EgoExo-Con: Exploring View-Invariant Video Temporal Understanding

Minjoon Jung[1], Junbin Xiao[2,3⋆], Junghyun Kim[1], Byoung-Tak Zhang[1], and Angela Yao[3]

[1] Seoul National University
[2] University of Science and Technology of China
[3] National University of Singapore
{mjjung, jhkim, btzhang}@bi.snu.ac.kr, junbinxiao@ustc.edu.cn, ayao@comp.nus.edu.sg

Abstract. Do Video-LLMs have *consistent* temporal understanding when videos capture the same event from different viewpoints? To study this question, we introduce EgoExo-Con(sistency), a benchmark of synchronized egocentric and exocentric video pairs with human-refined queries that ensure all concepts are visible in both viewpoints. EgoExo-Con emphasizes two temporal understanding tasks: Temporal Verification and Temporal Grounding. It evaluates not only correctness but consistency across viewpoints. Our analysis reveals two critical limitations of existing Video-LLMs: (1) models often fail to maintain consistency, with results far worse than their single-view performances. (2) When naively finetuned with synchronized videos of both viewpoints, the models show improved consistency but often underperform those trained on a single view. For improvements, we propose View-GRPO, a novel reinforcement learning framework that effectively strengthens view-specific temporal reasoning while encouraging consistent comprehension across viewpoints. Our method demonstrates its superior temporal understanding capabilities, especially for improving cross-view consistency. All resources have been made available at EgoExo-Con.

1 Introduction

Recent advances in video large language models (Video-LLMs) have shown impressive capabilities in question answering [39, 40, 59] and temporal grounding [16, 32, 41, 54], and are stepping towards fine-grained and long-range reasoning [8, 35, 42, 57]. However, most benchmarks [9, 28, 44, 61] and methods assume a fixed or minimally varying viewpoint, *e.g.* third-person view (exo) videos. This begs a critical question: *Do Video-LLMs achieve consistent temporal understanding across different camera perspectives?*

Often, videos of the same event appear strikingly different when captured from different perspectives. A cooking demonstration filmed from a head-mounted camera (ego view) looks unlike a side-mounted tripod shot (exo view). Yet, the

⋆ Corresponding Author

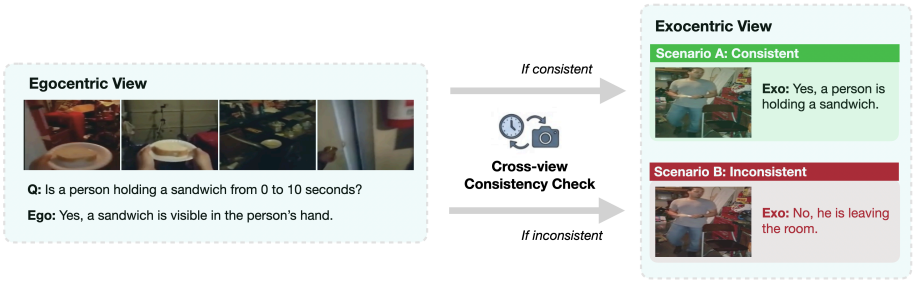

Fig. 1: Evaluation setup in EgoExo-Con. The same query is applied to synchronized egocentric and exocentric video pairs, and model responses are checked for consistency.

underlying temporal dynamics, such as cutting vegetables and stirring a pot, are identical. For humans, this view variation rarely impedes understanding; we easily track the sequence of actions and localize their temporal moments across viewpoints. This makes temporal reasoning particularly critical: while appearance cues can vary drastically with viewpoint, the temporal structure of events is invariant. Thus, evaluation of cross-view consistency is essential and can be effectively carried out through temporal understanding tasks; however, such capabilities remain largely underexplored in current Video-LLMs.

To study this, we introduce EgoExo-Con, a benchmark comprising 1,148 synchronized videos and 2,269 human-refined temporal-bounded event queries, to evaluate whether models can provide consistent predictions across viewpoints - a key indicator of view-invariant video-language understanding. The benchmark focuses on two temporal understanding tasks: *temporal verification* [19] and *temporal grounding* [10]. Temporal verification is a question-answering (QA) task that asks whether a given event occurs within a specific video moment, while temporal grounding requires identifying the relevant video moment (start and end timestamps) corresponding to an event query. As shown in Fig. 1, we ask the same event but with synchronized videos of different viewpoints, and check if the tested models can output correct and consistent responses.

We evaluate both the advanced closed-source models [5, 37] and open-source Video-LLMs comprising general-purpose [2, 4, 20, 55] and time-aware variants [16, 19, 32, 54]. Our results reveal that all models, especially the open-source ones, struggle with cross-view consistency. They generally exhibit a modest performance gap between individual ego and exo videos, but achieve consistency scores barely over half their single-view performance in both tasks. This indicates that the relatively stable performances across viewpoints may be sourced from view-specific biases rather than robust cross-view temporal understanding.

Our further investigations show that naive multi-view supervised fine-tuning (SFT) with synchronized video-language data is insufficient. In fact, it even underperforms the counterpart with single-view training, indicating that naively merging viewpoints may introduce conflicting priors that undermine temporal signals and consistency. Furthermore, the improvements are unstable and ineffective even when the visual encoder is unfrozen during training. These collec-

tively suggest that *viewpoint variation remains a significant challenge for current Video-LLMs in robust video temporal understanding.*

To that end, we propose View-GRPO, a reinforcement learning (RL) framework for view-invariant video temporal understanding. It encourages the model to learn consistent cross-view temporal reasoning as well as certain view-specific details. Experiments demonstrate that View-GRPO yields more robust and consistent video understanding on our benchmark and generalizes well to other video understanding benchmarks. In summary, EgoExo-Con establishes a new paradigm for evaluating and improving view-invariant temporal understanding in Video-LLMs. We hope it will foster future research on models that truly capture the essence of dynamic events independent of perspective. Our primary contributions are as follows:

- We study the robustness of Video-LLMs in cross-view video temporal understanding and introduce EgoExo-Con, a synchronized ego–exo benchmark constructed with manual annotation efforts.
- We reveal that current Video-LLMs achieve cross-view consistency barely better than half of their single-view performance, and naively blending perspectives for training could introduce conflicting priors, undermining consistency.
- We propose View-GRPO, a reinforced approach that explicitly strengthens temporal reasoning while encouraging view-invariant comprehension, significantly improving robust and consistent video temporal understanding.

2 Related Work

Video Large Language Models. Video-LLMs [2, 5, 20, 39, 55, 59] integrate pretrained video representations into powerful LLMs [2, 11, 51] to enable chatting about videos. While advances to date are mostly achieved in short and coarse-grained question answering [49, 53], more recent Video-LLMs [13, 16, 19, 23, 29, 31, 32, 43, 54] have explored grasping fine-grained temporal moments ("when"). However, all of these models solve videos captured in a single camera viewpoint (either exo or ego) and do not evaluate whether temporal reasoning remains stable across views of the same event. This work thus fills such gap by conducting a comprehensive analysis.

Ego-Exo Benchmarks. Most benchmarks target either exocentric [9, 10, 46, 47, 53] or egocentric [3, 7, 28, 45, 52] video understanding, with only a few offering paired views: CharadesEgo [36], LEMMA [18], EgoExo-4D [12], Assembly101 [33], EgoExo-Fitness [22], and EgoExOR [30]. Yet, they are either domain-specific or do not evaluate cross-view temporal reasoning. A concurrent effort, EgoExoBench [14], also explores Video-LMMs in cross-view temporal reasoning, but it primarily targets action ordering via multi-choice selection. While it highlights viewpoint gaps, it neither explicitly addresses prediction consistency nor proposes a concrete methodology. In contrast, we propose EgoExo-Con to evaluate temporal understanding and prediction consistency across synchronized viewpoints with View-GRPO for improvements.

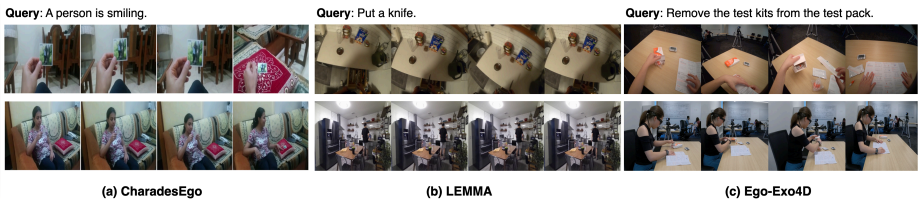

Fig. 2: Examples of queries and corresponding video moments from existing datasets. (a) and (b) highlight fundamental limitations, with the egocentric view (top) in (a) being insufficient due to differing focuses, and the exocentric view (bottom) in (b) being ambiguous due to occlusion and distance. Although the query in (c) is identifiable from both viewpoints, we enrich it with details.

Ego-Exo Learning. Research on egocentric–exocentric video understanding primarily studies representation alignment and cross-view adaptation. For instance, prior efforts [27, 36, 38, 50] in action recognition have proposed self-supervised methods based on contrastive objectives for view-invariant representation learning. Another line of studies [21, 26, 56] has explored distilling knowledge from one view to another. However, these efforts rarely examine whether the temporal reasoning of Video-LLMs remains consistent when identical events are observed from different viewpoints. In this work, we investigate this and propose a reinforcement learning approach to improve temporal reasoning consistency across viewpoints, inspired by the recent success of RL learning in video reasoning [8, 24, 41, 58].

3 EgoExo-Con Dataset

3.1 Limitations in Existing Datasets

Unfortunately, the original temporal queries in datasets often do not meet our requirements. In Fig. 2, (a) queries in CharadesEgo are template-based, with categorical actions, and (b) queries in LEMMA rely on atomic actions and objects, both of which tend to miss details. More critically, viewpoint-induced ambiguities hinder reliable evaluation for cross-view consistency: key elements may be visible from one viewpoint but obscured from another due to varied focus and temporal alignment. For instance, the query "A person is smiling" in Fig. 2-(a) is not visible from the egocentric video. To this end, we carefully curate and re-annotate queries from existing datasets to ensure cross-view visibility and sufficient descriptive detail for reliable evaluation.

3.2 Data Collection

We source data from three datasets: CharadesEgo [36], LEMMA [18], and EgoExo-4D [12], as they cover diverse and general domains, whereas other benchmarks are often restricted to specific domains (*e.g.*, fitness [22], toy assembly [33], and surgery [30]). CharadesEgo and LEMMA feature daily human-object interactions,

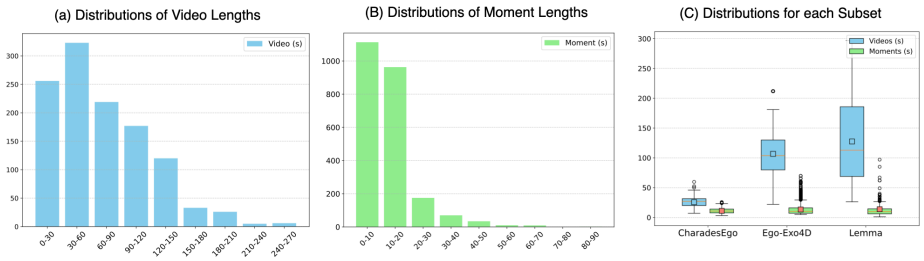

Fig. 3: Statistics of EgoExo-Con. We show the distribution of video lengths (a), moment lengths (b), and each subset statistics (c). Each subset exhibits distinct temporal scales across videos and moments.

while EgoExo-4D spans diverse skilled tasks, such as bike repairing and rock climbing. Among them, we collect synchronized video data annotated with query-timestamp pairs to support our focus on temporal understanding tasks. To ensure reliable evaluation, we carefully balance video diversity with model feasibility. For example, our preliminary experiments find that current models hardly perform effective temporal localization for long videos, making it infeasible to analyze consistency. Thus, we segment videos longer than five minutes into multiple clips surrounding the ground-truth moments, with each video clip lasting for at least two minutes, thus maximally preserving content diversity while keeping the task manageable for existing Video-LLMs.

3.3 Data Filtering and Refinement

We reformulate collected queries in multiple stages. We convert the per-frame Human-Object Interaction (HOI) labels (*e.g.*, put + cup, fridge) in LEMMA into natural-language queries. Since HOI labels are grounded to short 1–2 second intervals, we aggregate consecutive annotations into longer spans, extract salient verbs and nouns, and verbalize them into natural-language queries using simple rules for targets and prepositions. Next, we utilize a strong model (*i.e.*, GPT-4o [1]) to enrich the original queries across all datasets. Specifically, given sampled frames from the target moments, the model verifies whether a query contains elements that cannot be reliably inferred from one or both viewpoints and produces refined alternatives. Additionally, the model generates a *misaligned query* containing irrelevant content, which serves as a negative sample for temporal verification, thus balancing answers for "yes" and "No" in the verification task. The full prompt is shown in Appendix.

3.4 Human Verification

Finally, we perform human validation for all samples. Four human evaluators review the generated queries alongside associated synchronized video pairs. They confirm whether refined queries are accurately grounded in both viewpoints, and

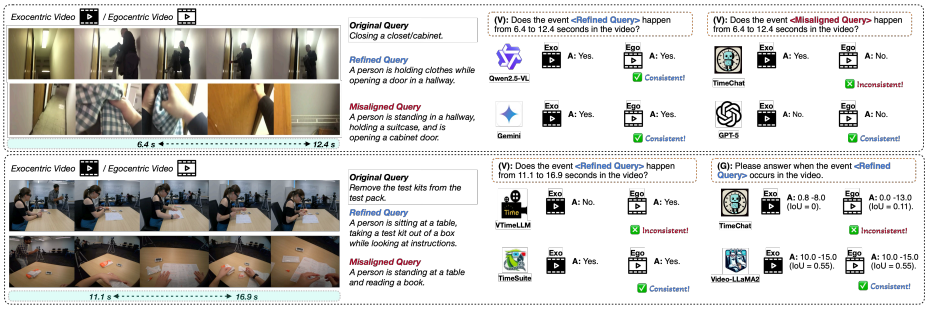

Fig. 4: Examples of test data and the corresponding model responses. We create refined and misaligned queries from each original query, use them for temporal verification (V) and grounding (G), and assess cross-view answer consistency.

misaligned queries intentionally conflict with the visual content. Specifically, we guide annotators to evaluate each ego-exo pair with three criteria: (1) temporal synchronization between the views, (2) visual quality (pairs with blur, poor lighting, resolution, etc. were rejected), and (3) action discriminability (the queried action is visible in both views). Specifically, for (3), we asked annotators to "verify whether all key entities and actions described in the query could be reliably inferred from both viewpoints." Queries passing this validation are retained, while ambiguous or low-quality samples are refined further or discarded (*e.g.*, if the video itself is too noisy). Uncertain cases are cross-checked with the authors to ensure reliability. Fig. 4 presents examples of refined and misaligned queries generated from the original query and its associated video.

3.5 Dataset Statistics

Eventually, we obtain 1,148 synchronized videos (426, 558, and 164 from CharadesEgo, EgoExo4D, and LEMMA, respectively) and 2,269 tightly aligned queries with timestamps. As shown in Fig. 3, each subset introduces distinct challenges specific to its domain and contributes a diverse range of video and moment lengths. Across the entire dataset, average duration and moment length are 86.4 seconds and 12.8 seconds, respectively. The average query length is 13.0 tokens, while the corresponding misaligned queries contain 15.9 tokens on average.

EgoExo-Con is a carefully curated video dataset that enables the evaluation of fine-grained, view-invariant temporal understanding, supported by extensive human verification and refinement. Among benchmarks that satisfy such strict criteria and involve human effort, EgoExo-Con remains competitive in scale. For instance, Video-MME [9] is a representative VideoQA benchmark consisting of 900 videos and 2,700 queries; however, those are not temporally annotated. EgoExo-Learn [17] contains 747 videos with human demonstrations capturing processes across viewpoints, yet the videos are not synchronized.

4 Evaluation

4.1 Models

Baseline. We evaluate a series of open-source models and categorize them as general-purpose or time-aware models, depending on whether they are designed for generic question-answering or specifically tuned to provide answers with corresponding video timestamps. Four general-purpose models: VideoChat2 [20], Qwen2.5-VL [2], Video-LLaMA2 [4], and Video-LLaMA3 [55], and four time-aware models: VTimeLLM [16], TimeChat [32], TimeSuite [54], and TimeChat-VT [19], are included. We provide details of each model in Appendix. Additionally, we include two powerful closed-source models: GPT-5 [37] and Gemini-2.5 Flash [5]. We also benchmark human performance as a reference. We invite four evaluators and present each viewpoint independently to avoid biased predictions, reporting the average of their scores. Note that closed-source models and human performance are reported on a $\simeq$30% subset of the full benchmark, which was uniformly sampled from each split to control evaluation costs. Additionally, we benchmark a random method that always returns "yes" or the entire video span for temporal verification and grounding, respectively.

Evaluation Metric. We use accuracy in percentage for temporal verification (V) and R@1, Intersection over Union (IoU)=0.5 for temporal grounding (G). For grounding, predictions are considered correct if their IoU with the ground-truth moment exceeds 0.5. A model is evaluated separately for each viewpoint: V-Ego and V-Exo measure binary accuracy for egocentric and exocentric videos, respectively. Similarly, G-Ego and G-Exo denote grounding performance. Consistency metrics, V-EgoExo and G-EgoExo, measure whether a model correctly verifies or grounds specific moments (*i.e.*, $0.5 \leqslant$ IoU) for both synchronized videos; consistent but wrong answers are not considered.

4.2 Performance Analysis on EgoExo-Con

Our analyses on the results of EgoExo-Con in Table 1 are as follows: **(1) Single view vs. Cross view.** While all models show a modest performance gap between individual ego and exo videos, they struggle with cross-view consistency in both tasks. Open-source models in particular achieve barely half of their single-view performance. This demonstrates that the relatively stable performances across viewpoints are largely due to view-specific bias cues but not robust cross-view reasoning. **(2) Time-aware vs. General-purpose.** Time-aware models generally lead on grounding. TimeSuite attains the strongest grounding consistency, and TimeChat-VT is competitive while also giving the best verification consistency. This advantage likely comes from better instruction tuning. However, VTimeLLM underperforms most general-purpose models, showcasing that time-aware models are not necessarily superior to general-purpose ones in temporal reasoning. **(3) Closed-sourced vs Human.** Closed-source models generally outperform open-source ones, reflecting their stronger capabilities. Yet a substantial gap

Table 1: Existing model performance on EgoExo-Con. F: input frames. Ego: include ego data for training.

Methods	# F	Ego	EgoExo-Con					
			V-Exo	V-Ego	V-ExoEgo	G-Exo	G-Ego	G-ExoEgo
Human	-	-	92.1	91.3	89.4	72.4	73.0	67.3
Closed-source								
GPT-5 [37]	32	-	60.5	61.3	52.5	34.5	32.8	20.1
Gemini-2.5 Flash [5]	1 fps	-	70.4	70.1	52.3	42.0	45.9	20.8
Random	-	-	50.0	50.0	50.0	12.5	12.5	12.5
General-purpose								
VideoChat2 [20]	16	✓	46.0	45.1	23.4	5.6	5.3	4.0
Qwen2.5-VL [2]	1 fps	✗	54.3	<u>56.3</u>	33.0	14.2	11.4	6.9
Video-LLaMA2 [4]	8	✓	53.3	52.1	27.9	12.0	11.5	7.5
Video-LLaMA3 [55]	1 fps	✓	<u>56.7</u>	54.6	<u>36.6</u>	27.7	**28.0**	16.2
Time-aware								
VTimeLLM [16]	100	✗	48.9	48.5	23.5	12.6	11.1	6.5
TimeChat [32]	96	✗	48.9	48.4	25.1	21.3	20.5	12.8
TimeSuite [54]	128	✓	47.4	48.5	25.6	**28.2**	<u>27.3</u>	**18.7**
TimeChat-VT [19]	96	✗	**62.1**	**61.4**	**42.1**	<u>27.8</u>	26.2	<u>16.3</u>

(37% 47%) in cross-view consistency remains compared to humans, with scores approaching random, underscoring the challenge of EgoExo-Con and the room for improvement. **(4) Training with ego data is not sufficient.** Models including egocentric videos for training do not consistently yield higher consistency than others trained on exocentric videos alone, showing that simply mixing ego and exo data does not benefit consistency. **(5) Temporal reasoning outweighs increasing frames for better results.** Video-LLaMA2 (8 frames) outperforms VideoChat2 (16 frames) across all metrics. TimeChat-VT (96 frames) also outperforms several models that use more/less context, suggesting reasoning and temporal modeling matter more than sheer frame count.

Furthermore, we analyze model behaviors across subsets by plotting the ego–exo performance gap in Fig. 5. Patterns are consistent across grounding and verification: in CharadesEgo, most models perform better on exocentric views (blue), whereas in LEMMA, they tend to favor egocentric views (red); EgoExo-4D shows mixed but generally smaller gaps. These trends likely reflect domain characteristics. As shown in Fig. 4, when a person performs in a fixed position, egocentric videos often provide favorable cues such as clearer hand–object interactions. Conversely, when a person moves around or changes location, egocentric views can become more challenging due to rapid scene shifts, while exocentric views offer greater stability. Such trends are particularly pronounced in CharadesEgo. Detailed results across subsets are provided in Appendix. Although the relative effectiveness of each viewpoint varies by domain, they do not significantly

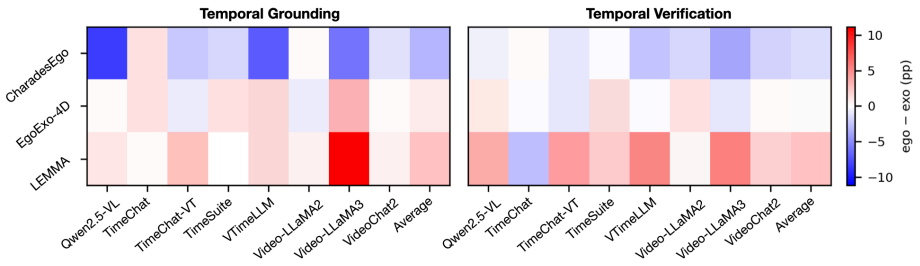

Fig. 5: Heatmaps of the performance gap. All values are reported in percentage points. Red and blue indicate higher performances on *ego* and *exo* perspectives, respectively. i.e., *a blue cell indicates that the corresponding model performs better on exocentric views than on egocentric ones, while red indicates better performance on egocentric views compared to exocentric ones.*

Table 2: Fine-tuned performance. The left and right tables report model performance for temporal verification (V) and temporal grounding (G). The first row in each model reports zero-shot performance, while the subsequent rows present results after fine-tuning on either (+Ego, +Exo) or both viewpoints (+ EgoExo). Notably, models trained on both viewpoints are not always the best-performing, and then they barely outperform those trained on a single viewpoint.

Methods	CharadesEgo			EgoExo-4D		
	V-Exo	V-Ego	V-ExoEgo	V-Exo	V-Ego	V-ExoEgo
VideoChat2	46.3	44.4	22.2	41.4	41.5	19.8
+ Ego	56.2	**59.2**	**36.4**	46.5	49.2	29.7
+ Exo	**56.6**	57.5	35.5	46.1	47.3	29.4
+ EgoExo	56.4	57.1	34.7	**48.3**	**50.1**	**30.1**
Video-LLaMA2	54.0	52.4	28.2	51.5	52.8	28.1
+ Ego	57.0	**58.2**	**31.7**	**60.2**	**60.3**	38.8
+ Exo	57.6	56.1	31.4	59.8	60.1	**39.2**
+ EgoExo	**58.5**	57.3	31.0	57.5	59.6	39.1

Methods	CharadesEgo			EgoExo-4D		
	G-Exo	G-Ego	G-ExoEgo	G-Exo	G-Ego	G-ExoEgo
TimeChat	44.9	46.1	30.1	4.9	6.3	3.3
+ Ego	**62.0**	**62.1**	**48.3**	9.7	**13.1**	**4.9**
+ Exo	58.8	60.0	47.1	10.4	10.7	4.2
+ EgoExo	60.3	61.8	46.5	**10.6**	11.5	4.3
TimeSuite	63.4	56.2	44.8	5.8	8.3	2.3
+ Ego	61.0	61.5	54.0	6.7	6.2	4.8
+ Exo	**74.6**	**68.7**	**59.5**	**10.5**	**9.9**	**5.7**
+ EgoExo	67.8	61.1	51.3	9.6	8.9	5.3

affect overall consistency. Overall, across models, consistency scores lag far behind single-view metrics, underscoring that achieving robust, view-invariant temporal understanding remains an open challenge.

4.3 Supervised Performance on EgoExo-Con

In this section, we study whether supervised fine-tuning (SFT) with synchronized ego-exo video data improves performance. We first collect 3.6k from CharadesEgo and 2.3k videos from EgoExo-4D training sets. We exclude LEMMA due to its limited size (*i.e.*, 243 videos for training). Then we finetune two general-purpose models: VideoChat2 and Video-LLaMA2 for temporal verification, and two time-aware models: TimeChat and TimeSuite for temporal grounding. All models apply LoRA [15] fine-tuning with their official configurations. More implementation details are presented in Appendix.

Table 2 reports results of three different settings: training on either viewpoint (Ego, Exo) or both viewpoints together (ExoEgo). All of them consistently improve over zero-shot baselines. Interestingly, despite utilizing twice the data, training

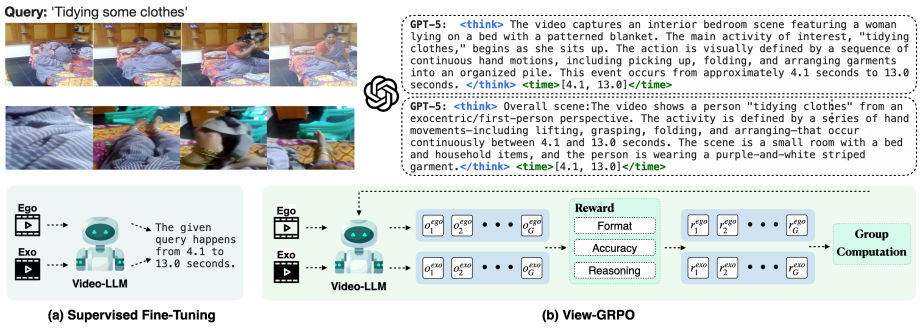

Fig. 6: Overview of our approach. (a) In supervised fine-tuning, the model is trained to directly predict the same query answers (*e.g.*, video moments) for synchronized video pairs. (b) View-GRPO trains a model to provide viewpoint-specific reasoning chains, which are generated by GPT-5 (top).

on both viewpoints (ExoEgo) yields only marginal gains and often underperforms the models trained on a single view. Specifically, in CharadesEgo, TimeSuite shows a notable 8.1% gap in consistency between training on both viewpoints and training only on exocentric videos. Furthermore, the improvements remain limited even when unfreezing the visual encoder (refer to Appendix), suggesting that simply increasing model capacity is insufficient and highlighting the fundamental challenge of view-invariant temporal understanding.

The above findings are consistent with observations in cross-view learning [22, 36], where naively blending perspectives does not always bring improvement. Without explicit alignment, conflicting priors across tasks or domains undermine temporal signals and consistency rather than improvement.

5 Method

We propose a reinforcement learning (RL) framework that guides models toward developing viewpoint-specific reasoning while encouraging shared consistency. Rather than simply enforcing identical outputs, our approach encourages the model to internalize the robust reasoning traces with structured guidance, while simultaneously preserving agreement between the outputs for each viewpoint. We build on Group Relative Policy Optimization (GRPO), which is particularly well-suited as it leverages relative rewards instead of absolute scores.

5.1 Background: Group Relative Policy Optimization (GRPO)

GRPO [34] is a reinforcement learning algorithm designed to refine large language model outputs by leveraging relative ranking among multiple candidate responses. Instead of treating each response independently with absolute rewards, GRPO evaluates a set of responses produced for the same prompt, assigning rewards

in relation to the group. This group-wise normalization helps reduce reward variance and makes optimization more stable compared to approaches relying solely on absolute scores or pairwise comparisons.

Given a prompt p, the model generates G candidate responses $o = \{o_1, \ldots, o_G\}$. Each response receives a reward value $r(o_i)$. GRPO then standardizes these scores within the group and optimizes a weighted objective:

$$R(o) = \sum_{i=1}^{G} \frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} \cdot \frac{r(o_i) - \text{mean}(\{r(o_i)\}_{i=1}^{G})}{\text{std}(\{r(o_i)\}_{i=1}^{G})}, \tag{1}$$

where $\pi_\theta(o)$ denotes the current policy and $\pi_{\theta_{\text{old}}}(o)$ is the previous policy. To prevent divergence from the base model, KL-divergence regularization is added:

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(p)} \Big[R(o) - \beta D_{\text{KL}}\big(\pi_\theta \parallel \pi_{\text{ref}}\big) \Big], \tag{2}$$

where π_{ref} is the base model and β controls the strength of the regularization. Please refer to the original paper for more details.

5.2 Learning Temporal Reasoning across Viewpoints

To adapt GRPO for cross-view reasoning, we first curate training data, including temporal reasoning chains for both egocentric and exocentric views. Fig. 6 (Top) illustrates the generation of video reasoning data. We prompt GPT-5 to produce step-by-step reasoning chains for each video in order to solve the given task, ensuring viewpoint-specific reasoning while aligning the final answers. We discard samples lacking valid reasoning in either view, as these typically indicate ambiguous or low-quality data. Specifically, we exclude cases where the model explicitly states its failure in the answer, or where the predicted moment has a temporal IoU (tIoU) below 0.7 with the ground-truth. After filtering, we retain 3.3k videos with 30k reasoning instances, which we name as View30K.

We then design reward functions comprising three major components:

1. Format Reward. For structured reasoning and easy answer extraction, responses must follow the template: `<think>...</think><answer>...</answer>`. Formally:

$$r_{\text{format}}(o) = \begin{cases} 1, & \text{if } o \text{ follows the required format,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

2. Accuracy Reward. We unify task-specific accuracy into r_{acc}. For temporal grounding, the reward is the temporal Intersection-over-Union (tIoU) between ground truth $[t_s, t_e]$ and prediction $[t'_s, t'_e]$. For verification, it is binary correctness:

$$r_{\text{accuracy}}(o) = \begin{cases} \frac{|[t_s,t_e] \cap [t'_s,t'_e]|}{|[t_s,t_e] \cup [t'_s,t'_e]|}, & \text{if } \textit{grounding}, \\ \mathbb{1}[o = o^*], & \text{if } \textit{verification}. \end{cases} \tag{4}$$

3. Reasoning Reward. We design a unified reasoning reward that encourages both semantic alignment with reference reasoning o^* and cross-view structural consistency with its counterpart reasoning $\tilde{o}$. Specifically, for a given video instance with two viewpoints, we denote by o the reasoning generated from one viewpoint, and by $\tilde{o}$ the reasoning generated from its counterpart viewpoint. Without loss of generality, if $o = o^{ego}$, then $\tilde{o} = o^{exo}$, and vice versa. First, given a generated reasoning o and a reference reasoning o^*, we employ an LLM judge to compute a semantic reward:

$$r_{\text{sem}}(o) = \text{Judge}(o, o^*) \in [0, 1]. \tag{5}$$

This term encourages the generated reasoning to remain faithful to the reference explanation. Consequently, we extract action and object sets from o and $\tilde{o}$ using spaCy, denoted as $\mathcal{A}$ and $\mathcal{O}$, respectively. We then measure their structural consistency via the Jaccard overlap:

$$r_{\text{struct}}(o) = \frac{1}{2}\big(\text{Jac}(\mathcal{A}(o), \mathcal{A}(\tilde{o})) + \text{Jac}(\mathcal{O}(o), \mathcal{O}(\tilde{o}))\big). \tag{6}$$

The final reasoning reward integrates both components:

$$r_{\text{reasoning}}(o) = \lambda * r_{\text{sem}}(o) + (1 - \lambda) * r_{\text{struct}}(o), \tag{7}$$

where $\lambda \in [0, 1]$ balances reference alignment and cross-view consistency.

Overall, the total reward integrates all the components to balance correctness and reasoning quality:

$$r(o) = r_{\text{accuracy}}(o) + r_{\text{format}}(o) + r_{\text{reasoning}}(o), \tag{8}$$

thus enabling models to learn correct and consistent cross-view temporal reasoning. We name the overall approach View-GRPO, as shown in Fig. 6-(b) (Bottom).

5.3 Implementations

We use Qwen2.5-VL [2] and InternVL3.5 [39] as a backbone. During training, all experiments set the same frame sampling rate (*i.e.*, 2 FPS) and freeze the visual encoder and update only the parameters of the LLM. The training involves 8 $\times$ A100 GPUs, and further implementation details are in Appendix.

5.4 Analyses on View-GRPO

Table 3 shows the performance of our method, View-GRPO, on EgoExo-Con compared to standard SFT and GRPO implementations. View-GRPO demonstrates superior performance, achieving significant improvements over baselines beyond those obtained with SFT and GRPO. Notably, the most significant improvements are often on cross-view consistency (V-ExoEgo and G-ExoEgo), although it also benefits individual views. We conjecture that the reasoning reward plays a central role, as it delivers noticeably higher consistency compared to naive GRPO. Additionally, View-GRPO remains effective with the different backbone,

Table 3: Performance of View-GRPO on EgoExo-Con.

Methods	EgoExo-Con					
	V-Exo	V-Ego	V-ExoEgo	G-Exo	G-Ego	G-ExoEgo
Qwen2.5-VL-3B	51.0	52.5	28.1	10.1	9.9	7.9
+ SFT	52.7	51.2	30.6	16.3	16.6	12.9
+ GRPO	52.5	52.9	30.3	15.4	16.3	10.3
+ **View-GRPO**	**54.6** ↑3.6	**54.4** ↑1.9	**34.2** ↑6.1	**18.9** ↑8.8	**17.9** ↑8.0	**14.9** ↑7.0
Qwen2.5-VL-7B	54.3	56.3	33.0	14.2	11.4	6.9
+ SFT	57.6	58.0	41.4	18.3	17.8	14.9
+ GRPO	55.2	57.6	39.8	18.6	16.1	14.3
+ **View-GRPO**	**58.6** ↑4.3	**58.2** ↑1.9	**45.1** ↑12.1	**22.0** ↑7.8	**21.6** ↑10.2	**18.7** ↑11.8
InternVL3.5-8B	64.4	64.7	50.7	12.8	6.7	3.0
+ SFT	70.3	71.1	57.5	18.1	15.1	7.8
+ GRPO	71.2	71.8	58.9	18.0	14.7	8.4
+ **View-GRPO**	**73.1** ↑8.7	**74.4** ↑9.7	**62.4** ↑11.7	**20.5** ↑7.7	**16.8** ↑10.1	**10.6** ↑7.6

Table 4: Ablation on reasoning reward.

r_{sem}	r_{struct}	EgoExo-Con					
		V-Exo	V-Ego	V-ExoEgo	G-Exo	G-Ego	G-ExoEgo
		55.2	57.6	39.8	18.6	16.1	14.3
✓		58.3	58.1	44.7	21.5	21.0	18.3
✓	✓	**58.6**	**58.2**	**45.1**	**22.0**	**21.6**	**18.7**

Table 5: Performance on Video-MME and TVGBench.

Methods	Video-MME	TVGBench		
	w/o subs	R1@0.3	R1@0.5	R1@0.7
TimeChat	30.2	22.4	11.9	5.3
TimeSuite	46.3	31.1	18.0	8.9
Qwen2.5-VL	61.1	28.1	19.5	10.5
View-GRPO	**69.7**	**42.0**	**25.0**	**13.9**

InternVL3.5. By encouraging models to produce faithful, step-by-step temporal explanations tailored to each viewpoint while converging toward consistent temporal conclusions, the model reduces reliance on view-specific biases and instead learns shared temporal abstractions.

We further analyze our reasoning reward design in depth. While LLMs are commonly used as judges [48, 60], the role in optimizing models in View-GRPO remains underexplored despite its effectiveness, as they often introduce potential bias and uncertainty in video evaluation [6, 25]. To provide insights into this, we employ judge models of different scales for temporal grounding and analyze their impact on optimization. In Fig. 7-(a), format and accuracy rewards remain relatively stable across scales. However, Qwen2.5-0.5B produces overly high reasoning rewards from the very first training steps in Fig. 7-(b), raising concerns about calibration

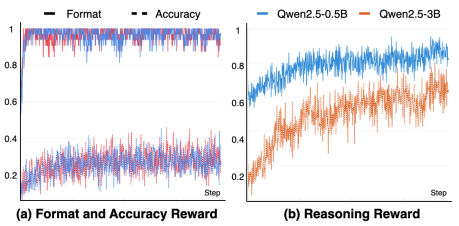

Fig. 7: Reward of different LLM judges. Qwen2.5-0.5B raises calibration concerns due to its overly high reasoning rewards from early steps.

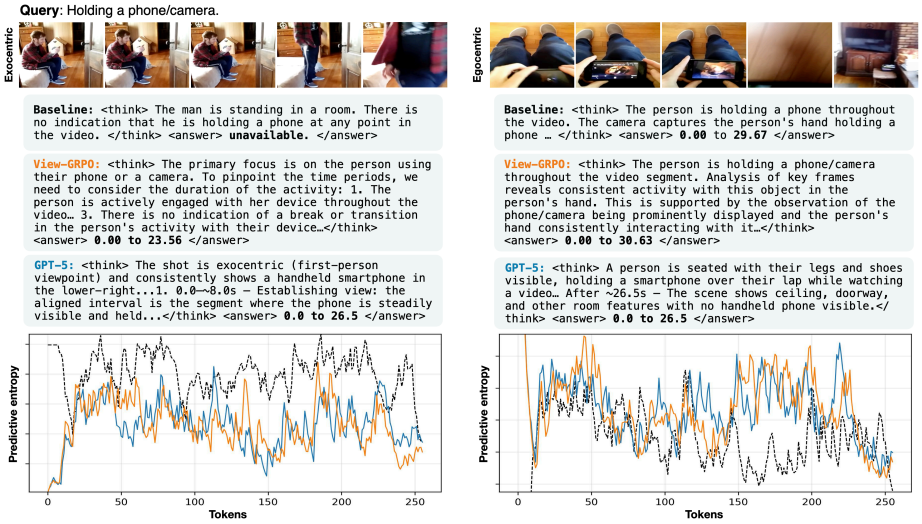

Fig. 8: Visualization of View-GRPO. We visualize the model-generated reasoning and the reference reasoning for each viewpoint, along with their token-level entropy distributions over the generation length. The model produces step-by-step temporal reasoning with accurate grounding predictions, where the high overlap in token entropy patterns indicates close alignment with GPT-5's reasoning traces.

and reliability. We find that this behavior leads to a measurable consistency degradation (*i.e.*, –3% in G-EgoExo). Table 4 shows the effect of different components in $r_{\text{reasoning}}$. The semantic reward r_{sem} consistently improves performance over the baseline, and incorporating the structural consistency r_{struct} leads to further gains. By r_{sem} for conceptual alignment with reference reasoning and r_{struct} for cross-view alignment, the model achieves more consistent temporal reasoning across viewpoints. We further investigate performance across varying reasoning lengths to assess the potential impact of hallucinations in reasoning in Appendix. Overall, such results underscore the critical influence of LLM-judges and reward calibration on model optimization.

Additionally, we evaluate our method on additional benchmarks: Video-MME [9] and TVGBench [41]. We select Video-MME as a representative benchmark for VideoQA, while TVGBench encompasses diverse temporal grounding datasets with highly varied query formulations. Table 5 demonstrates that our method possesses strong generalization capabilities across diverse video benchmarks. We attribute this robustness to strengthened temporal reasoning and improved cross-view consistency through View-GRPO. We discuss potential directions for improvement from several perspectives: data, architecture, and learning, in Appendix. Overall, our method demonstrates its effectiveness and potential for consistent and view-invariant temporal understanding.

5.5 Qualitative Analysis

Fig. 8 visualizes a qualitative comparison of reasoning traces and token-level predictive entropy across viewpoints. We compare the baseline model (*i.e.*, Qwen2.5-VL-7B), GPT-5 for reference, and our method under synchronized video pairs. Although the baseline provides an accurate answer for the egocentric video, it fails to identify the target moment for the exocentric video, showing unreliable reasoning behavior and inconsistency. In contrast, View-GRPO provides the step-by-step temporal reasoning traces with accurate grounding and closely follows the reference reasoning, exhibiting similar peaks and transitions. This suggests that View-GRPO internalizes robust reasoning patterns while aligning its consistent answers across viewpoints.

6 Conclusion

In this work, we introduced EgoExo-Con that comprises synchronized egocentric and exocentric videos paired with human-refined queries and evaluated whether models perform consistent temporal understanding across viewpoints. EgoExo-Con revealed that Video-LLMs struggle with cross-view consistency, lagging behind single-viewpoint performance, and that naively training on both viewpoints does not reliably help. To address this, we proposed View-GRPO that encourages viewpoint-specific temporal reasoning while promoting cross-view alignment, demonstrating its effectiveness over alternative training strategies. We hope EgoExo-Con and View-GRPO pave the way toward truly robust and view-invariant temporal understanding.

Acknowledge

This research is supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (MOE-T2EP20125-0037). This work was partly supported by the IITP (RS-2021-II212068-AIHub/10%, RS-2021-II211343-GSAI/15%, RS-2022-II220951-LBA/15%, RS-2022-II220953-PICA/20%), NRF (RS-2024-00353991-SPARC/20%, RS-2023-00274280-HEI/10%), and KEIT (RS-2024-00423940/10%) grant funded by the Korean government.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., Liu, Y.: Egothink: Evaluating first-person perspective thinking capability of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14291–14302 (2024)
4. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
6. Cores, D., Dorkenwald, M., Mucientes, M., Snoek, C.G., Asano, Y.M.: Lost in time: A new temporal benchmark for videollms. arXiv preprint arXiv:2410.07752 (2024)
7. Di, S., Xie, W.: Grounded question-answering in long egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12934–12943 (2024)
8. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
9. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075 (2024)
10. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
11. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
12. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
13. Guo, Y., Liu, J., Li, M., Tang, X., Chen, X., Zhao, B.: Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. arXiv preprint arXiv:2405.13382 (2024)
14. He, Y., Huang, Y., Chen, G., Pei, B., Xu, J., Lu, T., Pang, J.: Egoexobench: A benchmark for first-and third-person view video understanding in mllms. arXiv preprint arXiv:2507.18342 (2025)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)

16. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)
17. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Dong, L., Wang, Y., Wang, L., et al.: Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22072–22086 (2024)
18. Jia, B., Chen, Y., Huang, S., Zhu, Y., Zhu, S.C.: Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In: European Conference on Computer Vision. pp. 767–786. Springer (2020)
19. Jung, M., Xiao, J., Zhang, B.T., Yao, A.: On the consistency of video large language models in temporal comprehension. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13713–13722 (2025)
20. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
21. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: Ego-exo: Transferring visual representations from third-person to first-person videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6943–6953 (2021)
22. Li, Y.M., Huang, W.J., Wang, A.L., Zeng, L.A., Meng, J.K., Zheng, W.S.: Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In: European Conference on Computer Vision. pp. 363–382. Springer (2024)
23. Li, Z., Di, S., Zhai, Z., Huang, W., Wang, Y., Xie, W.: Universal video temporal grounding with generative multi-modal large language models. arXiv preprint arXiv:2506.18883 (2025)
24. Liao, Z., Xie, Q., Zhang, Y., Kong, Z., Lu, H., Yang, Z., Deng, Z.: Improved visual-spatial reasoning via r1-zero-like training. arXiv preprint arXiv:2504.00883 (2025)
25. Liu, M., Zhang, W.: Is your video language model a reliable judge? arXiv preprint arXiv:2503.05977 (2025)
26. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In: European Conference on Computer Vision. pp. 407–425. Springer (2024)
27. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Viewpoint rosetta stone: Unlocking unpaired ego-exo videos for view-invariant representation learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15802–15812 (2025)
28. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
29. Meinardus, B., Rodriguez, H., Batra, A., Rohrbach, A., Rohrbach, M.: Chrono: A simple blueprint for representing time in mllms. arXiv preprint arXiv:2406.18113 (2024)
30. Özsoy, E., Mamur, A., Tristram, F., Pellegrini, C., Wysocki, M., Busam, B., Navab, N.: Egoexor: An ego-exo-centric operating room dataset for surgical activity understanding. arXiv preprint arXiv:2505.24287 (2025)
31. Qian, L., Li, J., Wu, Y., Ye, Y., Fei, H., Chua, T.S., Zhuang, Y., Tang, S.: Momentor: Advancing video large language model with fine-grained temporal reasoning. arXiv preprint arXiv:2402.11435 (2024)

32. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323 (2024)
33. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)
34. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
35. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024)
36. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7396–7404 (2018)
37. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
38. Wang, Q., Zhao, L., Yuan, L., Liu, T., Peng, X.: Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3307–3317 (2023)
39. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
40. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. arXiv preprint arXiv:2403.10517 (2024)
41. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025)
42. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)
43. Wang, Y., Meng, X., Liang, J., Wang, Y., Liu, Q., Zhao, D.: Hawkeye: Training video-text llms for grounding text in videos. arXiv preprint arXiv:2403.10228 (2024)
44. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
45. Xiao, J., Huang, N., Qiu, H., Tao, Z., Yang, X., Hong, R., Wang, M., Yao, A.: Egoblind: Towards egocentric visual assistance for the blind people. arXiv preprint arXiv:2503.08221 (2025)
46. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
47. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13204–13214 (2024)

48. Xie, T., Zhao, S., Wu, C.H., Liu, Y., Luo, Q., Zhong, V., Yang, Y., Yu, T.: Text2reward: Reward shaping with language models for reinforcement learning. arXiv preprint arXiv:2309.11489 (2023)
49. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017)
50. Xue, Z.S., Grauman, K.: Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. Advances in Neural Information Processing Systems **36**, 53688–53710 (2023)
51. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
52. Ye, H., Zhang, H., Daxberger, E., Chen, L., Lin, Z., Li, Y., Zhang, B., You, H., Xu, D., Gan, Z., et al.: Mmego: Towards building egocentric multimodal llms for video qa. In: The Thirteenth International Conference on Learning Representations (2025)
53. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence (2019)
54. Zeng, X., Li, K., Wang, C., Li, X., Jiang, T., Yan, Z., Li, S., Shi, Y., Yue, Z., Wang, Y., et al.: Timesuite: Improving mllms for long video understanding via grounded tuning. arXiv preprint arXiv:2410.19702 (2024)
55. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
56. Zhang, H., Chu, Q., Liu, M., Wang, Y., Wen, B., Yang, F., Gao, T., Zhang, D., Wang, Y., Nie, L.: Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. arXiv preprint arXiv:2503.09143 (2025)
57. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024)
58. Zhang, X., Wen, S., Wu, W., Huang, L.: Tinyllava-video-r1: Towards smaller lmms for video reasoning. arXiv preprint arXiv:2504.09641 (2025)
59. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
60. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
61. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13691–13701 (2025)