame </ground>;} \\ a &\leftarrow \text{<answer>final answer </answer>.} \end{aligned} \tag{2}$$

After each turn with the grounding action, the grounded video clip v_k is concatenated to the history.

3.2 TwG-51K Dataset Construction

To facilitate training, we construct the TwG-51K dataset, comprising 50,744 multiple-choice video QA samples in total. Among these, 8,195 samples have ground-truth question-relevant video grounding annotations, while 42,549 samples lack such annotations. (i) For samples with video grounding annotations, we collect Grounded Question Answering (GQA) data from NExT-GQA [55] and CG-Bench [3], where the video clip that contains important clues for the answer are annotated. NExT-GQA provides 10.5K explicit grounding annotations, and we filter out samples whose videos are shorter than 20 seconds and finally obtain 7,123 samples from NExT-GQA. CG-Bench features long videos between 10 and 80 minutes with video grounding annotations, which consists of 12,129 questions from 1,219 videos. We collect videos from CG-Bench-mini and filter out samples where the IoU between the ground-truth video grounding clip and the entire video is smaller than 0.01 to facilitate training, and finally obtain 1,072 samples. (ii) To scale and diversify training data, we randomly sample a subset of LLaVA-Video-178K [67], a commonly used multi-domain general video instruction tuning dataset consisting of 1.3M questions from 178K videos, and finally obtain 42,549 multiple-choice training samples.

3.3 Multi-Grained Video Representation

Since the initial video is to provide global clues while the dynamically grounded video is to provide detailed information, we adopt multi-grained video representation for different videos. For the initial video input V, we represent it with F_{coarse} frames and f_{coarse} tokens per frame in a coarse-grained way, while F_{fine} frames and f_{fine} tokens per frame in a fine-grained way for the dynamically grounded video segment v_k (generally, $F_{\text{coarse}} > F_{\text{fine}}, f_{\text{coarse}} < f_{\text{fine}}$). To better support videos with various durations and reduce the search scope, in the grounding action, the start and end frames are the frame indexes in the initial video representation, *i.e.* from 0 to $F_{\text{coarse}} - 1$. When cutting the corresponding video clip, they are converted into real seconds.

3.4 Two-stage Reinforced Curriculum Strategy

Since the base video LLM is not capable of benefiting from the *think-with-grounding* stably, to reduce the training difficulty, we design a two-stage curriculum. For the first stage, we train the model with short video GQA data with grounding labels, *i.e.* NExT-GQA part in TwG-51K, serving as cold start, to activate the model to follow the *think-with-grounding* pattern and reach the answer via zooming in the correct video clip. Next, for the second stage, we train the model on the entire TwG-51K dataset, most of which lack grounding labels, to improve the model's generalization via thinking with adaptive grounding.

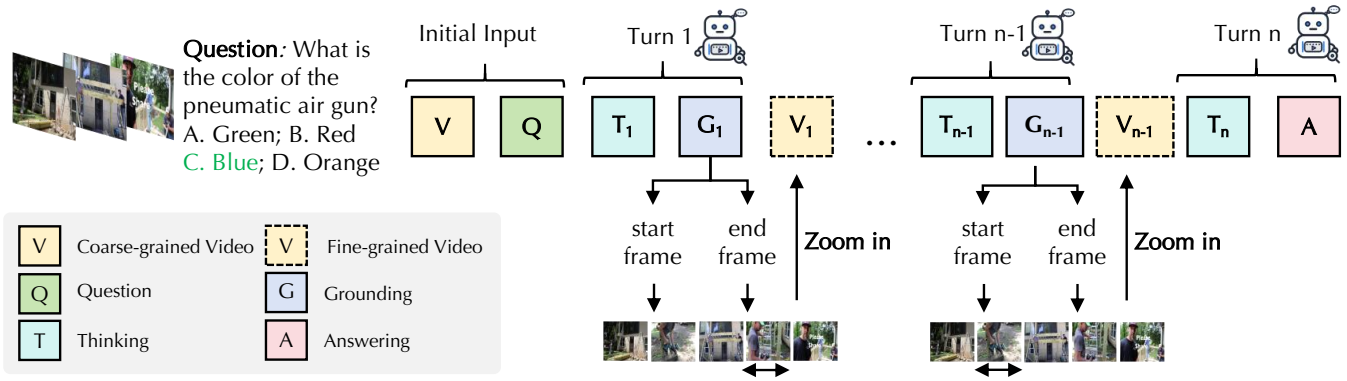


Figure 2: The trajectory of *think-with-grounding*. Initially, the coarse-grained video and question are given and in each turn, the model gives a thinking process and an action based on the history interaction. If the action is grounding, the grounded video clip is zoomed in with a fine-grained representation and added to the context. If the action is answering, the reasoning process stops.

We skip the SFT and adopt curriculum reinforced learning for both stages, since the model is observed to have the potential to think with grounding, which is demonstrated in subsequent experiments. Benefiting from our RL-only solution, our Video-TwG can be easily improved without heavy SFT annotation works in a lighter way.

3.5 TwG-GRPO Algorithm

As shown in Fig. 3, we propose a TwG-GRPO algorithm to support RL training in various data, which is based on Group Relative Policy Optimization (GRPO) [35], for the sake of memory-friendly nature. We adopt multi-turn optimization with trajectory-level reward for computation efficiency, where a trajectory τ with k turns follows the format: [{user: V, p},{assistant: t_1,g_1) },{user: v_1, p_1},...,{assistant: t_{k-1}, g_{k-1} },{user: v_{k-1}, p_{k-1} },{assistant: t_k , a }]. For each sample, we collect rollouts and Algorithm 1 shows the rollout process in our multi-turn *think-with-grounding* setting. After rolling out, we follow the GRPO algorithm to calculate the trajectory advantages in the group:

$$A_i = \frac{R_i - \mu(\{R_j\})}{\sigma(\{R_j\})}, \tag{3}$$

where the $\mu(\{R_j\})$ and $\sigma(\{R_j\})$ is the mean and standard deviation of the rewards in each group. The policy video LLM model is optimized by:

$$r_i(\theta) \triangleq \frac{\mathcal{M}_\theta(\tau_i \mid q)}{\mathcal{M}_{\theta_{\text{old}}}(\tau_i \mid q)}, \qquad \bar{r}_i(\theta) \triangleq \text{clip}\big(r_i(\theta),\, 1-\epsilon,\, 1+\epsilon\big),$$

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\big(r_i(\theta)A_i,\ \bar{r}_i(\theta)A_i\big)\right] - \beta\, D_{\text{KL}}(\mathcal{M}_\theta \,\|\, \mathcal{M}_{\text{ref}})\,. \tag{4}$$

Besides, we deploy dynamic resampling [62] technique when a group does not provide valid learning signals to improve training stability and accelerate convergence.

Since vanilla accuracy and format rewards can not handle complex *think-with-grounding* reasoning for various data, especially unlabeled general QA data, the rewards are carefully designed, which is elaborated on in the following. We adopt accuracy and format rewards according to convention, and for a trajectory with the grounding action, we calculate a grounding-related reward to give feedback to the grounding action, where we design Fine-grained Grounding Reward and Self-Confirmed Pseudo Reward for the training sample with and without grounding labels, respectively. To avoid the optimization conflicts between accuracy and grounding, we apply the accuracy-gated mechanism to give higher priority to accuracy.

Accuracy and Format Rewards. Following [14], the accuracy and format rewards are included. The accuracy reward evaluates the correctness of the answer in the last turn.

$$R_{\text{acc}} = \begin{cases} 1.0 & \text{final answer is correct} \\ 0.0 & \text{otherwise.} \end{cases} \tag{5}$$

The format reward evaluates if a trajectory follows the grounding or answering action format in all turns.

$$R_{\text{format}} = \begin{cases} 0.2 & \text{all turns follow the right format} \\ 0.0 & \text{otherwise.} \end{cases} \tag{6}$$

Fine-grained Grounding Reward. For data with video grounding annotations, the $R_{\text{grounding}}$ is calculated by

$$\begin{aligned} R_{\text{soft}} &= \text{IoU}(g_{\text{last}}, g_{\text{gt}}) \\ R_{\text{hard}} &= 0.5 \cdot \mathbb{I}(\text{IoU}(g_{\text{last}}, g_{\text{gt}}) > 0.0) \\ R_{\text{grounding}} &= R_{\text{soft}} + R_{\text{hard}}, \end{aligned} \tag{7}$$

where g_{last}, g_{gt} represent the last grounding action and the ground truth label. The complementary nature of these reward signals motivates their integration: R_{soft} offers discriminative power yet is relatively weak, whereas R_{hard} delivers high-confidence guidance but is less discriminative. Their combination effectively offsets their respective limitations.

Self-Confirmed Pseudo Reward. When dealing with large amounts of data without grounding labels, it is difficult to calculate $R_{\text{grounding}}$ to evaluate the quality of the grounding action. Distinguishing between better and worse grounding actions is non-trivial, as it has

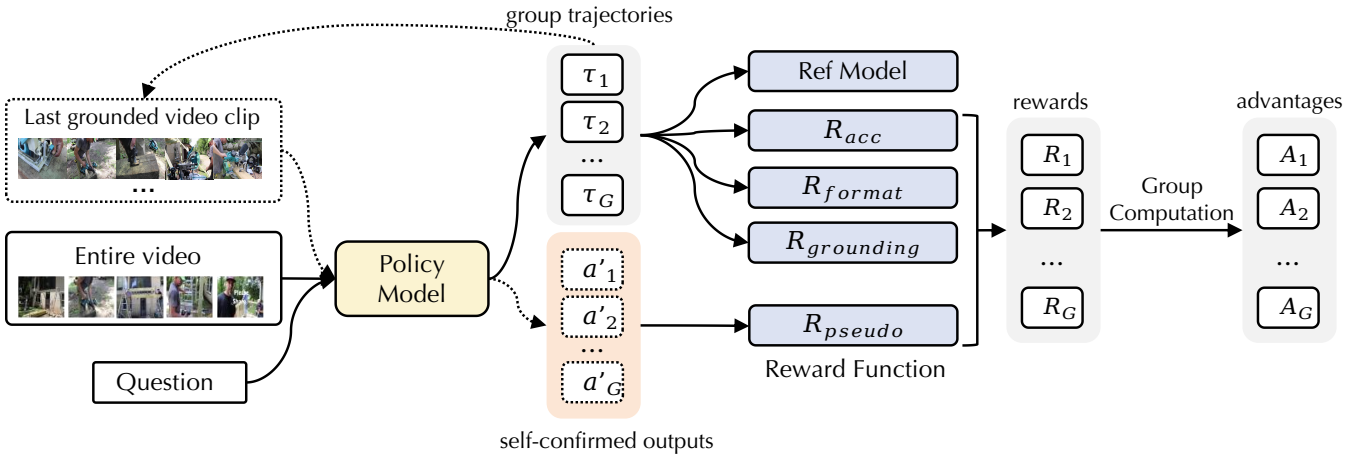


Figure 3: The illustration of our proposed TwG-GRPO algorithm.

been observed that the model can sometimes produce correct answers while grounding video clips that do not contribute to answering the question, leading to unnecessary computational overhead. To address this issue, we propose a Self-Confirmed Pseudo Reward to approximate the effect of $R_{\text{grounding}}$ in encouraging high-quality grounding actions.

The core idea is to use the model itself as the judge. Specifically, we prompt the model to answer the question Q based solely on the last grounded video clip v_{last}. If the model can produce the correct answer using only this clip, then v_{last} is considered more likely to be helpful for answering Q compared to when it fails to do so.

Specifically, for each trajectory containing grounding actions, let the last grounding action be g_{last}, the corresponding video clip be v_{last} (which follows the fine-grained video representation defined in Section 3.1 so as to accelerate with cache), and the ground-truth answer be a_{gt}. We prompt the policy model $\mathcal{M}$ to generate an answer using only v_{last} and $p'(Q)$ where p' is the prompt for this self-confirming instruction:

$$a' = \mathcal{M}(v_{\text{last}}, p'(Q)). \tag{8}$$

Then, the R_{pseudo} is calculated through

$$R_{\text{pseudo}} = \begin{cases} 0, & \text{if the self-confirmed answer is correct,} \\ -\gamma, & \text{otherwise,} \end{cases} \tag{9}$$

where γ is a hyper-parameter, and the reward is implemented in the form of a penalty.

Accuracy-Gated Mechanism. Since the correctness of the answer is of higher priority. Accordingly, the grounding-related reward is awarded only if the trajectory leads to a correct final answer to avoid the optimization conflict.

Above all, the reward can be calculated by

$$R = \begin{cases} R_{\text{acc}} + R_{\text{format}}, & \text{(a)} \\ R_{\text{acc}} + R_{\text{format}} + \mathbb{I}(R_{\text{acc}} > 0) \cdot R_{\text{grounding}}, & \text{(b)} \\ R_{\text{acc}} + R_{\text{format}} + \mathbb{I}(R_{\text{acc}} > 0) \cdot R_{\text{pseudo}}, & \text{(c)} \end{cases} \tag{10}$$

where (a), (b), and (c) correspond to trajectories without grounding actions, with grounding actions for labeled GQA data, and with grounding actions for general QA data, respectively. Algorithm 1 shows the reward calculation briefly.

4 Experiments

4.1 Implementations

We employ Qwen2.5-VL-7B [2] as the base video LLM. For efficient training, we adopt LoRA [15] with a rank of 128 and an alpha of 256. The learning rate is set to 1×10^{-6}. During GRPO training, a batch size of 32 with a group size of 8 is used. The KL constraint coefficient β is set to 0.005, and the coefficient γ in the Self-Confirmed Pseudo Reward is set to 0.1. For RL sampling, the temperature, top-p, top-k, and repetition penalty are configured as 1.0, 0.9, 50, and 1.0, respectively, with a maximum turn count $K = 3$. For coarse-grained video representation during training, we sample $F_{\text{coarse}} = 64$ frames, allocating up to $f_{\text{coarse}} = 16$ tokens per frame. For fine-grained representation, we sample $F_{\text{fine}} = 16$ frames with a maximum of $f_{\text{fine}} = 64$ tokens per frame. During inference, the frame counts remain fixed as above, while the maximum number of tokens per frame is adjusted via input video resolution to align to other baselines on the input length. We evaluate our Video-TwG under two resolution settings: (i) LR (Low Resolution): Matches the training configuration, allowing at most 16 tokens per frame for coarse-grained and 64 tokens per frame for fine-grained representations. (ii) HR (High Resolution): Allows at most 128 tokens per frame for coarse-grained and 512 tokens per frame for fine-grained representations. In evaluation, temperature is set to 0 and we set a maximum turn count $K = 3$ as well. If a prediction cannot be automatically parsed, we revert to the training sampling hyperparameters and retry up to three times. For a fair comparison, both our base model Qwen2.5-VL-7B and other evaluated models that support reasoning are assessed in reasoning mode. Our training and evaluation framework is built upon ms-swift [68] and VLMEvalKit [9]. We leverage vLLM [19] to accelerate the sampling process. Our training can be implemented on 2 H800 80G GPUs with early-stop strategy.

Table 1: Results of our Video-TwG and baselines on three mainstream long video benchmarks: Video-MME, LongVideoBench, MLVU. For MLVU, we evaluate the multiple choices questions. * means our reproduced results. ↑ is the relative gain against the Qwen2.5-VL-7B. The best and second are marked with bold and underline.

Method	Video-MME				LongVideoBench					MLVU-MC
	short	medium	long	overall	(8s,15s]	(15s,60s]	(180s,600s)	(900s,3600s]	overall	
Video-LLaVA-7B	45.3	38.0	36.2	39.9	-	-	-	-	39.1	47.3
ShareGPT4Video-8B	48.3	36.3	35.0	39.9	-	-	-	-	39.7	46.4
VideoChat2-7B	-	-	33.2	39.5	-	-	-	-	-	-
LongVA-7B	61.1	50.4	46.2	52.6	-	-	-	-	47.8	56.3
Long-LLaVA-7B	61.9	51.4	45.4	52.9	-	-	-	-	-	-
LongLLaVA-9B	59.6	50.3	42.7	50.9	-	-	-	-	51.9	-
Video-XL-7B	64.0	53.2	49.2	55.5		-	-	-	50.7	-
Video-R1-7B*	69.3	53.6	49.1	57.3	-	-	-	-	55.2	59.8
Long-VILA-R1-7B*	-	-	-	34.6	-	-	-	-	35.5	36.6
VideoChat-R1-7B*	-	-	-	54.4	-	-	-	-	40.8	57.0
VideoAgent	53.3	49.7	37.8	46.9	54.6	55.2	45.1	43.5	47.1	52.5
VideoTree	55.5	49.2	39.3	48.0	61.0	57.5	48.4	44.6	49.7	51.6
Qwen2.5-VL-7B(LR)	56.2	43.2	40.4	46.6	51.3	57.0	42.5	39.7	44.4	47.5
Video-TwG(LR)	63.7$_{7.5\uparrow}$	49.4$_{6.2\uparrow}$	47.7$_{7.3\uparrow}$	53.6$_{7.0\uparrow}$	59.8$_{8.5\uparrow}$	62.8$_{5.8\uparrow}$	47.3$_{4.8\uparrow}$	44.1$_{4.4\uparrow}$	49.7$_{5.3\uparrow}$	54.6$_{7.1\uparrow}$
Qwen2.5-VL-7B(HR)	71.2	53.8	46.7	57.2	63.5	69.8	49.5	45.6	52.4	55.3
Video-TwG(HR)	**71.9**$_{0.7\uparrow}$	**57.1**$_{3.3\uparrow}$	**50.0**$_{3.3\uparrow}$	**59.7**$_{2.5\uparrow}$	**73.5**$_{10.0\uparrow}$	**72.7**$_{2.9\uparrow}$	**54.1**$_{4.6\uparrow}$	**47.2**$_{1.6\uparrow}$	**56.3**$_{3.9\uparrow}$	**60.3**$_{5.0\uparrow}$

4.2 Baselines

The baselines used in our experiments are categorized into four groups: (i) General video LLMs, including Video-LLaVA [24], ShareGPT4Video [4], and VideoChat2 [21]; (ii) Long-context video LLMs, such as LongVA [65], Long-LLaVA [61], LongLLaVA [46], and Video-XL [38]; (iii) Video agentic systems for long video understanding, represented by VideoAgent [47] and VideoTree [50]; and (iv) Video reasoning models, including Video-R1 [10], VideoChat-R1 [22], and Long-VILA-R1 [5].

4.3 Benchmarks

We evaluate our method on multiple-choice questions across three mainstream benchmarks for long video understanding: Video-MME [12], LongVideoBench [53], and MLVU [70]. Video-MME consists of 2,700 manually annotated multiple-choice questions derived from 900 videos. The videos are divided into three groups based on duration: short (0-2 minutes), medium (4-15 minutes), and long (30-60 minutes). LongVideoBench contains 3,763 human-labeled questions spanning 6,678 videos, with an average duration of 472 seconds. The videos are categorized into four duration ranges: 8-15 seconds, 15-60 seconds, 180-600 seconds, and 900-3,600 seconds. We evaluate models on its validation split, which includes 1,337 multiple-choice questions. MLVU features videos ranging from 3 to 120 minutes, with an average duration of 15 minutes. It provides both Dev and Test splits, comprising multiple-choice and open-ended questions. In this work, we focus on the multiple-choice questions in the Dev split, which contains 2,174 questions covering seven distinct question types. For all benchmarks, subtitles are not used during evaluation.

4.4 Main Results

Table 1 summarizes the performance of Video-TwG and a wide range of baselines on three representative benchmarks for long video understanding. Overall, Video-TwG achieves consistently strong performance across datasets with diverse video durations, outperforming general video LLMs, long-context video LLMs, as well as recent reasoning and agent-based approaches.

Compared with general-purpose video LLMs, Video-TwG yields substantial improvements on all benchmarks. On Video-MME, Video-TwG (HR) achieves an overall accuracy of 59.7%, significantly outperforming general video LLMs that typically obtain around 40%. It also attains 50.0% accuracy on long videos, exceeding prior methods by a large margin, indicating that models primarily designed for short or medium-length videos struggle to generalize to long-duration scenarios. Video-TwG also consistently outperforms long-context models including LongVA, Long-LLaVA, LongLLaVA, and Video-XL. For example, on Video-MME, Video-TwG (HR) exceeds Video-XL by over 4 points in overall accuracy, suggesting that effective long-range reasoning is crucial beyond simply increasing the input context length. Recent video reasoning models, such as Video-R1, enhance long video understanding by introducing explicit reasoning mechanisms. While effective, these methods rely on reasoning under fixed video context which is not enough for very long temporal contexts, since the model is observed to make up facts confidently when the video information is not enough, which is shown in the case study. In contrast, Video-TwG achieves stronger performance via grounding and zooming in, indicating more effective long-range temporal modeling. Agent-based systems,

Algorithm 1 The process of rollout and reward calculation in our TwG-GRPO algorithm.

Input: The policy video LLM $\mathcal{M}_\theta$; video V; question Q; ground truth answer a_{gt}; ground truth video grounding label g_{gt} if there is; self-confirmed pseudo reward hyperparameter γ; The maximum number of turns K; multi-grained video representation arguments $(F_{\text{coarse}}, f_{\text{coarse}})$ and $(F_{\text{fine}}, f_{\text{fine}})$.
Output: The trajectory τ; reward R.

$v \leftarrow \text{VideoProcessor}(V, F_{\text{coarse}}, f_{\text{coarse}})$
$\tau \leftarrow [v, p(Q)]$
for k **in** $1, ..., K$ **do**
  $o_k \leftarrow \mathcal{M}_\theta(\tau)$ ▷ Get the current output.
  if o_k matches "grounding" format **then**
    $t_k, g_k \leftarrow \text{ParseGrounding}(o_k)$
    $v_k \leftarrow \text{VideoProcessor}(\text{VideoCut}(V, g_k), F_{\text{fine}}, f_{\text{fine}})$
    $\tau \leftarrow \tau + [t_k, g_k, v_k, p_k(Q)]$ ▷ Conduct video grounding.
  else if o_k matches "answering" format **then**
    $t_k, a \leftarrow \text{ParseAnswering}(o_k)$
    $\tau \leftarrow \tau + [t_k, a]$ ▷ Obtain the answer and stop.
    Break
  else
    $\tau \leftarrow \tau + [o_k]$ ▷ Cannot parse it and stop.
    Break
  end if
end for
$R \leftarrow R_{\text{acc}} + R_{\text{format}}$
if τ has grounding actions **then**
  $g_{\text{last}} \leftarrow \text{ExtractLastGrounding}(\tau)$
  if the sample has label g_{gt} **then**
    $R_{\text{grounding}} \leftarrow \text{GroundingReward}(g_{\text{last}}, g_{\text{gt}})$ ▷ Eq. (7)
    $R \leftarrow R + \mathbb{I}(R_{\text{acc}} > 0) \cdot R_{\text{grounding}}$
  else
    $v_{\text{last}} \leftarrow \text{VideoProcessor}(\text{VideoCut}(V, g_{\text{last}}), F_{\text{fine}}, f_{\text{fine}})$
    $a' \leftarrow \mathcal{M}_\theta(v_{\text{last}}, p'(Q))$
    $R_{\text{pseudo}} \leftarrow \text{PseudoReward}(a', a_{\text{gt}}, \gamma)$ ▷ Eq. (9)
    $R \leftarrow R + \mathbb{I}(R_{\text{acc}} > 0) \cdot R_{\text{pseudo}}$
  end if
end if
return τ, R

including VideoAgent and VideoTree, address long video understanding through explicit planning or hierarchical decomposition. Despite their improvements over earlier baselines, they depend on complex multi-stage pipelines that are prone to error accumulation and heavy computations, like captioning. Video-TwG consistently outperforms VideoTree by 11.7, 6.6 and 8.7 points, and VideoAgent by 12.8, 9.2 and 7.8 points on these benchmarks, respectively, while maintaining a simpler inference paradigm.

Furthermore, under both low-resolution (LR) and high-resolution (HR) settings, Video-TwG consistently outperforms its backbone Qwen-2.5-VL-7B across all benchmarks. The improvements are more pronounced in the low-resolution setting, while models trained under low-resolution still generalize effectively to high-resolution inference. This indicates that the gains primarily stem from the proposed training paradigm, making Video-TwG particularly effective for real-world and resource-constrained scenarios.

Table 2: Ablation experiments of different variants of the base Qwen2.5-VL-7B on Video-MME.

Method	Video-MME			
	short	medium	long	overall
Base[1]	56.2	43.2	40.4	46.6
Base-TwG-CoT[1]	56.9	43.2	42.1	47.4
Base-Think[2]	59.2	46.4	43.2	49.6
Base-MultiGrained[4]	63.1	45.9	42.7	50.6
Base-Think-MultiGrained [5]	63.0	46.9	45.3	51.7
Video-TwG	63.7	49.4	47.7	53.6

[1] Base: Qwen2.5-VL-7B.
[2] Base-TwG-CoT: Base is prompted to perform *think-with-grounding* using the same prompts as Video-TwG.
[3] Base-Think: Base is trained on the TwG-51K dataset to conduct text-only reasoning without grounding.
[4] Base-MultiGrained: Base receives both coarse- and fine-grained representations of the entire video as input.
[5] Base-Think-MultiGrained: Base-Think receives both coarse- and fine-grained representations of the entire video as input.

4.5 Ablation Study

We further conduct ablation study to explore the contributions of the core components of our Video-TwG method on the Video-MME benchmark. By default, we adopt the low-resolution setup for its friendly computational costs.

Analysis on the *think-with-grounding* paradigm. In Tab. 2, we show several experiments on the base Qwen2.5-VL-7B model for analysis on the *think-with-grounding* paradigm itself. These experiments are conducted to answer these questions. *(i) Can the model benefit from think-with-grounding?* We directly prompt the base model to conduct *think-with-grounding* with the same prompts as applied in Video-TwG (Base-TwG-CoT in Tab. 2). We can see that in the short and medium splits, the improvement is trivial but it has a 1.7 performance gain in the long video split, showing that the base model has the potential to conduct *think-with-grounding* reasoning to improve the performance, especially in long videos. However, there is still a significant gap between Base-TwG-CoT and our Video-TwG, proving the effectiveness of our method to further enhance *think-with-grounding* ability. *(ii) Can think-with-grounding go beyond traditional text-only reasoning?* We train the base model on the same TwG-51K dataset to make the model conduct text-only reasoning without grounding actions with the GRPO algorithm, whose results are shown in "Base-Think" in Tab. 2. Although it achieves improvement compared to the base, there is still a significant gap compared to our method. It is 4.5 lower than our method in long videos. Our method allows the model to dynamically load fine-grained video to context, which improves the model performance, while text-only RL only helps the model to better grasp the clues in the given fixed multimodal input. *(iii) Since think-with-grounding dynamically visits more video input, can such paradigm go beyond directly scaling the video input length?* In the Video-MME benchmark,

our Video-TwG is observed to conduct 0.65 grounding actions on average per sample. To answer this question, we increase the video input length for the base model, as shown in Base-MultiGrained and Base-Think-MultiGrained in Tab. 2, and we feed the Base and Base-Think with both the coarse- and fine-grained representation of the entire video, which is equivalent to conducting 1 grounding action per sample on the video context length. It shows that such vanilla scaling nearly catches up with our Video-TwG in short videos but there is a significant gap of 5.0(Base-MultiGrained) and 2.4(Base-Think-MultiGrained) in long videos, showing the necessity of intelligent selection of video clip, especially in long videos.

Analysis on our two-stage curriculum training strategy. We compare our Video-TwG with two variants, when it's trained with stage 1 or 2 only. The results are shown in Fig. 4, which shows that the performance drops significantly when either of training stages is skipped. It's noted that if skipping the stage 1, the model's grounding action vanishes quickly in the early stage of the training during stage 2, indicating the necessity of stage 1 as a good cold-start, and the model's performance in long videos is 1.4 inferior to when it is trained in stage 1. When equipped with stage 1 and stage 2, the model gradually grasp the *think-with-grounding* ability and achieves the best performance.

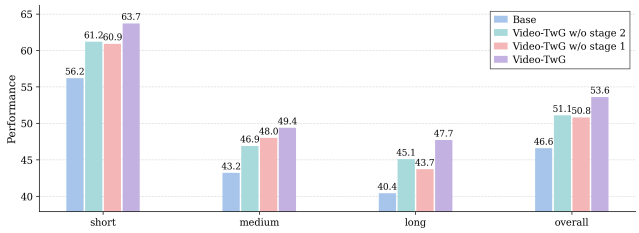
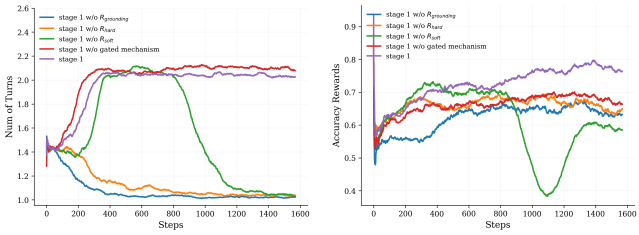


Figure 4: Ablation results on Video-MME of our models with different training stages.



(a) The number of turns **(b) The accuracy reward**

Figure 5: Training curves of several variants of our Video-TwG in stage 1. The curves are smoothed for aesthetics. It's the same below.

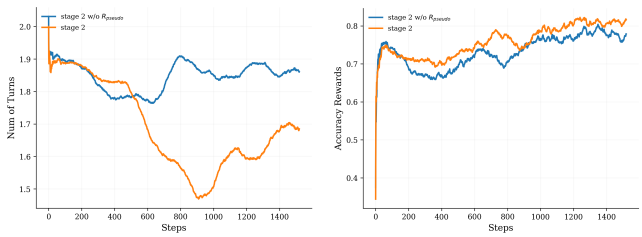
Analysis around the grounding reward in stage 1 training. We explore the roles of several core factors in stage 1 training, the accuracy-gated mechanism, the grounding reward $R_{\text{grounding}}$, the soft grounding reward R_{soft}, and the hard grounding reward R_{hard}. Fig. 5 shows the curves of the number of turns and the accuracy reward during the stage 1 training. We can see that without $R_{\text{grounding}}$ or without R_{hard} in $R_{\text{grounding}}$, the grounding action vanishes along with the training process, showing that the signal provided by R_{soft} is too weak to boost the model to learning *think-with-grounding*. Only with R_{hard} in $R_{\text{grounding}}$, the training is unstable: the model collapses at around step 800, when the model quickly forgets the grounding action. Compared to R_{soft}, R_{hard} fails to provide diversified and differentiated rewards in a group, which makes the training process less stable compared to $R_{\text{soft}} + R_{\text{hard}}$. Without the accuracy-gated mechanism, the model learns slightly more quickly to think with grounding; however, it achieves much lower accuracy rewards than our standard model in stage 1. The grounding and accuracy targets may conflict on some samples, thereby affecting the performance. We conduct a quantitative evaluation for the model without the accuracy-gated mechanism in stage 1, as shown in Tab. 3, which shows it is 1.5 and 0.5 lower than the standard model in the long and overall splits.

Table 3: The ablation results of the accuracy-gated mechanism in stage 1 training.

Method	Video-MME			
	short	medium	long	overall
stage 1 w/o gated mechanism	60.2	48.1	43.6	50.6
stage 1	61.2	46.9	45.1	51.1

Table 4: The ablation results of the self-confirmed pseudo reward in stage 2 training.

Method	Video-MME			
	short	medium	long	overall
Video-TwG w/o R_{pseudo}	63.4	48.0	47.2	52.9
Video-TwG	63.7	49.4	47.7	53.6



(a) The number of turns. **(b) The accuracy rewards.**

Figure 6: Training curves of several variants of our Video-TwG in stage 2.

Analysis on the Self-Confirmed Pseudo Reward R_{pseudo} in stage 2 training. Fig. 6 illustrates the training dynamics of the number of turns and accuracy reward during stage 2. Both settings exhibit a similar trend: the number of grounding actions first decreases and then increases. This suggests that grounding is not always necessary for all questions, the model learns to answer questions with fewer grounding actions in the early stage of training,

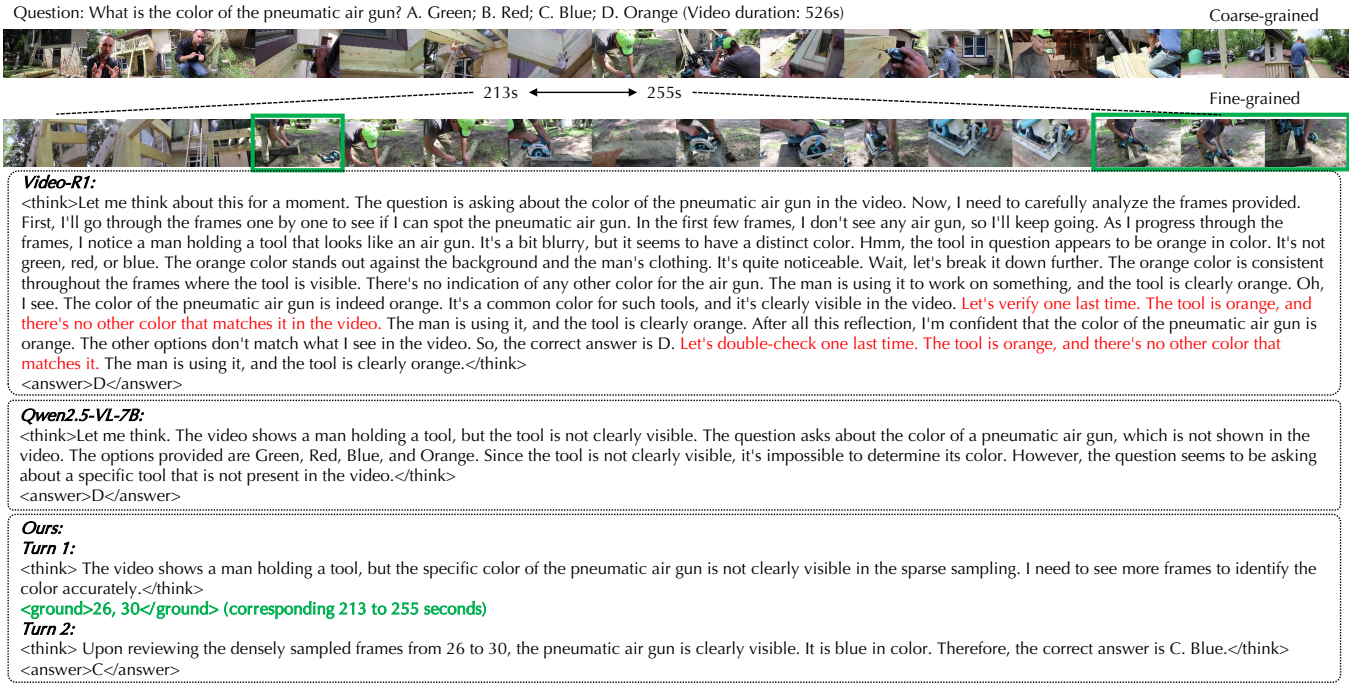


Figure 7: The case study on Video-R1, Qwen2.5-VL-7B, Our Video-TwG. The location of the crucial video clues for this question is marked with green boxes.

leading to a reduction in turns. As training progresses, the model realizes that additional grounding can further improve accuracy, and the number of turns begins to rise. Correspondingly, accuracy rewards start to increase more steadily after the turning points (around 700 steps for "stage 2 w/o R_{pseudo} " and 900 steps for "stage 2"). With the introduction of the Self-Confirmed Pseudo Reward R_{pseudo}, the model learns to invoke grounding actions more adaptively, using fewer grounding steps on average while achieving higher accuracy rewards throughout training. As shown in Tab. 4, removing R_{pseudo} leads to a slight performance drop, demonstrating its effectiveness in reducing redundant grounding actions without sacrificing overall performance.

We further analyze the number of samples containing grounding actions for Video-TwG w/o R_{pseudo} and Video-TwG in Tab. 5. Video-TwG reduces the number of such samples by 15.7%, approximately 300, indicating more selective grounding behavior. To assess the grounding quality, we evaluate QA performance when only the last grounded video clip is provided as input as an approximation to the grounding quality. For objectivity, we employ InternVL-3.5-38B [44] as an external evaluator, sampling 32 frames with 256 tokens per frame. Results in Tab. 6 show that grounding clips produced by standard Video-TwG yield better QA performance, with gains of 2.1 and 1.6 on the long and overall splits, respectively.

Overall, these results indicate that R_{pseudo} encourages more effective and cautious grounding strategies, particularly for reducing useless exploration in long videos. The model learns to avoid unnecessary grounding actions that do not contribute to answering the question, leading to both higher efficiency and better grounding quality.

Table 5: The number of samples that has grounding action on Video-MME benchmark.

Method	Video-MME			
	short	medium	long	overall
Video-TwG w/o R_{pseudo}	682	687	725	2094
Video-TwG	615	584	566	1765

Table 6: The evaluation results on the grounded video clip on Video-MME by InternVL-3.5-38B for our Video-TwG w/ and w/o R_{pseudo}.

Method	Video-MME			
	short	medium	long	overall
Video-TwG w/o R_{pseudo}	65.5	55.3	51.4	57.3
Video-TwG	68.1	54.5	53.5	58.9

4.6 Case Study

In this section, we present a qualitative case study comparing the responses of Video-R1, Qwen2.5-VL-7B, and our method, which is shown in Fig. 7. Given a fixed and sparse video context from a long video, Video-R1 produces substantially longer reasoning traces than Qwen2.5-VL-7B; however, its reasoning process is chaotic and leads to hallucinated conclusions, confidently identifying the tool as orange despite being incorrect. The base model Qwen2.5-VL-7B

recognizes that key visual clues are missing in the initial coarse-grained video input, which ultimately results in an incorrect answer. In contrast, our method explicitly expresses uncertainty in its initial prediction and performs a grounding action in the first turn. After receiving fine-grained visual input from the grounded video clip, it successfully identifies the correct answer.

5 Conclusion

In this work, we introduce Video-TwG, a curriculum reinforced framework that enables video LLMs to think with grounding in a multi-turn reasoning process, zooming into question-relevant clips only when necessary. To support scalable training, we construct the TwG-51K dataset and propose a two-stage reinforced curriculum strategy that cold-starts on short-video GQA with grounding labels and then scales to the full dataset to improve generalization. Moreover, to handle complicated *think-with-grounding* reasoning for various kinds of data, we develop a TwG-GRPO algorithm which features the fine-grained grounding reward, self-confirmed pseudo reward, and accuracy-gated mechanism. Extensive experiments on Video-MME, LongVideoBench, and MLVU demonstrate consistent advantages over strong baselines. Our results validate the effectiveness of our two-stage reinforced curriculum strategy: training with low-resolution inputs consistently enhances the base model's *think-with-grounding* ability across different inference resolutions, while TwG-GRPO better leverages diverse unlabeled data to improve grounding quality and reduce redundant groundings without sacrificing QA performance. It should be noted that answering some questions might need multiple non-contiguous segments, where it's not efficient to conduct one segment grounding per action, which is a limitation of our method. In future work, we plan to go beyond trajectory-level optimization by developing fine-grained process reward modeling and step-level training, and achieve multiple-segment grounding simultaneously, aiming to build more powerful and efficient long video reasoning models.

Acknowledgments

This work was supported by the National Natural Science Foundation of China No. 62425604 and Beijing National Research Center For Information Science And Technology Grant Number No. BNR2026TD03005.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*. 5803–5812.

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

[3] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. 2024. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075* (2024).

[4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems* 37 (2024), 19472–19495.

[5] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. 2025. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966* (2025).

[6] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. 2024. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188* (2024).

[7] Andong Deng, Zhongpai Gao, Anwesa Choudhuri, Benjamin Planche, Meng Zheng, Bin Wang, Terrence Chen, Chen Chen, and Ziyan Wu. 2025. Seq2Time: Sequential Knowledge Transfer for Video LLM Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13766–13775. https://openaccess.thecvf.com/content/CVPR2025/html/Deng_Seq2Time_Sequential_Knowledge_Transfer_for_Video_LLM_Temporal_Grounding_CVPR_2025_paper.html

[8] Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. 2025. ReWind: Understanding Long Videos with Instructed Learnable Memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13734–13743.

[9] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11198–11201.

[10] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776* (2025).

[11] Wei Feng, Xin Wang, Hong Chen, Yu-Wei Zhan, Zihan Song, Bin Huang, Kecheng Zheng, and Wenwu Zhu. [n. d.]. Cross-granularity Video LLM for Long Video Understanding. ([n. d.]).

[12] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24108–24118.

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*. 5267–5275.

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

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

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

[17] Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. 2025. Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://neurips.cc/virtual/2025/poster/119777

[18] Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. 2025. On the Consistency of Video Large Language Models in Temporal Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13713–13722. https://openaccess.thecvf.com/content/CVPR2025/html/Jung_On_the_Consistency_of_Video_Large_Language_Models_in_Temporal_CVPR_2025_paper.html

[19] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[20] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. 2023. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 2 (2023), 1–33.

[21] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.

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

[23] Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. 2025. Universal Video Temporal Grounding with Generative Multi-modal Large Language Models. arXiv:2506.18883 [cs.CV] https://arxiv.org/abs/2506.18883

[24] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*. 5971–5984.

[25] Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. 2025. BOLT: Boost Large Vision-Language Model Without Training for Long-form Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3318–3327.

[26] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. 2025. When Thinking Drifts: Evidential Grounding for Robust Video Reasoning. In *Advances in Neural*

Information Processing Systems (NeurIPS). https://openreview.net/forum?id=qDm3fpLYDW

[27] Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093* (2024).

[28] Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. 2024. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems* 37 (2024), 23464–23487.

[29] Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, Cordelia Schmid, and Tobias Weyand. 2025. MINERVA: Evaluating Complex Video Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 23968–23978. https://openaccess.thecvf.com/content/ICCV2025/html/Nagrani_MINERVA_Evaluating_Complex_Video_Reasoning_ICCV_2025_paper.html

[30] Baoqi Pei, Yifei Huang, Jilan Xu, Yuping He, Guo Chen, Fei Wu, Jiangmiao Pang, and Yu Qiao. 2025. EgoThinker: Unveiling Egocentric Reasoning with Spatio-Temporal CoT. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=9Zwl2Ly28N

[31] Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. 2025. Enrich and Detect: Video Temporal Grounding with Multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 24297–24308. https://openaccess.thecvf.com/content/ICCV2025/html/Pramanick_Enrich_and_Detect_Video_Temporal_Grounding_with_Multimodal_LLMs_ICCV_2025_paper.html

[32] Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. 2025. Vamba: Understanding hour-long videos with hybrid mamba-transformers. *arXiv preprint arXiv:2503.11579* (2025).

[33] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549* (2025).

[34] Saul Santos, António Farinhas, Daniel C McNamee, and André FT Martins. 2025. Infinity-Video: A Training-Free Approach to Long Video Understanding via Continuous-Time Memory Consolidation. *arXiv preprint arXiv:2501.19098* (2025).

[35] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[36] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434* (2024).

[37] Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Yi-Fan Zhang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Shaohui Lin, et al. 2025. Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuracy. *arXiv preprint arXiv:2502.05177* (2025).

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

[39] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18221–18232.

[40] Zihan Song, Xin Wang, Zi Qian, Hong Chen, Longtao Huang, Hui Xue, and Wenwu Zhu. 2025. Modularized self-reflected video reasoner for multimodal llm with application to video question answering. In *Forty-second International Conference on Machine Learning*.

[41] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive Keyframe Sampling for Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 29118–29128.

[42] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025. DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18992–19001.

[43] Han Wang, Yuxiang Nie, Yongjie Ye, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. 2025. Dynamic-VLM: Simple Dynamic Visual Token Compression for VideoLLM. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20812–20823.

[44] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265* (2025).

[45] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. 2025. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22958–22967.

[46] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. 2024. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889* (2024).

[47] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*. Springer, 58–76.

[48] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. 2024. Hawkeye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228* (2024).

[49] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. 2025. Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://neurips.cc/virtual/2025/poster/116757

[50] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3272–3283.

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[52] Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1884–1894.

[53] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems* 37 (2024), 28828–28857.

[54] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. 2025. Number it: Temporal Grounding Videos like Flipping Manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13754–13765. https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Number_it_Temporal_Grounding_Videos_like_Flipping_Manga_CVPR_2025_paper.html

[55] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13204–13214.

[56] Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. 2025. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478* (2025).

[57] Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2025. Vca: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20168–20179.

[58] Zuhao Yang, Yingchen Yu, Yunqing Zhao, Shijian Lu, and Song Bai. 2025. TimeExpert: An Expert-Guided Video LLM for Video Temporal Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 24286–24296. https://openaccess.thecvf.com/content/ICCV2025/html/Yang_TimeExpert_An_Expert-Guided_Video_LLM_for_Video_Temporal_Grounding_ICCV_2025_paper.html

[59] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[60] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. 2025. Re-thinking Temporal Search for Long-Form Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8579–8591.

[61] Yin Song and Chen Wu and Eden Duthie. 2024. aws-prototyping/long-llava-qwen2-7b. https://huggingface.co/aws-prototyping/long-llava-qwen2-7b

[62] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476* (2025).

[63] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. 2025. TimeSuite: Improving MLLMs for Long Video Understanding via Grounded Tuning. In *International Conference on Learning Representations (ICLR)*. arXiv:2410.19702 [cs.CV] https://openreview.net/forum?id=nAVejJURqZ Poster.

[64] Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. 2025. VTimeCoT: Thinking by Drawing for Video Temporal Grounding and Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 24203–24213. https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_VTimeCoT_Thinking_by_Drawing_for_

Video_Temporal_Grounding_and_Reasoning_ICCV_2025_paper.html

[65] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852* (2024).

[66] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895* (2025).

[67] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713* (2024).

[68] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517 [cs.CL] https://arxiv.org/abs/2408.05517

[69] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362* (2025).

[70] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13691–13701.