EGOR1: INTERACTION-CENTRIC LATENT REASONING FOR EGOCENTRIC FUTURE INTERACTION FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Future interaction forecasting is a key capability for predictive multimodal intelligence, with broad applications in egocentric AI systems such as AR/VR assistance, human–robot collaboration, and embodied world models. However, despite the rapid progress of multimodal large language models (MLLMs) on general video understanding, their ability to predict imminent interaction events remains largely underexplored, and existing egocentric forecasting benchmarks are primarily designed for specialist models rather than MLLM-based reasoning. To bridge this gap, we introduce EGOR1-BENCH1K, a new benchmark for evaluating future interaction forecasting in MLLMs. Given historical egocentric video frames and a multiple-choice query, the model must select the most likely next interaction event and localize the interacted object in the last observed frame. On top of this benchmark, we propose **EGOR1**, an interaction-centric latent reasoning framework that transforms latent reasoning from generic future summarization into progressive interaction-state grounding. Specifically, EGOR1 first distills semantically complete future interaction state through a Multimodal State Teacher, and then performs object-centric latent reasoning to progressively ground the predicted interaction onto the target object. We further apply a GRPO-based reinforcement learning stage to improve robustness and generalization. Extensive experiments show that future interaction forecasting remains challenging for current MLLMs, while EGOR1 consistently improves both future event prediction and object grounding over strong MLLM and latent reasoning baselines. These results suggest that structured latent reasoning is a promising route toward predictive egocentric intelligence in multimodal foundation models.

1 INTRODUCTION

Future interaction forecasting requires inferring what interaction will happen next from historical observations, extending video understanding beyond recognizing the current moment. This capability supports predictive egocentric AI: proactive AR/VR assistants can anticipate actions and target objects for low-latency feedback and interaction prefetching; human–robot collaboration benefits from timely anticipation for safe coordination; and embodied world models must forecast future interaction outcomes. It is therefore a key step from passive perception toward predictive and interaction-aware intelligence.

Recent multimodal large language models (MLLMs) perform strongly on image and video understanding. Benchmarks such as VIDEO-MME and MLVU evaluate video reasoning, long-video comprehension, and temporal question answering (Fu et al., 2025; Zhou et al., 2025), but do not explicitly require forecasting the next interaction from historical egocentric evidence and grounding it onto the target object. This predictive ability remains largely underexplored.

Classical egocentric forecasting benchmarks provide an important foundation, but they only partially cover the capability studied in this work. EPIC-KITCHENS establishes action anticipation as future verb–noun prediction from preceding observations, while EGO4D extends forecasting to short-term object interaction anticipation and long-term action anticipation, including future interacted objects, verbs, and associated temporal cues (Grauman et al., 2022; Damen et al., 2018). These

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

benchmarks are highly valuable, yet they are mainly designed for specialist anticipation models and structured evaluation pipelines, rather than MLLM-native reasoning interfaces. This leaves an important gap between classical egocentric forecasting and modern multimodal foundation models.

To bridge this gap, we introduce EGOR1-BENCH1K. Given historical egocentric video frames and a multiple-choice query, the model must select the most likely next interaction event and localize the interacted object in the last observed frame. This jointly evaluates semantic forecasting and spatial grounding through an MLLM-compatible interface that supports structured latent reasoning and post-training.

Strong MLLM baselines show that direct prompting and supervised fine-tuning remain challenged by cluttered scenes, plausible distractors, and multiple salient candidate objects. These failures suggest a missing task-aligned intermediate reasoning process: the model must infer a coherent *future interaction state* from history and then determine *which object* that state will act upon.

This motivates our method, EGOR1, an interaction-centric latent reasoning framework for future interaction forecasting. The core idea of EGOR1 is that latent reasoning for this task should not merely summarize future visual evidence as a few generic latent tokens. Instead, it should first distill a semantically complete future interaction state and then progressively ground that state onto the target object. Concretely, EGOR1 introduces two stages of latent reasoning: *future interaction state distillation*, which constructs state-level supervision for the <|context_latent|> tokens using a multimodal state teacher built from future interaction frames and paired interaction text; and *object-centric latent reasoning*, which further generates <|object_latent|> tokens to contract the global interaction state onto the most likely target object, together with contrastive grounding to distinguish the interacted object from background and distractors. On top of this structured latent reasoning process, we further apply Group Relative Policy Optimization (GRPO) as a final reinforcement learning stage, using both option correctness and box IoU as rewards to improve robustness and generalization.

Our experiments show that EGOR1 substantially outperforms both general MLLM baselines and generic latent reasoning methods on EGOR1-BENCH1K. More importantly, the results consistently support the main thesis of this paper: future interaction forecasting is not well served by generic future summarization, but instead benefits from a reasoning process that moves from future interaction state to target grounding.

In summary, our contributions are as follows:

- We introduce EGOR1-BENCH1K, a new benchmark for evaluating future interaction forecasting in MLLMs, where models must jointly predict the next interaction event and localize the corresponding target object.
- We identify the key limitations of standard MLLM prompting, supervised fine-tuning, and generic latent reasoning for this task, and show that the core challenge lies in modeling future interaction state and resolving target ambiguity.
- We propose EGOR1, an interaction-centric latent reasoning framework that combines future interaction state distillation, object-centric latent reasoning, and GRPO-based post-training, leading to strong improvements in both future event prediction and object grounding.

2 RELATED WORK

2.1 EGOCENTRIC ACTION ANTICIPATION AND FORECASTING

Egocentric future prediction spans action anticipation, object interaction anticipation, and short-term forecasting. EPIC-KITCHENS evaluates future $(\textit{verb}, \textit{noun})$ prediction, including top-k performance, long-tail distributions, and cross-user generalization (Damen et al., 2018; 2022). EGO4D adds a richer forecasting suite, including future active objects, interaction verbs, time-to-contact, and localization, while EGO-EXO4D broadens first- and third-person skilled-activity understanding (Grauman et al., 2022; 2024).

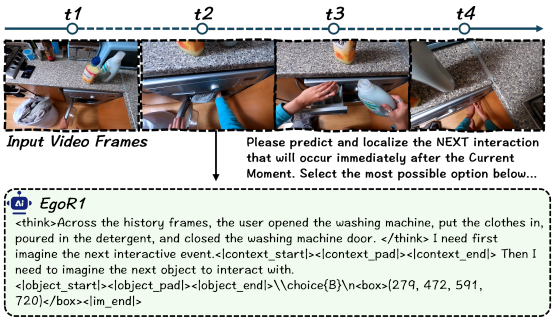


Figure 1: Egocentric Future Interaction Forecasting.

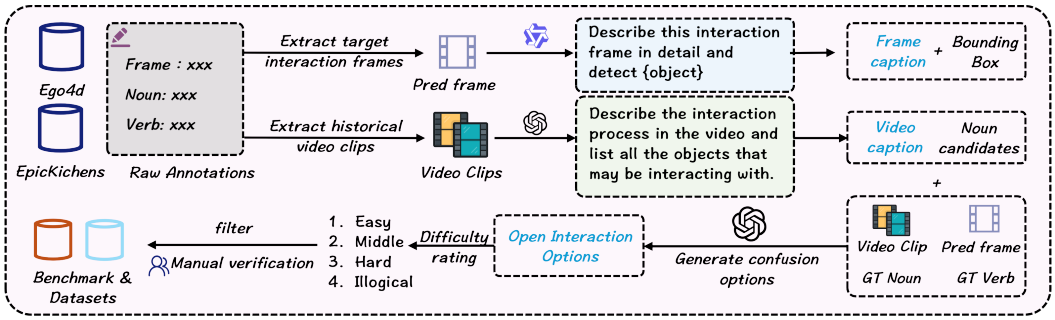


Figure 2: Data Construction Pipeline.

At the modeling level, however, most prior work has been developed for specialist anticipation architectures rather than general-purpose multimodal foundation models. Representative methods focus on recurrent anticipation, temporal transformers and memory, hand/object dynamics, next-active-object prediction, or interaction regions through dedicated architectures and task-specific heads (Furnari & Farinella, 2019; Girdhar & Grauman, 2021; Wu et al., 2022; Ragusa et al., 2023; Thakur et al., 2024; Pasca et al., 2024; Pei et al., 2025; Mur-Labadia et al., 2024; Roy et al., 2024). A complementary line predicts action targets and hand trajectories or uses auxiliary hand–object supervision (Li et al., 2022; Bao et al., 2023; Zhang et al., 2023; 2024). A common feature of this line of work is that forecasting is treated as a narrowly structured prediction problem, often optimized separately for verbs, nouns, boxes, or temporal offsets. Such formulations have significantly advanced egocentric forecasting, but they are not naturally aligned with the generative interfaces, multimodal prompting styles, and open-ended reasoning mechanisms of modern MLLMs. In contrast, our work targets future interaction forecasting as an MLLM-native reasoning problem, where models must jointly infer the next interaction event and ground it onto the corresponding object.

2.2 Video Understanding Benchmarks for MLLMs

General video-capable MLLMs include Video-ChatGPT and LLaVA-OneVision (Maaz et al., 2024; Li et al., 2024a). Video-MME evaluates diverse video domains, temporal spans, and multimodal inputs (Fu et al., 2025), while EgoSchema, MVBench, TempCompass, LongVideoBench, and MLVU cover long-form, multi-task, temporal, and long-context understanding (Mangalam et al., 2023; Li et al., 2024b; Liu et al., 2024; Wu et al., 2024; Zhou et al., 2025).

These benchmarks emphasize comprehension, temporal question answering, event-level reasoning, and multimodal perception (Fu et al., 2025; Zhou et al., 2025), but do not explicitly evaluate the joint forecasting and grounding of the next egocentric interaction. Our benchmark targets this missing setting. Egocentric video–language pretraining methods such as LaViLa and EgoVLPv2 provide transferable representations (Zhao et al., 2023; Pramanick et al., 2023), but do not directly evaluate this joint future selection and localization setting.

2.3 Latent Visual Reasoning and Machine Mental Imagery

Our method is most closely related to recent progress in latent visual reasoning, which enables multimodal models to reason through continuous visual representations rather than purely textual chains of thought. A representative example is Mirage, which introduces the notion of *machine mental imagery* by interleaving latent visual tokens with text tokens during decoding (Yang et al., 2025). Instead of explicitly synthesizing images, Mirage supervises compact latent visual tokens distilled from future image embeddings, first grounding them to visual targets and then relaxing the supervision to better support downstream reasoning. This framework demonstrates that visual imagination can be performed efficiently in latent space, making it a natural starting point for tasks that require prediction beyond current observations.

More recent work further develops this direction. LVR formulates latent visual reasoning as autoregressive reasoning directly in the visual embedding space, jointly modeling latent visual generation and text generation, and further shows that reinforcement learning can stabilize and enhance such hybrid reasoning (Li et al., 2025). Monet studies the efficiency and supervision quality of latent visual reasoning in greater depth, proposing a three-stage distillation-based training pipeline together with latent-aware policy optimization (Wang et al., 2026). Taken together, these works validate latent visual reasoning as a promising scaffold for multimodal reasoning when explicit image generation is unnecessary or prohibitively expensive.

Despite their effectiveness, existing latent visual reasoning methods remain largely task-agnostic in how future latents are constructed and supervised. Their latent thoughts are typically optimized to capture compact future visual summaries or generic predictive evidence for broad multimodal reasoning tasks (Yang et al., 2025; Li et al., 2025; Wang et al., 2026). Future interaction forecasting, however, requires a more structured predictive process: the model must first infer a semantically coherent future interaction state and then progressively ground that state onto the target object. Our work builds on the latent reasoning paradigm, but departs from prior methods by explicitly tailoring future latents to interaction-state prediction and object-centric grounding.

3 Benchmark and Dataset

Building on egocentric anticipation datasets (Damen et al., 2018; Grauman et al., 2022) and complementing general MLLM video evaluation (Fu et al., 2025), EgoR1-Bench1K jointly evaluates next-interaction prediction from historical observations and grounding onto the target object.

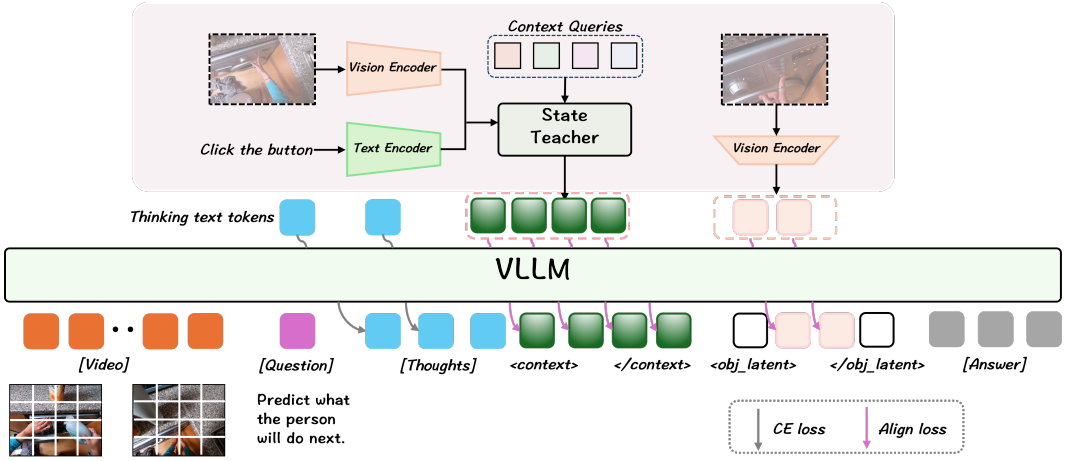


Figure 3: **Overview of EgoR1.**

3.1 Task Definition

Each benchmark sample consists of a sequence of historical egocentric video frames, a prediction query, and five candidate future interaction options. The model is asked to answer the query by

selecting the most likely future interaction option and localizing the interacted object in the last observed frame. Formally, the input is

$$(X, q, \mathcal{O}), \tag{1}$$

where $X = \{I_1, \ldots, I_T\}$ denotes the historical frames, q is the prediction question, and

$$\mathcal{O} = \{o_1, o_2, o_3, o_4, o_5\} \tag{2}$$

contains one correct option and four distractor options. The output consists of (1) the selected option index and (2) a bounding box in absolute image coordinates indicating the target object in the last observed frame. This formulation jointly evaluates semantic future interaction prediction and spatial grounding.

3.2 Data Construction Pipeline

As shown in Figure 2, we construct the benchmark from egocentric video data collected from Ego4D and EPIC-KITCHENS, together with their coarse interaction annotations such as verb–noun categories and interaction timestamps (Grauman et al., 2022; Damen et al., 2018). For each interaction instance, we extract the preceding 16 seconds of video as the historical input and use the annotated interaction contact frame as the future target frame. This construction ensures that the model must forecast the imminent interaction from pre-contact evidence rather than recognize an interaction that is already visually obvious.

We then synthesize the benchmark sample in several stages. First, we use Qwen2.5-VL-72B (Bai et al., 2025b) to localize the interacted object in the future interaction frame and to generate a frame-level caption and a concise interaction phrase for the ground-truth option, e.g., *"grab the plastic storage box"*. Second, we use GPT-5.2 (OpenAI, 2025) to summarize each historical video clip and generate four distractor interaction options. We constrain the distractors to be both semantically plausible and visually grounded: the distractor objects must appear in the last observed frame, and the resulting interaction phrases must remain reasonable under the observed scene context. Third, we combine the distilled frame-level annotations and the original video clip into a unified benchmark sample, and assign each sample a difficulty label and a scene category using GPT-5.

Because future interaction forecasting should remain a genuine prediction problem rather than collapse into near-duplicate recognition, we further perform manual filtering. In particular, we remove overly easy cases where the future interaction frame is visually too similar to the last observed frame and therefore leaks the answer. We also discard samples labeled as *Illogical*, namely cases where the future interaction is not reasonably inferable from the observed history. Finally, we subsample the remaining data to obtain a benchmark set with a controlled difficulty distribution of Easy:Middle:Hard $= 3 : 2 : 1$, resulting in a 1K-sample evaluation set. For each retained sample, we additionally distill a chain-of-thought trajectory with GPT-5.2, which can be used for reasoning analysis or supervision.

3.3 Dataset Statistics

Table 1 summarizes the main statistics of our dataset. Since our current focus is benchmark evaluation, we report the most relevant quantities for future interaction forecasting, including split sizes, source videos, average history length, and difficulty distribution.

3.4 Evaluation

We evaluate the benchmark with two metrics: **Joint Action–Object Option Accuracy** and **mean IoU (mIoU)**. The first measures whether the model selects the correct future interaction option, and the second measures how accurately it localizes the interacted object in the last observed frame.

Joint Action–Object Option Accuracy. Let c_i^* and $\hat{c}_i$ denote the ground-truth and predicted option for sample i, respectively. The accuracy is defined as

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{c}_i = c_i^*], \tag{3}$$

where N is the number of evaluation samples.

Object Grounding mIoU. Let $\hat{b}_i$ and b_i^* denote the predicted and ground-truth target-object boxes for sample i. We compute

$$\mathrm{IoU}(\hat{b}_i, b_i^*) = \frac{\mathrm{area}(\hat{b}_i \cap b_i^*)}{\mathrm{area}(\hat{b}_i \cup b_i^*)}, \tag{4}$$

and report the mean IoU over the evaluation set:

$$\mathrm{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{IoU}(\hat{b}_i, b_i^*). \tag{5}$$

Statistic	Value
Train samples	27094
Val samples	4481
Test samples	1296
Source videos	5124
Average history length	16 s
Easy/Middle/Hard	12281/14496/6094

Table 1: Statistics of constructed dataset.

4 METHOD

4.1 METHOD OVERVIEW

As shown in Figure 3, EGOR1 builds on QWEN2.5-VL-7B, comprising a vision encoder E_v, a multimodal projector P, and a language model f_θ. Given historical frames $X = \{I_1, \ldots, I_T\}$ sampled at 1 FPS and question q, it predicts the next interaction and localizes its object in the last observed frame. The structured answer contains an option $x \in \{1, \ldots, 5\}$ and a target-object <box>. Unlike direct decoding or generic latent future summarization, this task requires reasoning about both the future event and its target.

EGOR1 implements *progressive interaction-state grounding* through two groups of latent tokens. The <|context_latent|> tokens capture a semantically complete future interaction state, while <|object_latent|> tokens refine it toward object-centric evidence and target grounding. This coarse-to-fine design first infers the likely next interaction state and then focuses on the most probable target object.

4.2 FUTURE INTERACTION STATE DISTILLATION

Decisive forecasting cues span local object evidence, manipulation phase, evolving intention, and global scene state. Coarse future summaries may retain appearance but miss these structured semantics. *Future interaction state distillation* therefore constructs a semantically complete target for the <|context_latent|> tokens from the future interaction frame and its paired interaction text.

To this end, a lightweight Multimodal State Teacher receives the future contact frame I_f and paired interaction text c_f only during training. Frozen encoders produce $V_f = E_v(I_f)$ and $T_f = E_t(c_f)$, and K_s learned queries aggregate both modalities through a Q-Former-style module (Li et al., 2023):

$$Z^{s,*} = M_\phi(Q^s; V_f, T_f) \in \mathbb{R}^{K_s \times d}. \tag{6}$$

The teacher aligns these states with a text-guided visual summary and the interaction text, and adds in-batch contrastive alignment (van den Oord et al., 2018). The complete teacher architecture and objective are provided in the supplementary material.

After teacher pretraining, the historical-only model generates $\hat{Z}^s$ and aligns it token-wise to the frozen target:

$$\mathcal{L}_{s\text{-align}} = \frac{1}{K_s} \sum_{k=1}^{K_s} \left(1 - \cos(\hat{z}_k^s, z_k^{s,*})\right). \tag{7}$$

This turns the <|context_latent|> tokens into explicit carriers of future interaction state. The future frame and its derived supervision are never available at inference.

4.3 OBJECT-CENTRIC LATENT REASONING

State latents predict *what interaction state the history is evolving toward*, but not necessarily *which object* it will act upon. In many egocentric scenes, a correct global prediction still leaves substantial ambiguity at the object level: several objects may be visually salient, semantically plausible, and spatially close to the user's hands, while only one of them is the true interaction target. EGOR1 therefore performs a *state-to-target transition*, contracting the global interaction state into object-centric evidence for target grounding.

Conditioned on the predicted state, the model generates object latents

$$\hat{Z}^o = f_\theta^o(X, q, \mathcal{O}, \hat{Z}^s). \tag{8}$$

During training, RoIAlign (He et al., 2017) extracts a positive feature r^+ from the annotated interacted-object region in the future frame. The pooled object latent $\bar{z}^o$ is aligned with r^+ and contrasted against background and non-interacted object regions:

$$\mathcal{L}_{o\text{-align}} = 1 - \cos(P_o(\bar{z}^o), r^+), \tag{9}$$

$$\mathcal{L}_{\text{ctr}} = -\log \frac{e^{s^+}}{e^{s^+} + \sum_m e^{s_m^-}}. \tag{10}$$

Positive alignment grounds the predicted state, whereas hard negatives distinguish nearby objects that afford the same action.

Training uses two supervised stages. Stage 1 optimizes

$$\mathcal{L}_{\text{SFT1}} = \mathcal{L}_{\text{CE}} + \lambda_s \mathcal{L}_{s\text{-align}} + \lambda_o \mathcal{L}_{o\text{-align}} + \lambda_c \mathcal{L}_{\text{ctr}}, \tag{11}$$

and Stage 2 retains $\mathcal{L}_{\text{CE}}$ while relaxing direct latent penalties, allowing the self-generated state and object tokens to adapt to final answer decoding. Region sampling and the complete objectives are given in the supplementary material.

4.4 REINFORCEMENT LEARNING FOR GENERALIZATION

Although the two-stage supervised fine-tuning procedure already teaches the model to produce grounded latent reasoning and structured predictions, it is still optimized under teacher-forced likelihood objectives. As a result, the model may overfit frequent action–object patterns or rely on shortcut correlations in the training distribution, which limits its robustness under unseen scenes, unseen objects, and hard distractor cases. To further improve generalization, we perform a final reinforcement learning stage using *Group Relative Policy Optimization* (GRPO) (Shao et al., 2024). Importantly, the policy is optimized over the generated text tokens, while both semantic correctness and localization quality are incorporated through the reward.

For each input, GRPO samples a group of structured responses and parses every response into an option $\hat{c}^{(g)}$ and a box $\hat{b}^{(g)}$. The reward jointly evaluates both outputs:

$$R^{(g)} = \mathbf{1}[\hat{c}^{(g)} = c^*] + \lambda_{\text{iou}} \text{IoU}(\hat{b}^{(g)}, b^*), \tag{12}$$

$$A^{(g)} = \frac{R^{(g)} - \mu_R}{\sigma_R + \epsilon}. \tag{13}$$

Relative advantages compare semantically and spatially distinct candidates within the same group. The complete clipped GRPO objective and KL regularization are provided in the supplementary material.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Benchmark and input setting. All main results are reported on EGOR1-BENCH1K. Each sample consists of a 16-second egocentric video clip, from which frames are uniformly sampled at 1 FPS as the visual input. The model receives the sampled historical frames, the multiple-choice prediction and localization question, and the five candidate options, and outputs the selected option together with the target-object box in the last observed frame.

Training stages. Our training follows the three-stage pipeline introduced in Section 4.3. We first perform grounded supervised fine-tuning with future interaction state distillation and object-centric latent reasoning enabled. We then conduct a second supervised stage with relaxed latent supervision. Finally, we apply GRPO-based reinforcement learning for generalization, where the reward is composed of the option correctness and the predicted-box IoU.

Implementation details. EGOR1 is built on top of QWEN2.5-VL-7B (Bai et al., 2025b). The number of state latents is set to 4. We train the model on 8 NVIDIA H100 GPUs for 2 epochs using Adafactor (Shazeer & Stern, 2018) with a learning rate of 1×10^{-7} and gradient accumulation step 8. More detailed hyperparameters will be provided in the appendix.

5.2 COMPARED METHODS

We compare three baseline groups under the same option-accuracy and box-mIoU protocol: the proprietary GPT-5.2 model (OpenAI, 2025); zero-shot and CoT-supervised QWEN2.5-VL-7B and QWEN3-VL-8B (Bai et al., 2025b;a); and the latent-reasoning methods MIRAGE and LVR (Yang et al., 2025; Li et al., 2025). This comparison separates general multimodal capability, task-specific supervision, and generic latent visual reasoning.

5.3 MAIN RESULTS

Table 2 reports the main results on EGOR1-BENCH1K. Zero-shot MLLMs are weak on both outputs, and supervised CoT improves the Qwen models but leaves a clear gap, showing that generic video supervision alone is insufficient. Both latent baselines exceed direct decoding, supporting an explicit intermediate predictive representation. EGOR1 performs best with 0.8485 accuracy and 0.4219 mIoU, improving over LVR by 0.0585 and 0.0260, respectively. The gains on both metrics indicate that interaction-state distillation helps infer what happens next, while progressive object grounding identifies where the interaction will occur.

Compared with generic latent reasoning baselines such as MIRAGE and LVR, the gains of EGOR1 show that the key is not just to introduce latent reasoning, but to make it interaction-centric. By distilling future interaction state and progressively grounding it onto the target object, EGOR1 directly addresses semantic uncertainty in future event prediction and target ambiguity in object grounding.

Category	Method	Acc. ↑	mIoU ↑
Proprietary	GPT-5.2	0.3502	0.2802
General	Qwen2.5-VL-7B Zero-shot	0.2398	0.0909
	Qwen2.5-VL-7B CoT-SFT	0.6269	0.3216
	Qwen3-VL-8B Zero-shot	0.2651	0.1676
	Qwen3-VL-8B CoT-SFT	0.6441	0.3061
Latent	Mirage	0.7757	0.3615
	LVR	0.7900	0.3959
	EgoR1	**0.8485**	**0.4219**

Table 2: Main results on EGOR1-BENCH1K.

5.4 ABLATION STUDY

5.4.1 COMPONENT ABLATION

Figure 4 shows that the zero-shot base is weak on both event prediction and grounding, motivating explicit predictive structure.

Future Interaction State Distillation gives the largest isolated accuracy gain, confirming its role in event-level forecasting, whereas Object-Centric Latent Reasoning produces the strongest isolated mIoU by resolving the target object. Their combination improves both metrics, supporting the proposed state-to-target progression. GRPO is weak when applied directly to the base model but improves the full supervised model to 0.8485 accuracy and 0.4219 mIoU. Thus, reinforcement learning

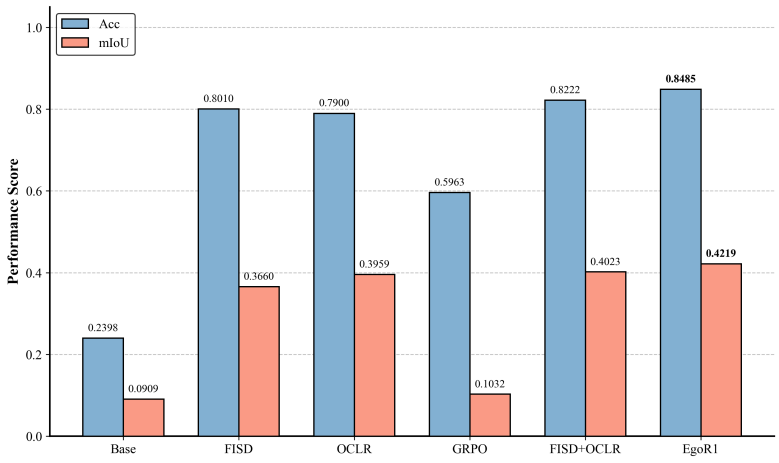


Figure 4: Illustration of component ablation.

is most effective as post-training refinement after the predictive state and object grounding have been established. Complete numerical variants and extended interpretation are reported in the supplementary material.

The two components are complementary: state distillation primarily improves the model's ability to infer what will happen next, whereas object-centric reasoning improves its ability to determine which object will be involved. Their joint gains therefore support a structured state-to-target process rather than either component in isolation.

5.4.2 Ablation on the Multimodal State Teacher

We further study how the design of future interaction state supervision affects the final forecasting performance. Table 3 compares three variants: simple average pooling, a visual-only learned teacher, and our Multimodal State Teacher. The visual-only learned teacher gives modest gains over average pooling, while the multimodal teacher improves both metrics to 0.8485/0.4219. The larger gain appears only when future appearance is paired with interaction text, supporting semantically complete interaction-state supervision rather than generic visual compression. Further analysis is moved to the supplementary material.

State Supervision Variant	Acc. (%) ↑	mIoU (%) ↑
Average Pooling	0.7757	0.3615
Visual-only Learned Teacher	0.7975	0.3669
Multimodal State Teacher	**0.8485**	**0.4219**

Table 3: Ablation on future interaction state supervision.

6 Conclusion

In this paper, we introduce EgoR1-Bench1K, a new benchmark for evaluating future interaction forecasting in multimodal large language models, where the model must jointly predict the next interaction event and localize the corresponding target object from historical egocentric observations. We further propose EgoR1, an interaction-centric latent reasoning framework that transforms future latent reasoning from generic future summarization into progressive interaction-state grounding. By combining future interaction state distillation, object-centric latent reasoning, and a final GRPO-based reinforcement learning stage, EgoR1 consistently improves both future event prediction and object grounding over strong MLLM and latent reasoning baselines. We hope this work can serve as a useful step toward predictive egocentric intelligence and inspire future research on interaction-aware multimodal reasoning, proactive assistance, and embodied decision making.

AI USE STATEMENT

We used generative AI tools to assist benchmark construction. Qwen2.5-VL-72B generated interacted-object boxes, frame-level captions, and interaction phrases from future contact frames. GPT-5.2 summarized historical clips, generated distractor options, and distilled chain-of-thought trajectories for reasoning analysis or supervision. GPT-5 assigned difficulty labels and scene categories. Model-assisted annotations underwent manual verification and filtering as described in Section 3.2 and Supplementary Material, Appendix A. Generative AI assistance was also used for LaTeX template migration and language editing of this submission.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *IEEE/CVF International Conference on Computer Vision*, pp. 13702–13711, 2023.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision*, pp. 720–736, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling–unrolling LSTMs and modality attention. In *IEEE/CVF International Conference on Computer Vision*, pp. 6252–6261, 2019.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 13505–13515, 2021.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language–image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21003–21012, 2022.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL*, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 12585–12602, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J. Guerrero, Giovanni Maria Farinella, and Antonino Furnari. AFF-ttention! affordances and attention models for short-term object interaction anticipation. In *European Conference on Computer Vision*, pp. 167–184, 2024.

OpenAI. Update to GPT-5 system card: GPT-5.2. OpenAI technical report, 2025. Available at `https://openai.com/index/gpt-5-system-card-update-gpt-5-2/`.

Razvan-George Pasca, Alexey Gavryushin, Muhammad Hamza, Yen-Ling Kuo, Kaichun Mo, Luc Van Gool, Otmar Hilliges, and Xi Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18286–18296, 2024.

Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, et al. Modeling fine-grained hand–object dynamics for egocentric video representation learning. *arXiv preprint arXiv:2503.00986*, 2025.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023.

Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. StillFast: An end-to-end approach for short-term object interaction anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3636–3645, 2023.

Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6740–6750, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Yang Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604, 2018.

Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Leveraging next-active objects for context-aware anticipation in egocentric videos. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8657–8666, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MeMViT: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13587–13597, 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-centric video representation for long-term action anticipation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6751–6761, 2024.

Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware egocentric video recognition model. In *IEEE/CVF International Conference on Computer Vision*, pp. 13901–13912, 2023.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.