EgoVITA: Learning to Plan and Verify for Egocentric Video Reasoning

Yogesh Kulkarni and Pooyan Fazli

Arizona State University, Tempe, AZ, USA
ykulka10@asu.edu, pooyan@asu.edu

https://people-robots.github.io/EgoVITA/

Abstract. Egocentric video understanding requires procedural reasoning under partial observability and continuously shifting viewpoints. Current multimodal large language models (MLLMs) struggle with this setting, often generating plausible but visually inconsistent or weakly grounded responses. We introduce **EgoVITA**, a framework that decomposes egocentric video reasoning into a structured *plan-then-verify* process. The model first generates an **egocentric plan**: a causal sequence of anticipated actions from a first-person perspective. This plan is then evaluated by an **exocentric verification** stage that uses third-person reasoning over the same video to verify its spatiotemporal and logical consistency, without exocentric video input. This decomposition enables cross-perspective feedback without requiring paired ego-exo supervision. To train this reasoning process, we adopt Group Relative Policy Optimization (GRPO) with two dense reward signals: one that grounds anticipated actions in subsequent visual observations and another that reinforces consistent third-person verification. EgoVITA achieves state-of-the-art performance on egocentric reasoning benchmarks, outperforming Qwen2.5-VL-7B by $+\mathbf{7.7}$ on EgoBlind and $+\mathbf{4.4}$ on EgoOrient, while maintaining strong generalization on exocentric video tasks with only $52k$ training samples.

Keywords: Egocentric Video Understanding · Procedural Reasoning · Reinforcement Learning

1 Introduction

Egocentric, or first-person, videos play a central role in applications such as life-logging, personal memory assistants [46], and assistive technologies [19, 43]. Unlike exocentric footage, egocentric videos are captured from the actor's viewpoint, resulting in continuous camera motion, partial observability, and frequent occlusions. Effective egocentric understanding therefore requires procedural reasoning to predict how actions unfold, how object states evolve, and how goals are achieved over time. Despite recent advances, multimodal large language models (MLLMs) remain weak in this setting. They often produce plausible yet incorrect descriptions [35], struggle with spatiotemporal reasoning [26, 34], and fail to maintain coherent object and goal representations [25, 32, 37, 48] across time.

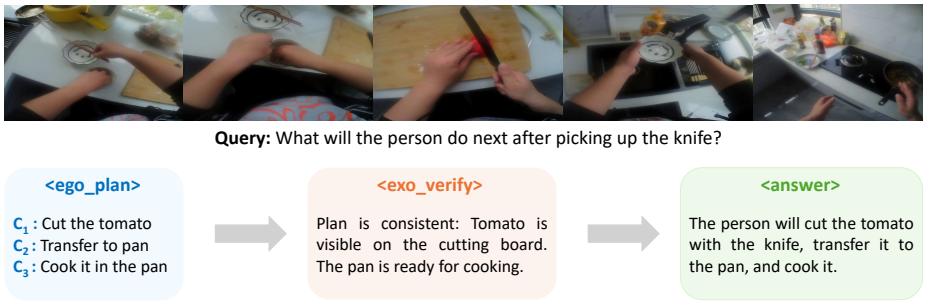

Fig. 1: EgoVITA Framework. Given a video and query, the model first generates an egocentric plan, then verifies it through third-person reasoning over the same video before producing a final answer.

These shortcomings reveal a key gap: current MLLMs lack explicit mechanisms to reason about how actions and object states evolve over time in first-person video. While supervised fine-tuning (SFT) on egocentric datasets can improve in-domain performance, it often degrades accuracy on standard exocentric benchmarks, highlighting the challenge of generalizing across viewpoints. Recent methods [31] attempt to overcome this by scaling to massive datasets, but rely on sparse, rule-based rewards that provide limited guidance for procedural reasoning and still suffer from catastrophic forgetting of exocentric performance.

We hypothesize that these failures stem not from insufficient visual information but from the lack of explicit mechanisms to validate first-person predictions against broader scene constraints. When a model plans from an egocentric view (e.g., "I will reach for the cup"), it must also assess whether that plan is physically plausible (e.g., "a cup is within arm's reach on the counter"). Such validation requires reasoning about scene layout, object affordances, and spatial relationships from a complementary third-person perspective. Recent work suggests that MLLMs can reason across viewpoints, demonstrating perspective-taking and perspective transformation abilities [4, 23, 50]. For instance, given a first-person video, a model may reinterpret the scene from an external observer's perspective (e.g., "A person is reaching toward a cup on the counter"), indicating the ability to shift between egocentric and exocentric descriptions.

Building on this, we introduce **EgoVITA** (**Ego**centric **V**ideo **I**ntelligence and **T**hinking **A**gent), a framework that decomposes egocentric reasoning into two distinct stages: (1) an **egocentric planning** stage that predicts a causal sequence of first-person actions, and (2) an **exocentric verification** stage that audits this plan using third-person reasoning over the same egocentric video, without requiring exocentric input. This separation provides two advantages. First, the planning stage can specialize in egocentric procedural reasoning. Second, by retaining a dedicated verification component grounded in third-person reasoning, the framework preserves exocentric performance and mitigates catastrophic forgetting without requiring synchronized ego-exo video pairs [18].

To train this plan-then-verify framework, we adopt reinforcement learning to provide trajectory-level supervision for multi-step reasoning. Specifically, we use Group Relative Policy Optimization (GRPO) [12], which evaluates multiple candidate trajectories per input and updates the policy based on their relative quality. Within this framework, we introduce two dense reward signals: Anticipatory Cross-Modal Grounding (ACMG), which grounds anticipated actions by encouraging their visual and temporal consistency with subsequent observations, and a *confidence reward*, which reinforces third-person validation of the predicted plan. In summary, our contributions are as follows:

1. We propose EgoVITA, a framework that decomposes egocentric reasoning into an egocentric planner and an exocentric verifier, introducing explicit feedback between first-person prediction and third-person validation.
2. We introduce two dense reward mechanisms, ACMG and a confidence reward, that promote anticipatory grounding of predicted actions and consistent third-person verification.
3. We demonstrate that EgoVITA achieves strong performance across both egocentric and exocentric benchmarks, improving Qwen2.5-VL-7B by $+\mathbf{7.7}$ on EgoBlind and $+\mathbf{4.4}$ on EgoOrient, while maintaining competitive performance on exocentric video tasks within a single unified model.

2 Related Work

Egocentric Video Understanding. Egocentric video reasoning is uniquely challenging due to rapid camera motion, frequent object occlusions [26, 33], and the need to infer the intentions of an unobserved, active person [31, 37]. As a result, recent MLLM research has focused mainly on creating *benchmarks* that highlight specific cognitive failures rather than developing new *methods*. These failures include deficits in long-term memory and procedural reasoning [38, 46, 47], fundamental temporal understanding [20, 34], spatial orientation [10, 17], and intention or Theory of Mind grounding [25, 32, 37]. Models also struggle with fine-grained perception of dynamic object states [48] or scene-text [21, 53], and often show hallucinations [35] or provide unsafe answers in assistive contexts [19, 43]. The few methods proposed, such as EgoThinker [31] and EgoVLM [39], employ reinforcement learning [22]. However, their reliance on simple, sparse rewards for output format and correctness limits long-term procedural reasoning and often causes catastrophic forgetting of general exocentric knowledge. To address this, we introduce a novel framework that leverages dense, predictive rewards to teach complex egocentric procedures while preserving exocentric generalization, tackling a key challenge in ego-exo knowledge transfer [15, 28, 51].

Ego-Exo Video Understanding. Applying MLLMs trained on large corpora of exocentric images and videos to egocentric video faces a fundamental *ego-exo gap* [14]. This domain shift introduces several challenges, including viewpoint disparity that disrupts feature alignment [28], difficulty in matching objects across views [30], and complications in relative spatial reasoning [10]. Prior work

attempts to bridge this gap by learning view-invariant representations [28], using audio as a "sound bridge" [15], or employing SFT pipelines for direct knowledge transfer [44, 51]. While these approaches improve cross-view alignment, they struggle to teach dynamic procedural reasoning and often suffer from catastrophic forgetting of exocentric knowledge. EgoVITA overcomes these limitations by explicitly separating reasoning into an egocentric planning phase and an exocentric verification phase. This design enables the model to learn robust first-person understanding while using verification to preserve third-person generalization. Unlike prior work [18] that relies on synchronized ego-exo video pairs, EgoVITA achieves scalable cross-view consistency without requiring paired data.

3 EgoVITA

We introduce EgoVITA, a reinforcement learning framework that decomposes egocentric video reasoning into two complementary components: an *egocentric planner* that performs procedural reasoning by predicting first-person action sequences, and an *exocentric verifier* that evaluates these plans from a third-person perspective for visual and logical consistency. We train EgoVITA in two phases. We first apply supervised fine-tuning (SFT) to initialize the structure of plan-then-verify. We then optimize the model using Group Relative Policy Optimization (GRPO) with two dense reward functions: Anticipatory Cross-Modal Grounding (ACMG), which encourages visual and temporal alignment between anticipated actions and subsequent visual observations, and a confidence reward, which reinforces consistent third-person validation. An overview of the framework is shown in Figure 1.

3.1 Problem Formulation

We formulate egocentric video reasoning as a sequential decision-making problem, where a multimodal large language model (MLLM) acts as a policy π_θ parameterized by weights θ. Given a video V and a text prompt X, the policy autoregressively generates a *reasoning trajectory*

$$Y = \langle y_1, \ldots, y_T \rangle, \quad y_t \sim \pi_\theta(y_t \mid V, X, y_{<t}), \tag{1}$$

where each y_t is a generated token. We structure the reasoning trajectory Y into three contiguous segments:

$$Y = \left(Y^{\text{plan}}, Y^{\text{verify}}, Y^{\text{ans}}\right),$$

corresponding to the egocentric plan `<ego_plan>`, the exocentric verification `<exo_verify>`, and the final answer block `<answer>`, respectively.

The quality of a reasoning trajectory Y is evaluated by a composite reward function $R(Y)$, which integrates signals encouraging alignment with future visual observations and consistency of third-person validation. The objective is to learn parameters θ that maximize the expected reward:

$$\theta^* = \arg\max_\theta \ \mathbb{E}_{Y \sim \pi_\theta(\cdot|V,X)}[R(Y)]. \tag{2}$$

3.2 Reasoning Decomposition

Egocentric Planning. Given a video V and query X, the model first generates an `<ego_plan>`: a temporally ordered sequence of action clauses $\{c_1, c_2, \ldots, c_K\}$ describing the anticipated actions of the camera wearer (e.g., 1. Cut the tomato. 2. Transfer to pan.). Each clause describes an anticipated action whose outcome can be validated against subsequent visual observations. By generating such procedural plans, the model is encouraged to reason about action ordering, object state transitions, and temporal dependencies from a first-person perspective.

Exocentric Verification. The generated plan is then audited by an `<exo_verify>` stage, which reasons about the scene from a third-person perspective. Operating on the same video, it evaluates the egocentric plan without requiring exocentric video or paired ego-exo data. Given the egocentric plan, the model assesses whether the predicted actions are consistent with scene layout, object affordances, and observable context (e.g., "The knife is visible on the counter. The cutting board is within reach. Plan is consistent."). This verification step evaluates the plausibility of the planned actions under a complementary perspective, rather than relying solely on first-person prediction.

Answer. Finally, the model produces a final `<answer>` conditioned on both the egocentric plan and its exocentric verification. This decomposition grounds the response in both first-person procedural reasoning and third-person validation.

3.3 Training

EgoVITA is trained in two phases. First, SFT initializes the model to produce plan-verify reasoning trajectories. Second, GRPO further optimizes the policy using dense reward signals. We describe each phase below.

Stage I: SFT We first apply SFT to initialize the policy to generate egocentric plans `<ego_plan>`, exocentric verifications `<exo_verify>`, and final answers `<answer>` under cross-entropy supervision.

We rely on human-annotated datasets as supervision. For egocentric planning, we use EgoProceL [3], which provides time-aligned action annotations labeled by human annotators for egocentric videos. We prompt Qwen2.5-VL-72B [2] to generate plan sequences conditioned on these annotations (prompt in Supplementary). For exocentric verification, we construct a dataset based on HD-EPIC [33], which contains human-labeled (action - timestamp) video annotations. Qwen2.5-VL-72B generates third-person chain-of-thought rationales grounded in these annotations (prompt in Supplementary). All generated reasoning traces are filtered and verified using Seed1.5-VL [13] to ensure quality. Both datasets are merged into a unified corpus D and trained with standard cross-entropy:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(V,X,Y^*)\sim D}\left[\log \pi_\theta(Y^* \mid V, X)\right]. \tag{3}$$

where (V, X, Y^*) denotes a video, prompt, and ground-truth reasoning sequence.

Stage II: GRPO After SFT, we further optimize the policy π_θ using GRPO, which performs trajectory-level policy updates based on relative reward comparisons. In egocentric video reasoning, multiple plausible reasoning trajectories may exist for a given input due to scene ambiguity and viewpoint dynamics. GRPO addresses this by sampling a group of candidate trajectories per input and updating the policy according to their relative rewards. For each video-prompt pair (V, X), the policy π_θ generates a group of $k = 8$ rollouts $\{Y_1, \ldots, Y_k\}$, where each trajectory includes a full `<ego_plan>`, `<exo_verify>`, and `<answer>` segments. Each Y_i is scored using a composite reward function $R(Y_i)$, which integrates alignment with future visual observations and consistency of third-person validation.

GRPO Objective. Group-level rewards $\{R(Y_i)\}_{i=1}^k$ are normalized to compute relative advantages:

$$\hat{A}(Y_i) = \frac{R(Y_i) - \mu_R}{\sigma_R + \epsilon}, \tag{4}$$

where μ_R and σ_R are the mean and standard deviation of the rewards within the sampled group. The GRPO objective is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(V,X)} \Bigg[\frac{1}{k} \sum_{i=1}^{k} \Bigg(\min \Big(r_i(\theta)\hat{A}(Y_i), \text{clip}\left(r_i(\theta), 1-\epsilon, 1+\epsilon\right) \hat{A}(Y_i) \Big) - \beta_{\text{KL}} \log \frac{\pi_\theta(Y_i \mid V, X)}{\pi_{\text{ref}}(Y_i \mid V, X)} \Bigg) \Bigg], \tag{5}$$

where the probability ratio is

$$r_i(\theta) = \frac{\pi_\theta(Y_i \mid V, X)}{\pi_{\theta_{\text{old}}}(Y_i \mid V, X)}. \tag{6}$$

and β_{KL} limits deviation from the reference policy π_{ref} (initialized from SFT).

Composite Reward (R). The reward is a weighted sum of four components:

$$R = w_f R_{\text{format}} + w_a R_{\text{answer}} + w_g R_{\text{ACMG}} + w_c R_{\text{confidence}}, \tag{7}$$

where:

1. **Format Reward (R_{format})** is a sparse binary signal that checks for correct tag usage and output structure,
2. **Answer Reward (R_{answer})** is a sparse reward measuring alignment between the generated answer and the ground-truth label,
3. **ACMG Reward (R_{ACMG})** is a dense reward that promotes alignment between predicted plan clauses and future visual frames, encouraging temporal and visual grounding, and
4. **Confidence Reward ($R_{\text{confidence}}$)** is a dense reward that reinforces consistent third-person validation of the predicted plan.

Sparse rewards (R_{format}, R_{answer}) provide binary pass/fail scores based on output structure and final correctness, while dense rewards (R_{ACMG}, $R_{\text{confidence}}$) provide continuous, fine-grained feedback that evaluates the quality of intermediate reasoning within the trajectory.

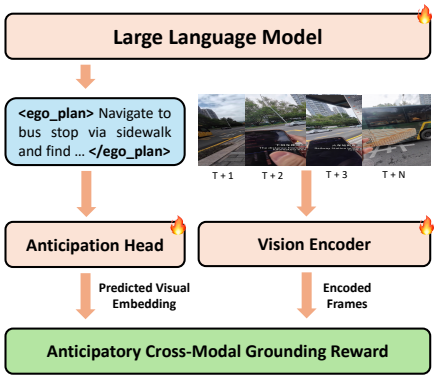

Fig. 2: ACMG Mechanism. Each plan clause is projected into a visual embedding space and matched against the embeddings of the next N frames. The clause-level reward is the maximum cosine similarity over this temporal window.

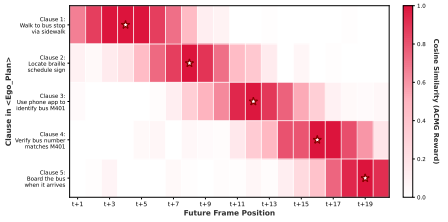

Fig. 3: Temporal Alignment. Heatmap of clause-to-frame similarity, with white stars ($\star$) marking peak matches. The diagonal pattern shows earlier clauses aligning with near-future frames and later clauses with subsequent frames, consistent with anticipatory temporal grounding.

Anticipatory Cross-Modal Grounding (ACMG). To encourage egocentric plans to align with future visual observations, we introduce the ACMG reward. This dense reward measures the visual and temporal alignment between each plan clause and subsequent observations in the video (Figure 2). For each clause c_i, we extract its final text hidden state h_i^{text} from the language decoder of the MLLM and project it into the visual embedding space via a small trainable MLP, referred to as the Anticipation Head. The output of this projection is a predicted future visual embedding, denoted as $\hat{v}_i = \text{MLP}(h_i^{\text{text}})$.

Let v_{t+n} denote the visual embedding of frame $t+n$, extracted from the visual encoder of the MLLM. We consider the next $N = 16$ frames following timestamp t as a temporal horizon. This window captures near-future action outcomes while limiting influence from distant events. The ACMG reward for each clause c_i is defined as the maximum cosine similarity between the predicted embedding and the embeddings of these future frames (Figure 3):

$$R_{\text{ACMG}}(c_i) = \max_{n \in \{1,\ldots,N\}} \frac{\hat{v}_i \cdot v_{t+n}}{\|\hat{v}_i\| \, \|v_{t+n}\|}. \tag{8}$$

The max operator selects the frame that best aligns with the predicted action, making the reward robust to small temporal offsets and variable action durations. The trajectory-level ACMG reward is computed by averaging across clauses:

$$R_{\text{ACMG}} = \frac{1}{K} \sum_{i=1}^{K} R_{\text{ACMG}}(c_i). \tag{9}$$

Ablation studies in the Supplementary Material validate the choice of the max operator and the temporal window size $N = 16$.

Confidence Reward. To improve the quality of <exo_verify>, we introduce a confidence reward $R_{\text{confidence}}$ that applies a relative preference signal over candidate verification trajectories.

Teacher-guided phase. During the first 200 RL steps, each of the $k = 8$ rollout verifications is compared to a teacher-generated reference verification. The confidence reward encourages the policy to assign a higher likelihood to the teacher output relative to its own rollout:

$$R_{\text{confidence}} = \gamma \Big[\log \pi_\theta \big(y^T \mid V, X, Y^{\text{plan}}\big) - \log \pi_\theta \big(y^P \mid V, X, Y^{\text{plan}}\big) \Big], \tag{10}$$

where γ is a scaling coefficient controlling the strength of the preference signal, and y^T and y^P denote the teacher and policy verification outputs, respectively. We use Qwen2.5-VL-72B [2] as a teacher model that generates exocentric verification rationales conditioned on the egocentric plan. This warm-up phase aligns the exocentric verifier with teacher-generated third-person audits conditioned on the predicted first-person plan.

Self-ranking phase. After the warm-up period, each rollout verification is scored using the composite reward (Eq. 7). The highest-scoring verification y^{chosen} and lowest-scoring verification y^{rejected} are compared using a log-probability preference objective:

$$R_{\text{confidence}} = \begin{cases} \gamma \cdot (\log \pi_\theta(y_{\text{chosen}} \mid V, X, Y^{\text{plan}}) \\ \quad - \log \pi_\theta(y_{\text{rejected}} \mid V, X, Y^{\text{plan}})), & \text{if scores differ;} \\ 1, & \text{otherwise.} \end{cases} \tag{11}$$

This update increases the likelihood of third-person audits that better reconcile the predicted first-person plan with first-person visual observations.

3.4 Exocentric Regularization

Training solely on egocentric data can lead to catastrophic forgetting, reducing performance on exocentric tasks. To mitigate this, we periodically interleave GRPO updates with a lightweight exocentric regularization step. Specifically, after every 200 RL iterations, we optimize the model on an auxiliary exocentric VideoQA dataset (MSR-VTT [45]) using a cross-entropy loss $\mathcal{L}_{\text{exo}}$. This loss is combined with the GRPO objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GRPO}} + \lambda_{\text{exo}} \mathcal{L}_{\text{exo}}, \tag{12}$$

where λ_{exo} controls the strength of the exocentric regularization term. We use MSR-VTT because its diverse third-person video content spans a broad range of everyday scenes, providing complementary supervision to counterbalance egocentric specialization. No additional exocentric signals are introduced during RL training. Ablation studies in the Supplementary Material analyze the effect of λ_{exo} and confirm that this regularization preserves exocentric performance while maintaining gains on egocentric reasoning tasks.

4 Experiments and Evaluation

Implementation Details. We implement EgoVITA using the open-source MS-SWIFT [52] library. In Stage I (SFT), we train on two datasets: EgoProceL [3] ($5k$ samples) for egocentric planning and HD-EPIC [33] ($7k$ samples) for exocentric verification. Chain-of-thought reasoning traces for both planning and verification are generated once using Qwen2.5-VL-72B [2], conditioned on ground-truth annotations. The SFT model is trained for one epoch with a learning rate of $5e-6$. In Stage II (GRPO), we sample rollouts from HD-EPIC and EgoIT ($40k$ samples) [46]. We set the policy temperature $\beta_{\text{KL}} = 0.1$, the confidence scaling coefficient $\gamma = 0.1$, and the exocentric regularization coefficient $\lambda_{\text{exo}} = 0.05$. We apply EgoVITA to three state-of-the-art MLLMs: Qwen2.5-VL-7B [2], InternVL-3.5-8B [42], and Qwen3-VL-8B [1]. All models are trained on 8 NVIDIA L40S GPUs using DoRA [27] adapters ($r = 8$, $\alpha = 16$) on the vision encoder, projector, and language decoder. The Anticipation Head (MLP) is used only during Stage II to compute R_{ACMG} and is discarded at inference, adding no overhead to the final model. The reward weights are set to $\{w_f, w_a, w_g, w_c\} = \{0.1, 0.3, 0.3, 0.3\}$. The format reward receives a lower weight as it saturates early in training, while the remaining three signals are weighted equally to balance their influence during optimization. At inference, EgoVITA runs at 84.3 token/s with a time-to-first-token of 0.6 s on a single L40S GPU, adding only 0.1 s overhead compared to the base Qwen2.5-VL model.

Benchmarks. We evaluate EgoVITA on several egocentric understanding benchmarks, including EgoBlind [43] (visual assistance for blind users), EgoPlan [6] (egocentric task planning), EgoThink [7] (first-person perspective thinking), EOC-Bench [48] (embodied cognition across past, present, and future), and EgoOrientBench [17] (object orientation understanding). We also evaluate on exocentric video understanding benchmarks: MVBench [24] (temporal perception-to-cognition QA), Video-MME [9] (comprehensive video understanding), LVBench [41] (long-video reasoning), and TOMATO [36] (visual temporal reasoning).

4.1 Results

We compare EgoVITA against four baselines: (1) the original foundation models, (2) the same models after SFT, (3) SFT models further trained with GRPO using only final-answer and format rewards, and (4) state-of-the-art MLLMs for egocentric video understanding.

EgoVITA improves all foundation models. As shown in Table 1, EgoVITA consistently enhances egocentric performance across all three foundation models: Qwen2.5-VL gains +**7.7** on EgoBlind, +**3.7** on EgoThink, and +**4.4** on EgoOrient. InternVL-3.5 improves by +**3.6** on EgoBlind and +**3.8** on EgoOrient, while the stronger Qwen3-VL achieves gains of +**3.5** and +**2.3** on the same tasks, suggesting that the improvements generalize across model capacities. These egocentric gains come without hurting exocentric performance, as all three improve or maintain their scores on exocentric benchmarks (e.g., InternVL gains +**2.7** on Video-MME).

Table 1: Evaluation of EgoVITA on egocentric and exocentric benchmarks. Best scores are in **bold**, and improvements from EgoVITA are highlighted in green with 95% confidence interval (CI) margins (±) via bootstrap. ‡ Improvement not statistically significant at the 0.05 level. ♢ Self-trained variant: SFT data and teacher confidence signals generated by Qwen3-VL-8B itself (no external 72B teacher).

	Egocentric Video Understanding					Exocentric Video Understanding			
Model	**EgoBlind**	**EgoPlan**	**EgoThink**	**EOC-Bench**	**EgoOrient**	**MVBench**	**Video-MME**	**LVBench**	**Tomato**
InternVL-3.5 Family									
InternVL-3.5	47.8	34.0	58.5	37.1	36.3	71.4	65.6	44.4	31.4
SFT	49.2	34.8	59.3	38.1	37.8	68.2	62.1	41.2	29.8
GRPO (Format + Answer)	49.8	35.1	59.6	38.6	38.4	68.9	63.2	41.8	30.1
EgoVITA	**51.4** (+3.6±0.5)	**35.9** (+1.9±0.5)	**60.8** (+2.3±0.5)	**39.7** (+2.6±0.6)	**40.1** (+3.8±0.7)	**73.8** (+2.4±0.6)	**68.3** (+2.7±0.6)	**47.1** (+2.7±0.7)	**33.2** (+1.8±0.6)
Qwen2.5-VL Family									
Qwen2.5-VL	29.7	30.2	48.2	41.6	47.6	65.2	62.2	42.8	29.3
SFT	34.8	31.3	49.1	41.8	48.1	63.2	60.1	40.2	29.0
GRPO (Format + Answer)	35.6	31.4	49.0	42.2	48.6	66.4	62.9	40.9	29.8
EgoVITA	**37.4** (+7.7±0.9)	**32.9** (+2.7±0.6)	**51.9** (+3.7±0.7)	**44.1** (+2.5±0.6)	**52.0** (+4.4±0.8)	**66.8** (+1.6±0.5)	**64.6** (+2.4±0.6)	**43.6** (+0.8±0.5)	**30.2** (+0.9±0.5)
Qwen3-VL Family									
Qwen3-VL	48.4	33.7	62.7	46.8	60.8	68.4	71.4	56.1	34.0
SFT	50.1	34.2	63.1	47.5	61.8	68.2	71.2	56.3	34.2
GRPO (Format + Answer)	50.8	34.5	63.4	47.9	62.3	68.1	71.1	56.4	34.3
EgoVITA	**51.9** (+3.5±0.7)	**35.0** (+1.3±0.5)	**63.9** (+1.2±0.5)	**48.6** (+1.8±0.6)	**63.1** (+2.3±0.6)	**69.2** (+0.8±0.5)	**72.2** (+0.8±0.5)	**56.5** (+0.4±0.4)‡	**34.5** (+0.5±0.4)
Self-Trained									
EgoVITA♢	50.2 (+1.8)	34.1 (+0.4)	63.0 (+0.3)	47.4 (+0.6)	61.5 (+0.7)	68.3 (-0.1)	71.5 (+0.1)	56.0 (-0.1)	33.9 (-0.1)

Table 2: Comparison with State of the Art. All results are reproduced by us.

	Egocentric Video Reasoning				Basic Exocentric Understanding			
Variant	**EgoBlind**	**EgoPlan**	**EOC-Bench**	**EgoSchema**	**MMMU**	**DocVQA**	**MME (Perception)**	**VQAv2**
Qwen2-VL	41.0	32.5	39.6	64.2	51.0	79.5	1677	80.4
SFT	44.2	33.1	40.8	65.4	49.8	78.1	1654	79.2
GRPO (Format + Answer)	45.6	33.6	41.4	66.1	50.4	78.9	1665	79.8
EgoThinker [31]	43.8	34.1	37.1	**68.2**	49.2 (-1.8)	69.0 (-10.5)	1644 (-33)	78.1 (-2.3)
EgoVITA	**46.8** (+5.8±0.3)	**34.6** (+2.1±0.4)	**42.7** (+3.1±0.2)	67.9 (+3.7±0.5)	**51.3** (+0.3±0.4)	**80.1** (+0.6±0.3)	**1682** (+5±7)	**81.2** (+0.8±0.6)

Furthermore, the self-trained variant using Qwen3-VL-8B (EgoVITA♢), where both SFT data and teacher confidence signals are generated by the same 8B model, still outperforms the Qwen3-VL baseline on all egocentric benchmarks. This suggests that the improvements are driven by the plan–verify training framework itself, rather than reliance on a larger external teacher.

Comparison with state-of-the-art. Work on egocentric video reasoning with MLLMs is limited. EgoThinker [31] built upon Qwen2-VL [40] is the closest related method that enables a controlled backbone comparison. Table 2 compares EgoVITA and EgoThinker across standard egocentric reasoning benchmarks and general exocentric tasks, including MMMU [49], DocVQA [29], MME [8], and VQAv2 [11]. Although EgoThinker is trained with approximately **5M** samples, it shows substantial performance degradation on several exocentric benchmarks (e.g., −**10.5** on DocVQA and −**1.8** on MMMU), suggesting limited cross-view retention under large-scale egocentric training. In contrast, EgoVITA achieves stronger egocentric performance (e.g., +**5.8** over the base model on EgoBlind) using **52k** training samples. Furthermore, the proposed dense reward design is associated with improved egocentric reasoning while maintaining competitive performance on exocentric benchmarks (e.g., +**0.6** on DocVQA), indicating improved cross-view stability relative to large-scale egocentric fine-tuning.

RL is essential. SFT improves output formatting but provides limited gains in egocentric reasoning and can degrade exocentric performance. For example, Qwen2.5-VL after SFT drops by −**2.0** on MVBench and −**2.1** on Video-MME,

Fig. 4: EgoVITA learns to "look ahead" in egocentric video. EAGLE [5] visualizations show that EgoVITA progressively concentrates attention on task-relevant objects (e.g., garlic, egg bowl, pan) and hand regions across frames, whereas Qwen2.5-VL shows scattered, temporally static activations over background elements (e.g., shelves, walls). The focused, temporally evolving attention patterns in EgoVITA are associated with correct action predictions (green), whereas the baseline often produces incorrect or visually inconsistent outputs (red).

indicating a trade-off between egocentric specialization and exocentric retention. In contrast, applying RL through EgoVITA recovers exocentric performance, improving MVBench and Video-MME by $+\mathbf{3.6}/+\mathbf{4.5}$ points over SFT while also further increasing EgoBlind by $+\mathbf{2.6}$ points. A similar trend is observed for InternVL-3.5, where RL improves Video-MME by $+\mathbf{6.2}$ points and EgoBlind by $+\mathbf{2.2}$ points over SFT.

Dense rewards matter. GRPO using only format and answer rewards provides limited gains over SFT, as these sparse signals primarily optimize final-output correctness. Adding the dense rewards (ACMG and Confidence) yields consistent additional improvements: $+\mathbf{1.8}$ on EgoBlind and $+\mathbf{3.4}$ on EgoOrient for Qwen2.5-VL, and $+\mathbf{1.6}/+\mathbf{1.7}$ on the same tasks for InternVL-3.5. These results suggest that grounding plans in future visual observations and reinforcing cross-perspective verification both contribute to improved egocentric reasoning.

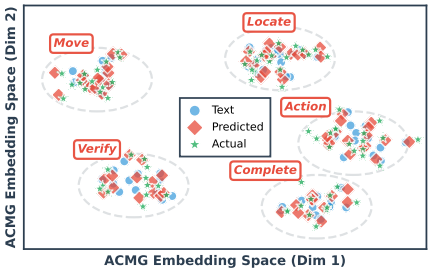

Fig. 5: ACMG Embedding Structure. t-SNE of text (blue), predicted (red), and actual (green) embeddings. Alignment within distinct action clusters indicates cross-modal grounding.

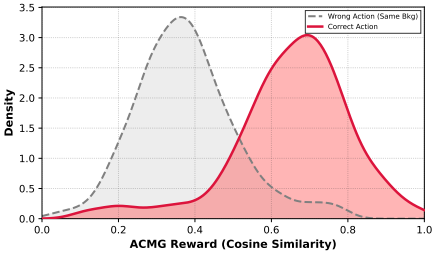

Fig. 6: Static Bias Check. Reward density for matched vs. same-scene/wrong-action pairs. The gap ($\Delta = 0.33$) shows sensitivity to temporal dynamics rather than static backgrounds

Visual Attribution Analysis. We apply EAGLE [5] to visualize which regions contribute to action prediction (Figure 4). Qwen2.5-VL shows largely temporally static saliency, with attention frequently concentrated on background elements and limited focus on task-relevant objects. This behavior is accompanied by incorrect action predictions in several cases. In contrast, EgoVITA demonstrates temporally progressive attribution patterns that shift across frames, increasingly focusing on task-critical objects (e.g., garlic, egg bowl) as the action unfolds. This behavior is consistent with the influence of R_{ACMG}, which encourages alignment between predicted actions and future visual observations. The results suggest that the dense grounding signal promotes greater sensitivity to regions that are predictive of subsequent task dynamics, rather than static scene features.

Semantic Structure Analysis. To examine if EgoVITA learns structured visual-linguistic representations, we categorize `<ego_plan>` clauses into five procedural steps (Move → Locate → Action → Verify → Complete) using GPT-4o [16]. Figure 5 shows a t-SNE projection of the corresponding embeddings, revealing two key properties. First, *between-cluster separation* is visible: embeddings corresponding to different procedural stages form separable clusters, suggesting structured differentiation across task phases. Second, *within-cluster alignment* is strong: the MLP-predicted visual embeddings (red diamonds ◊) align closely with embeddings of future frames (green stars ⋆), indicating consistent cross-modal alignment between predicted plan clauses and subsequent visual observations.

Robustness to Static Feature Bias. To evaluate whether ACMG captures temporal dynamics rather than static background features, we conduct a controlled analysis on 850 EgoPlan [6] samples (Figure 6). For each valid action clause, we pair it with "Wrong Action" frames from the same scene (identified by Qwen3-VL-30B-A3B [1]), controlling for static background and object identity while varying only the action. If R_{ACMG} relied primarily on static scene features, the score distributions would overlap. Instead, correct predictions yield a mean similarity of $\mu = 0.68$, whereas same-scene wrong-action pairs drop to $\mu = 0.35$

Table 3: Task-level breakdown. EgoBlind categories from [43] are Tool Use (TU), Information Reading (IR), Navigation (NV), Safety Warning (SW), Social Communication (SC), and Other Resources (OR). EOC-Bench temporal dimensions from [48] are Past (Pst), Present (Prs), and Future (Fut). EgoOrientBench tasks from [17] are Choose (Cho), Verify (Ver), and Freeform (FF).

	EgoBlind							EOC-Bench				EgoOrientBench			
Method	TU	IR	NV	SW	SC	OR	Avg	Pst	Prs	Fut	Avg	Cho	Ver	FF	Avg
Qwen2.5-VL	34.3	35.0	12.8	30.8	37.8	27.5	29.7	33.7	48.8	42.2	41.6	45.0	62.0	35.8	47.6
SFT	36.3	38.6	15.8	41.2	41.4	35.5	34.8	33.9	49.1	42.4	41.8	45.5	62.5	36.3	48.1
GRPO	36.7	39.2	16.3	42.9	41.9	36.6	35.6	34.2	49.4	43.0	42.2	46.0	63.0	36.8	48.6
EgoVITA	**37.4**	**40.5**	**17.4**	**46.6**	**43.2**	**39.3**	**37.4** (+7.7)	**36.5**	**50.2**	**45.6**	**44.1** (+2.5)	**51.0**	**65.5**	**39.5**	**52.0** (+4.4)

($\Delta = 0.33$). This separation indicates that the Anticipation Head is sensitive to action-dependent visual changes, suggesting that the reward emphasizes temporal variation rather than static scene content.

Task-level analysis. Table 3 shows that the gains from EgoVITA are concentrated in specific categories rather than arising uniformly from SFT distillation. On EgoBlind, the largest improvements over SFT occur in Safety Warning (+**5.4**) and Other Resources (+**3.8**). On EOC-Bench, Future reasoning benefits most (+**3.2**). On EgoOrientBench, the Choose task shows a substantial improvement (+**5.5**). These non-uniform, category-specific gains beyond SFT indicate that the improvements stem from the proposed plan-then-verify framework and its dense reward design rather than general distillation.

4.2 Ablation Studies

RQ1: Is teacher-guided initialization necessary? Table 4 isolates the effect of the 200-step teacher-guided initialization during the RL stage. The model achieves substantial gains (average +2.9) even without teacher supervision during RL. Since the SFT policy is initialized using human-annotated data, it already provides a reasonable starting point for self-ranking optimization. These results suggest that the improvements are primarily driven by the reward formulation rather than reliance on teacher guidance.

RQ2: Are ACMG and confidence rewards complementary? Table 5 examines the individual and combined effects of ACMG and the confidence reward. ACMG primarily improves performance on tasks requiring temporal grounding (e.g., EgoPlan +0.6), suggesting that aligning predicted plans with future visual observations benefits procedural reasoning. In contrast, the confidence reward yields gains on tasks emphasizing verification consistency (e.g., EOC-Bench +1.3), indicating its role in refining third-person audits of the egocentric plan. When combined, the two rewards produce larger improvements (e.g., EOC-Bench +1.9) than either component alone, suggesting that temporal grounding and verification-based preference signals provide complementary benefits.

RQ3: Is disentangled reasoning into `<ego_plan>` and `<exo_verify>` necessary? Table 6 shows that removing either component degrades performance.

Table 4: Teacher-Guided Warm-up. Effect of removing the 200-step teacher phase from the confidence reward.

Variant	Egocentric Benchmarks			
	EgoBlind	EgoPlan	EgoThink	EOC-Bench
Qwen2.5-VL-7B	29.7	30.2	48.2	41.6
No Teacher Warm-up	36.0 (+6.3)	31.8 (+1.6)	50.5 (+2.3)	43.0 (+1.4)
EgoVITA (Full)	**37.4** (+7.7)	**32.9** (+2.7)	**51.9** (+3.7)	**44.1** (+2.5)

Table 5: ACMG and confidence reward on Qwen2.5-VL. Removing either component degrades performance.

Variant	EgoPlan	EOC-Bench	Video-MME	Tomato
GRPO (Format + Answer)	31.4	42.2	62.9	29.8
+ ACMG only	32.0	43.1	63.4	29.9
+ Confidence only	32.3	43.5	63.8	30.0
EgoVITA (Full)	**32.9** (+1.5)	**44.1** (+1.9)	**64.6** (+1.7)	**30.2** (+0.4)

Table 6: Disentangled reasoning. Removing `<ego_plan>` harms egocentric reasoning, while removing `<exo_verify>` reduces exocentric performance.

Variant	EgoPlan	EOC-Bench	Video-MME	Tomato
GRPO (Format + Answer)	31.4	42.2	62.9	29.8
w/o `<ego_plan>`	29.6	40.1	63.0	29.7
w/o `<exo_verify>`	31.5	42.5	61.2	28.3
EgoVITA (Full)	**32.9** (+1.5)	**44.1** (+1.9)	**64.6** (+1.7)	**30.2** (+0.4)

Table 7: Anticipatory (ACMG) vs. Present Grounding. Comparison with a present-frame baseline (grounding on frame t). Both improve performance, with ACMG yielding substantially larger gains.

Model	EgoBlind	EgoPlan	EgoThink	EOC-Bench
Qwen2.5-VL-7B	29.7	30.2	48.2	41.6
Present	33.8 (+4.1)	31.2 (+1.0)	50.1 (+1.9)	42.8 (+1.2)
ACMG	**37.4** (+7.7)	**32.9** (+2.7)	**51.9** (+3.7)	**44.1** (+2.5)

Eliminating `<ego_plan>` significantly degrades egocentric performance (e.g., −1.8 on EgoPlan), indicating the importance of modeling action sequences before answer generation. Conversely, removing `<exo_verify>` reduces exocentric performance (e.g., −1.7 on Video-MME), suggesting that the verification stage contributes to maintaining general third-person reasoning capabilities. Together, these results indicate that the planning and verification components play complementary roles: `<ego_plan>` supports task-specific procedural reasoning, while `<exo_verify>` helps retain broader cross-view generalization. The full EgoVITA framework balances these components, achieving improved performance across both egocentric and exocentric benchmarks (e.g., +1.7 on Video-MME).

RQ4: Why does anticipatory grounding (ACMG) improve procedural reasoning? Table 7 contrasts present-frame grounding with anticipatory grounding. Models trained with present-frame alignment primarily capture the current static scene, yielding limited improvements (e.g., +1.0 on EgoPlan). In contrast, ACMG aligns anticipated actions with future observations over a temporal horizon, encouraging consistency between planned actions and their subsequent outcomes. This anticipatory grounding is associated with larger gains (e.g., +2.7 on EgoPlan and +7.7 on EgoBlind). As illustrated in Figure 7 (Left), ACMG correctly anticipates the "sweep cane" action before the blind person reaches the curb.

5 Conclusion

We presented EgoVITA, a framework for egocentric video reasoning that decomposes reasoning into a plan-then-verify process consisting of egocentric planning and exocentric verification. By coupling this structure with dense rewards for anticipatory grounding and third-person verification, EgoVITA learns to generate

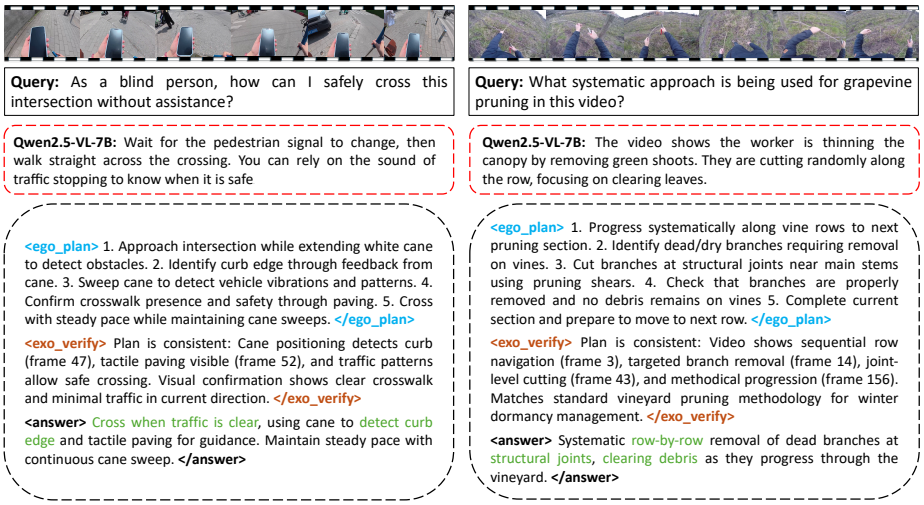

Fig. 7: EgoVITA Qualitative Examples. Left: For a blind person, the `<ego_plan>` generates sequential, task-oriented actions. **Right:** For a procedural task, `<ego_plan>` outlines a structured step-by-step plan. In both cases, `<exo_verify>` evaluates consistency with relevant visual frames, producing a grounded `<answer>`. In contrast, Qwen2.5-VL produces *unsafe* visually inconsistent responses.

procedurally coherent and visually grounded plans and verifications. Empirical results demonstrate improvements in egocentric reasoning while preserving performance on exocentric video understanding tasks, highlighting the effectiveness of cross-perspective reasoning without requiring paired ego-exo supervision. The ability to maintain exocentric performance while specializing in egocentric reasoning opens the possibility of a single unified model serving both first-person assistive applications and general video understanding.

Future work will extend this framework to streaming egocentric scenarios, where the model must plan and verify incrementally as new frames arrive rather than reasoning over an entire video. Adaptive temporal windows that expand for long, multi-step activities and contract for rapid actions could improve grounding accuracy across diverse tasks. Finally, iterative plan-verify reasoning could enable predicted plans to be refined through repeated cross-perspective feedback when visual evidence is ambiguous or multiple plausible future plans exist.

Acknowledgments

This research was supported by the National Eye Institute (NEI) of the National Institutes of Health (NIH) under award number R01EY034562. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We acknowledge Research Computing at Arizona State University for providing computing resources.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
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. Bansal, S., Arora, C., Jawahar, C.: My view is the best view: Procedure learning from egocentric videos. In: European Conference on Computer Vision (ECCV) (2022)
4. Bu, Y., Wu, X., Zhao, Z., Cai, Y., Hsu, D., Liu, Q.: Walk in others' shoes with a single glance: Human-centric visual grounding with top-view perspective transformation. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2025)
5. Chen, R., Guo, X., Liu, K., Liang, S.Y., Liu, S., Zhang, Q., Zhang, H., Cao, X.: Where mllms attend and what they rely on: Explaining autoregressive token generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
6. Chen, Y., Ge, Y., Ge, Y., Ding, M., Li, B., Wang, R., Xu, R., Shan, Y., Liu, X.: Egoplan-bench: Benchmarking multimodal large language models for human-level planning. International Journal of Computer Vision (IJCV) (2026)
7. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
8. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: Neural Information Processing Systems (NeurIPS) (2023)
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. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
10. Gholami, M., Rezaei, A., Weimin, Z., Zhang, Y., Akbari, M.: Spatial reasoning with vision-language models in ego-centric multi-view scenes. In: International Conference on Learning Representations (ICLR) (2026)
11. Ging, S., Bravo, M.A., Brox, T.: Open-ended vqa benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. In: International Conference on Learning Representations (ICLR) (2024)
12. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
13. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
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. In: Neural Information Processing Systems (NeurIPS) (2025)
15. Huang, S., Wu, J., Wei, X., Cai, Y., Jiang, D.S., Wang, Y.: Sound bridge: Associating egocentric and exocentric videos via audio cues. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)

16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Jung, J.H., Kim, E.T., Kim, S., Lee, J.H., Kim, B., Chang, B.: Is 'right'right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
18. Jung, M., Xiao, J., Kim, J., Zhang, B.T., Yao, A.: Egoexo-con: Exploring view-invariant video temporal understanding. arXiv preprint arXiv:2510.26113 (2025)
19. Kim, J., Park, J., Park, J., Lee, S., Chung, J., Kim, J., Joung, J.H., Yu, Y.: Guidedog: A real-world egocentric multimodal dataset for blind and low-vision accessibility-aware guidance. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2026)
20. Kulkarni, Y., Fazli, P.: Videopasta: 7k preference pairs that matter for video-llm alignment. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2025)
21. Kulkarni, Y., Fazli, P.: Videosavi: Self-aligned video language models without human supervision. In: Conference on Language Modeling (COLM) (2025)
22. Kulkarni, Y., Fazli, P.: Avatar: Reinforcement learning to see, hear, and reason over video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
23. Lee, P.Y., Je, J., Park, C., Uy, M.A., Guibas, L., Sung, M.: Perspective-aware reasoning in vision-language models via mental imagery simulation. In: International Conference on Computer Vision (ICCV) (2025)
24. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
25. Li, Y., Veerabadran, V., Iuzzolino, M.L., Roads, B.D., Celikyilmaz, A., Ridgeway, K.: Egotom: Benchmarking theory of mind reasoning from egocentric videos. arXiv preprint arXiv:2503.22152 (2025)
26. Liang, S., Zhong, Y., Hu, Z.Y., Tao, Y., Wang, L.: Fine-grained spatiotemporal grounding on egocentric videos. In: International Conference on Computer Vision (ICCV) (2025)
27. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation. In: International Conference on Machine Learning (ICML) (2024)
28. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Viewpoint rosetta stone: Unlocking unpaired ego-exo videos for view-invariant representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
29. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021)
30. Mur-Labadia, L., Santos-Villafranca, M., Bermudez-Cameo, J., Perez-Yus, A., Martinez-Cantin, R., Guerrero, J.J.: O-mama: Learning object mask matching between egocentric and exocentric views. In: International Conference on Computer Vision (ICCV) (2025)
31. Pei, B., Huang, Y., Xu, J., He, Y., Chen, G., Wu, F., Qiao, Y., Pang, J.: Egothinker: Unveiling egocentric reasoning with spatio-temporal cot. In: Neural Information Processing Systems (NeurIPS) (2025)

32. Peng, T., Hua, J., Liu, M., Lu, F.: In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. In: Neural Information Processing Systems (NeurIPS) (2025)
33. Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K.K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., et al.: Hd-epic: A highly-detailed egocentric video dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
34. Plizzari, C., Tonioni, A., Xian, Y., Kulshrestha, A., Tombari, F.: Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
35. Seth, A., Tyagi, U., Selvakumar, R., Anand, N., Kumar, S., Ghosh, S., Duraiswami, R., Agarwal, C., Manocha, D.: Egoillusion: Benchmarking hallucinations in egocentric video understanding. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2025)
36. Shangguan, Z., Li, C., Ding, Y., Zheng, Y., Zhao, Y., Fitzgerald, T., Cohan, A.: Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In: International Conference on Learning Representations (ICLR) (2025)
37. Sun, P., Xiao, J., Tse, T.H.E., Li, Y., Akula, A., Yao, A.: Visual intention grounding for egocentric assistants. In: International Conference on Computer Vision (ICCV) (2025)
38. Tian, S., Wang, R., Guo, H., Wu, P., Dong, Y., Wang, X., Yang, J., Zhang, H., Zhu, H., Liu, Z.: Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2026)
39. Vinod, A., Pandit, S., Vavre, A., Liu, L.: Egovlm: Policy optimization for egocentric video understanding. arXiv preprint arXiv:2506.03097 (2025)
40. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
41. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: International Conference on Computer Vision (ICCV) (2025)
42. 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)
43. 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. In: Neural Information Processing Systems (NeurIPS) (2025)
44. Xu, J., Huang, Y., Pei, B., Hou, J., Li, Q., Chen, G., Zhang, Y., Feng, R., Xie, W.: Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. In: International Conference on Learning Representations (ICLR) (2025)
45. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
46. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
47. Ye, H., Zhang, H., Daxberger, E., Chen, L., Lin, Z., Li, Y., Zhang, B., You, H., Xu, D., Gan, Z., et al.: Mm-ego: Towards building egocentric multimodal llms for video qa. In: International Conference on Learning Representations (ICLR) (2025)

48. Yuan, Y., Dang, R., Li, L., Li, W., Jiao, D., Li, X., Zhao, D., Wang, F., Zhang, W., Xiao, J., et al.: Eoc-bench: Can mllms identify, recall, and forecast objects in an egocentric world? In: Neural Information Processing Systems (NeurIPS) (2025)
49. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
50. Zhan, X., Huang, W., Sun, H., Fu, X., Ma, C., Cao, S., Jia, B., Lin, S., Yin, Z., Bai, L., et al.: Actial: Activate spatial reasoning ability of multimodal large language models. In: Neural Information Processing Systems (NeurIPS) (2025)
51. 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)
52. Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., et al.: Swift: a scalable lightweight infrastructure for fine-tuning. In: AAAI Conference on Artificial Intelligence (AAAI) (2025)
53. Zhou, S., Xiao, J., Li, Q., Li, Y., Yang, X., Guo, D., Wang, M., Chua, T.S., Yao, A.: Egotextvqa: Towards egocentric scene-text aware video question answering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)