From Passive Observer to Active Critic: Reinforcement Learning Elicits Process Reasoning for Robotic Manipulation

Yibin Liu[1,2⋆], Yaxing Lyu[3⋆], Daqi Gao[1⋆], Zhixuan Liang[4], Weiliang Tang[5], Shilong Mu[6], Xiaokang Yang[1], and Yao Mu[1⋆⋆]

[1] Shanghai Jiao Tong University
[2] Northeastern University
[3] Xiamen University Malaysia
[4] The University of Hong Kong
[5] The Chinese University of Hong Kong
[6] Xspark AI
{liuyibin@stumail.neu.edu.cn, muyao@sjtu.edu.cn}

Abstract. Accurate process supervision remains a critical challenge for long-horizon robotic manipulation. A primary bottleneck is that current video MLLMs, trained primarily under a Supervised Fine-Tuning (SFT) paradigm, function as passive "Observers" that recognize ongoing events rather than evaluating the current state relative to the final task goal. In this paper, we introduce **PRIMO R1** (**P**rocess **R**easoning **I**nduced **MO**nitoring), a 7B framework that transforms video MLLMs into active "Critics". We leverage outcome-based Reinforcement Learning to incentivize explicit Chain-of-Thought generation for progress estimation. Furthermore, our architecture constructs a structured temporal input by explicitly anchoring the video sequence between initial and current state images. Supported by the proposed **PRIMO Dataset** and **Benchmark**, extensive experiments across diverse in-domain environments and out-of-domain real-world humanoid scenarios demonstrate that PRIMO R1 achieves state-of-the-art performance. Quantitatively, our 7B model achieves a 50% reduction in the mean absolute error of specialized reasoning baselines, demonstrating significant relative accuracy improvements over 72B-scale general MLLMs. Furthermore, PRIMO R1 exhibits strong zero-shot generalization on difficult failure detection tasks. We establish state-of-the-art performance on the RoboFail benchmark with 67.0% accuracy, surpassing closed-source models like OpenAI o1 6.0%. The project website is: 10-oasis-01.github.io/primo-r1-website.

Keywords: Vision-Language Models · Language Models Reasoning · Embodied AI · Task Progress Estimation

1 Introduction

The pursuit of general-purpose robots capable of performing long-horizon manipulation tasks remains a central challenge in embodied AI. A critical bottleneck in acquiring such skills is to derive effective reward signals. While sparse rewards

⋆ Equal contribution. This work was done during Yibin Liu's and Yaxing Lyu's internships at ScaleLab in Shanghai Jiao Tong University.

⋆⋆ Corresponding author: muyao@sjtu.edu.cn

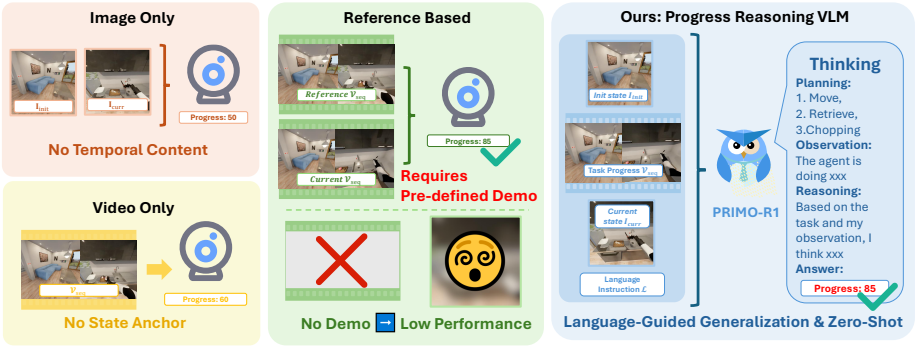

Fig. 1: Paradigm comparison: Prior approaches vs. our PRIMO R1.

(*e.g.*, binary success/failure) are easy to specify, they are often insufficient for efficient policy learning in complex environments. Conversely, dense rewards, which provide granular feedback on task progress, typically rely on laborious manual engineering or privileged access to ground-truth states unavailable in the real world. Recent advances in Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) [20, 21, 36] have sparked hope for learning universal reward functions directly from visual observations. However, deploying these models as reliable "process supervisors" reveals a fundamental limitation in their current paradigm.

Existing video MLLMs, despite their impressive capabilities in captioning and QA, primarily function as passive "Observers". They excel at describing *what* is happening but struggle with the rigorous quantitative reasoning required to judge *how well* the task is proceeding. Most prior approaches treat progress estimation as a standard regression or classification problem via supervised fine-tuning. Within this paradigm, models are optimized to recognize and describe ongoing events, rather than to measure the actual distance between the current state and the final task goal. Consequently, these "Observers" are brittle: they fail to generalize to unseen objects, cannot explain their predictions, and crucially, often assign high progress scores to failed attempts simply because the visual trajectory resembles a successful motion. This exposes a critical structural deficit: without explicit temporal boundary anchoring and continuous reasoning pathways, models are unable to align continuous visual trajectories with the discrete logical conditions required for task success.

To bridge the gap between passive perception and active evaluation, we argue that a reliable reward model must evolve from an Observer into an active **"Critic"**. We introduce **PRIMO R1** (**P**rocess **R**easoning **I**nduced **MO**nitoring), a 7B model framework that elicits explicit process reasoning from video MLLMs. Instead of supervising the model with a single scalar label, we leverage Reinforcement Learning (RL) to incentivize the generation of Chain-of-Thought. Furthermore, to address the loss of detail in continuous dynamic feature spaces, our architecture employs a structural prompting strategy by explicitly anchoring the video sequence between initial and current state images. This design provides clear visual boundary conditions that transform the reasoning task from generic

temporal perception into a structured state-alignment verification. By conditioning this reasoning process on diverse natural language task goals, we establish a structural connection between the objective input and the reasoning execution, effectively exploiting the inherent linguistic generalization capabilities of foundational MLLMs. To support this paradigm, we construct the **PRIMO Dataset**, encompassing both SFT and RL post-training data with CoT annotations, and **PRIMO Benchmark**, designed to systematically evaluate out-of-domain generalization across cross-task and cross-environment settings.

Our experiments reveal that optimizing a policy model for continuous progress reasoning intrinsically constructs the temporal context representations required for discrete failure detection. By enforcing rigorous temporal alignment and self-reasoning, PRIMO R1 achieves state-of-the-art performance across multiple domains. Quantitatively, our 7B model attains an average Mean Relative Accuracy (MRA) of 82.90 and a Mean Absolute Error (MAE) of 15.52, effectively outperforming 72B-scale general MLLMs by a margin of +9.10 absolute MRA points. Furthermore, it exhibits robust zero-shot generalization in execution anomaly verification, achieving 67.0% accuracy on the RoboFail benchmark and surpassing parameter-heavy closed-source models including GPT-4o and OpenAI o1.

Our contributions can be summarized as follows:

- We introduce **PRIMO R1**, a 7B reasoning model that effectively transforms video MLLMs from passive Observers into interpretable Critics. It achieves SOTA performance in task progress estimation and failure detection.
- We present the **PRIMO Dataset** for task progress detection, covering both SFT and RL post-training data with CoT annotations, alongside **PRIMO Bench**, which systematically evaluates the out-of-domain generalization capabilities of post-training methods in video-based MLLMs.
- We propose a structured temporal input strategy that explicitly anchors video sequences between initial and current state images. This boundary anchoring facilitates high-precision state alignment, achieving a 50% reduction in the mean absolute error compared to specialized baselines.
- We demonstrate that optimizing for progress reasoning intrinsically enables robust zero-shot generalization for failure detection. This is validated on the RoboFail benchmark, where PRIMO R1 achieves a state-of-the-art 67.0% accuracy, surpassing closed-source models like OpenAI o1 by 6.0%.

2 Related Work

2.1 Multimodal Large Language Models for Video Understanding

Early video MLLMs adapted static architectures via temporal aggregation [16, 22] and mitigated memory bottlenecks through context compression and hierarchical structures [6, 15, 25, 29, 39]. These architectures operate predominantly as passive "Observers," excelling at perceptual QA but lacking quantitative temporal reasoning. The structural transition toward an active "Critic" paradigm necessitates explicit temporal localization, prompting recent designs to integrate

evidence searching and timestamp encoding [11,24,35]. Our framework completes this transition by enabling progress reasoning for rigorous temporal judgment.

2.2 Vision-Based State Estimation and Reward Modeling

Semantic reward modeling leverages VLMs to encode universal value functions via representation distances [20,21] or frozen embeddings [30]. For explicit progress estimation, recent mechanisms employ frame-ordering [19], multi-modal integration [36, 38], and synthetic trajectory augmentation [37]. A primary structural limitation of VLAC [36], Robo-Dopamine [31], and PROGRESSLM [38] is their functional dependency on explicit reference demonstrations. Furthermore, framing estimation as direct regression via Supervised Fine-Tuning (SFT) restricts the model's capacity for causal failure analysis. Our method bypasses reference dependency by explicitly eliciting process reasoning chains.

2.3 Reinforcement Learning for Reasoning Elicitation

Inference-time scaling and outcome-based Reinforcement Learning (RL) induce Chain-of-Thought (CoT) behaviors without dense annotations [10]. This "R1 paradigm" has expanded into multimodal domains, enhancing static visual reasoning [28, 34] and dynamic temporal grounding [8, 32]. Parallel architectural optimizations include dynamic frame sampling [9] and current-state image anchoring for planning [5]. We map this capability to robotic process supervision, formulating task completion metrics as outcome rewards to elicit verifiable, self-correcting reasoning paths for task assessment.

3 Method

3.1 Problem Formulation

We formalize the task of robotic process supervision as a multi-modal state estimation problem. The input space consists of four key variables: an initial state image I_{init} capturing the environment prior to execution, a process video sequence $V_{seq} = \{v_1, v_2, \ldots, v_T\}$ representing the temporal state transitions, a current state image I_{curr} reflecting the latest observed outcome, and a language instruction $\mathcal{I}$ specifying the task goal (the specific structure and content of $\mathcal{I}$ are detailed in Figure 10). The objective is to learn a mapping function F that evaluates the visual tuple $(I_{init}, V_{seq}, I_{curr})$ conditioned on the instruction $\mathcal{I}$ as the semantic reference, outputting a scalar progress indicator $y \in [0, 100]$, where $y = 0$ denotes the initial state and $y = 100$ signifies successful state. As demonstrated in our ablation study (Table 4), explicitly modeling both boundary states (I_{init} and I_{curr}) alongside the temporal transition (V_{seq}) is a necessary structural prerequisite for accurate progress estimation across varying task horizons.

In the standard paradigm, existing video MLLMs function as passive "Observers", treating progress estimation as a direct regression or classification task.

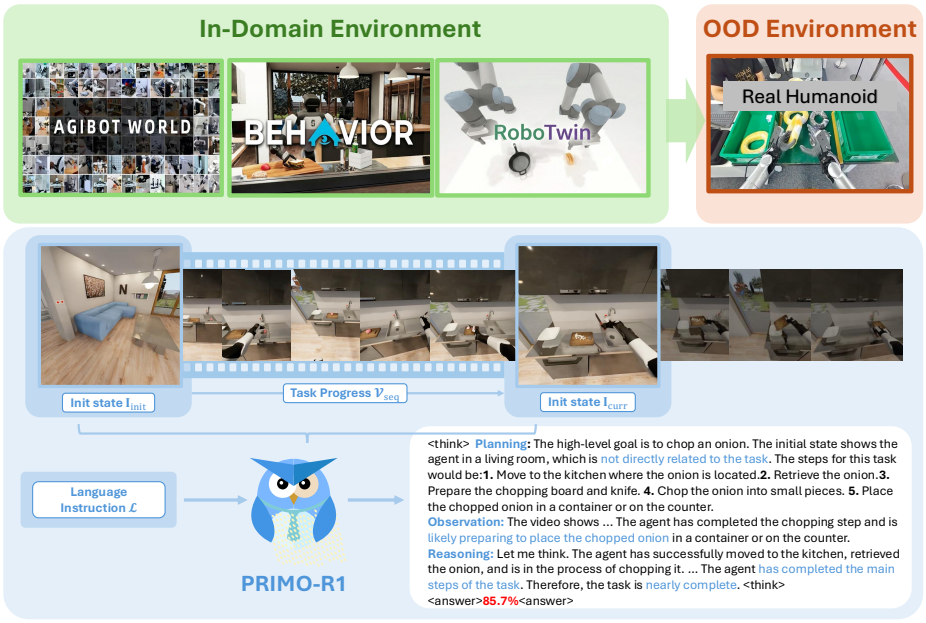

Fig. 2: Overall framework of PRIMO R1. Evaluated across in-domain simulations (AgiBot, BEHAVIOR, RoboTwin) and OOD real humanoid environments, the model processes a video sequence (V_{seq}) anchored by initial (I_{init}) and current (I_{curr}) states. It generates an explicit Chain-of-Thought to output the final progress estimate.

They directly model the distribution of the target y_{gt} conditioned on the input tuple $(I_{init}, V_{seq}, I_{curr}, \mathcal{I})$ via Supervised Fine-Tuning (SFT). This formulation isolates visual features at a surface level, bypassing the underlying causal structure of the state transitions.

To transform the model into an active "Critic", we reformulate the prediction process from direct scalar regression into a multi-step generative reasoning task. We define a policy π_θ that sequentially generates a latent reasoning chain (Chain-of-Thought) $\mathcal{C}$, followed by the final progress estimate $\hat{y}$. Rather than relying on dense annotations to supervise the intermediate variable $\mathcal{C}$, we optimize π_θ using Reinforcement Learning. The optimization objective maximizes the expected reward $R(\hat{y}, y_{gt})$, which is computed solely based on the accuracy of the final prediction $\hat{y}$. This structural dependency incentivizes the policy to self-organize the intermediate reasoning $\mathcal{C}$ to accurately align the temporal transition V_{seq} between the boundary states I_{init} and I_{curr}. Crucially, conditioning this generative reasoning process on diverse natural language task goals ($\mathcal{I}$) establishes a direct structural mapping between the semantic objective and the visual execution logic, explicitly exploiting the linguistic generalization capabilities of foundational MLLMs to process varying evaluation criteria. The complete architectural workflow of this framework is illustrated in Fig. 2.

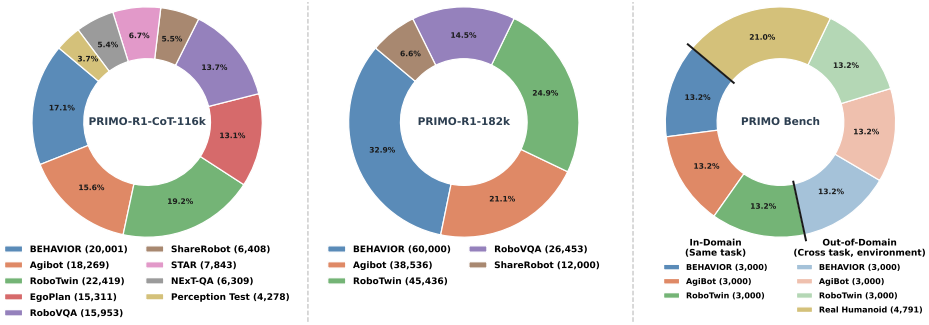

Fig. 3: Dataset distribution for SFT (left), RL (middle), and PRIMO Bench (right). Charts show sample counts and domain percentages (e.g., BEHAVIOR, AgiBot, RoboTwin). The PRIMO Bench highlights the data split between In-Domain and Out-of-Domain evaluation sets. See Appendix A for details.

3.2 PRIMO Dataset and Benchmark

To systematically elicit and evaluate the temporal reasoning capabilities of Video MLLMs for robotic process supervision, we present the **PRIMO Dataset** and the accompanying **PRIMO Bench**.

PRIMO Dataset for Post-Training. The PRIMO Dataset is meticulously constructed to support our two-stage post-training paradigm, covering both Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) data. Unlike standard video QA datasets, our data features fine-grained progress indicators annotated with Chain-of-Thought reasoning paths. The training corpus aggregates multi-source trajectories from a real-world environment (AgiBot) and two high-fidelity simulations (BEHAVIOR-1k and RoboTwin). Additionally, to maintain data diversity during the SFT phase, we incorporate several general video reasoning datasets to augment the training corpus, yielding a comprehensive collection partitioned into a 116k-sample SFT dataset (PRIMO-R1-CoT-116k) and a 182k-sample RL dataset (PRIMO-R1-182k), as illustrated in Figure 3.

PRIMO Bench for Generalization Evaluation. To systematically evaluate the robustness of post-training methods against varying degrees of distribution shift, we introduce **PRIMO Bench**, which categorizes evaluation into two splits:

- **In-Domain (ID) - Same Task:** Evaluates the model's estimation accuracy on task categories that were exposed during the training phase within the three seen environments.
- **Out-of-Domain (OOD) - Cross-Task & Cross-Environment:** Designed to test zero-shot generalization. *Cross-Task* evaluates the model on entirely unseen tasks within the familiar environments. *Cross-Environment* introduces a stringent unseen domain transfer challenge, evaluating the model on real-world trajectories collected via teleoperation of a different physical humanoid robot (Leju KUAVO-MY) in unstructured physical environments (e.g., factories and service scenarios).

Comprehensive details regarding raw data collection, semantic annotation synthesis, data processing methodologies, and exact dataset statistics of our dataset and benchmark are provided in Appendix A.

3.3 Process Reasoning RL with Group Relative Policy Optimization

To transform the Video MLLM from a passive observer into an active critic capable of self-correction, we employ Group Relative Policy Optimization (GRPO) [10]. Unlike standard Proximal Policy Optimization (PPO) [26], which relies on a computationally expensive value function critic to estimate the baseline, GRPO leverages the group statistics of sampled outputs to estimate the baseline. This is particularly advantageous for Video MLLMs, where the memory overhead of maintaining a separate value network alongside the policy model is prohibitive.

Group Sampling and Advantage Estimation. Formally, for a given task tuple $(I_{init}, V_{seq}, I_{curr}, \mathcal{I})$, we sample a group of G outputs $\{o_1, o_2, \ldots, o_G\}$ from the policy $\pi_{\theta_{old}}$. Each output o_i consists of a reasoning chain $\mathcal{C}_i$ (enclosed in `<think>` tags) and a final progress estimate $\hat{y}_i$. Instead of training a value function $V(x)$, GRPO computes the advantage A_i for each output o_i by normalizing its reward r_i against the group's distribution:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\}) + \epsilon}, \tag{1}$$

where ϵ is a small constant for numerical stability. This relative advantage encourages the model to generate reasoning paths that yield higher rewards than the average of its current stochastic explorations, effectively filtering out "hallucinated" progress estimates.

Rule-Based Reward Design. A core challenge in eliciting reasoning is defining an effective reward signal without dense annotation. We define a composite reward function $R(o_i, y_{gt}) = r_{\text{fmt}} + r_{\text{acc}}$ targeting both structure and precision:

(1) Format Reward (r_{fmt}). To explicitly induce a Chain-of-Thought, we enforce a strict structural constraint. The model receives a positive reward (e.g., $+1$) only if the output strictly follows the pattern `<think>reasoning</think>` followed by `<answer>prediction</answer>`. This prevents the policy from collapsing into direct guessing.

(2) Accuracy Reward (r_{acc}). Since the task progress y is continuous, treating it as a binary outcome is insufficient. To provide dense feedback for numerical reasoning, we adopt a *bounded linear decay* reward function:

$$r_{\text{acc}} = \max\left(0, 1 - \frac{|\hat{y}_i - y_{gt}|}{R_{\max}}\right), \tag{2}$$

where $R_{\max}$ (e.g., 100.0) represents the maximum error range. This formulation ensures the reward starts at 1.0 for an exact match and linearly decreases to 0.0 as the error approaches $R_{\max}$, strictly confining the score to the $[0, 1]$ interval.

Optimization Objective. The policy π_θ is updated to maximize the expected advantage while remaining close to the reference policy π_{ref} to prevent reward hacking or language degeneration. The GRPO objective is formulated as:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\left[\min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i\right)\right. \\ \left. -\beta \cdot \mathbb{D}_{\text{KL}}\left(\pi_\theta(o_i|x)||\pi_{\text{ref}}(o_i|x)\right)\right], \tag{3}$$

where $\rho_i = \frac{\pi_\theta(o_i|x)}{\pi_{\theta_{old}}(o_i|x)}$ is the probability ratio, and β controls the strength of the KL divergence penalty. By optimizing this objective, PRIMO R1 implicitly learns that generating detailed, causal reasoning in $\mathcal{C}$ is the most reliable strategy to maximize the accuracy reward in $\hat{y}$, thereby emerging as a robust Critic.

4 Experiments

In this section, we systematically evaluate the performance of PRIMO R1. The experiments are structured to assess two primary capabilities: continuous task progress estimation across both in-domain simulations and out-of-domain real-world environments, and zero-shot generalization in discrete execution failure detection. Furthermore, we conduct ablation studies to isolate the impact of temporal context modalities on estimation accuracy, followed by qualitative case studies analyzing the structural logic of the generated reasoning chains. Throughout our evaluations, we employ Qwen2.5-VL-7B-Instruct as the foundation model for all training phases. Detailed experimental setups, including hardware specifications, training parameters, and inference configurations like sampled frame count and frame resolution, are provided in Appendix H.

4.1 Evaluation Metrics

We evaluate task progress estimation using Mean Relative Accuracy (MRA) and Mean Absolute Error (MAE).

Mean Relative Accuracy (MRA). Given a prediction $\hat{y}$, ground-truth progress y, and a set of accuracy thresholds $\mathcal{T}$, Mean Relative Accuracy (MRA) is defined as

$$\text{MRA} = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}}\mathbb{I}\left(\frac{|\hat{y}-y|}{|y|} < 1-\tau\right), \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

Mean Absolute Error (MAE). MAE is defined as

$$\text{MAE} = \mathbb{E}\big[|\hat{y}-y|\big], \tag{5}$$

and is reported to provide a clear measure of absolute prediction error.

4.2 Main Results: Generalization in Progress Estimation

We present the evaluation of our proposed method against state-of-the-art baselines. The results are analyzed in two parts: a comprehensive performance comparison across all domains (Table 1) and an ablation study focusing on the impact of SFT and RL strategies on generalization (Table 2).

Table 1: Comparison on Progress Estimation. We report the Mean Relative Accuracy (MRA ↑, higher is better) and Mean Absolute Error (MAE ↓, lower is better) across four distinct environments. The best results are highlighted in **bold**. Specialized progress and reward models are evaluated under their own recommended input configurations and are therefore not strictly comparable cell-for-cell with models run under our unified protocol; the ProgressLM-3B-RL row in the Reasoning & Video group is the same model re-evaluated under our unified setting.

Model	AgiBot		Behavior		RoboTwin		Real Humanoid		Average	
	MRA (↑)	MAE (↓)	MRA (↑)	MAE (↓)	MRA (↑)	MAE (↓)	MRA (↑)	MAE (↓)	MRA (↑)	MAE (↓)
Closed-Source Models										
GPT-5 mini	74.27	24.81	79.60	20.08	80.52	18.34	67.14	32.59	75.38	23.96
GPT-4o	81.01	18.99	79.92	20.08	81.73	18.27	74.65	25.35	79.33	20.67
Gemini 2.5 Flash	73.54	26.41	78.01	21.99	81.04	18.58	67.37	32.63	74.99	24.90
Claude-Haiku-4.5	74.40	25.59	70.93	29.07	74.13	25.87	72.68	27.32	73.04	26.96
Open-Source General MLLMs										
Qwen2.5-VL-7B	77.43	22.56	69.91	30.06	67.37	32.62	56.46	34.73	67.79	29.99
InternVL 3.5 8B	78.52	21.47	72.19	27.15	70.81	29.18	65.44	34.55	71.74	28.09
Qwen2.5-VL-72B	78.00	22.00	79.49	20.50	75.41	24.59	62.29	28.10	73.80	23.80
Reasoning & Video MLLMs										
ProgressLM-3B-RL	42.90	30.83	46.43	28.23	31.84	36.48	32.00	34.90	38.29	32.61
Video R1 7B	72.42	27.58	70.63	29.37	70.86	29.14	53.57	31.87	66.87	29.49
Robobrain 7B	72.99	25.91	72.52	26.97	70.41	28.85	55.83	28.51	67.94	27.56
Cosmos-Reasoning 7B	72.48	27.01	67.06	32.35	73.14	25.85	59.39	31.41	66.52	29.12
Specialized Progress & Reward Models (own input configuration)										
VLAC	84.92	15.08	76.86	23.14	67.04	32.96	70.77	29.23	74.90	25.10
GVL (Gemini 2.5 Flash)	56.81	35.84	58.11	32.31	61.18	27.91	48.77	38.37	56.22	33.61
Robo-Dopamine-3B	81.18	18.82	57.78	42.22	82.96	17.04	68.43	31.57	72.59	27.41
Robo-Dopamine-4B	83.88	16.12	65.30	34.70	67.22	32.78	73.89	26.11	72.57	27.43
ProgressLM	81.13	18.79	80.89	15.97	67.03	32.97	**84.24**	**15.76**	78.32	20.87
PRIMO R1 (Ours)	**87.67**	**12.33**	**87.08**	**12.90**	**84.52**	**15.48**	72.32	21.37	**82.90**	**15.52**

Overall Performance. Table 1 summarizes the performance of all models across the four evaluation environments: AgiBot, Behavior, RoboTwin, and Real Humanoid. Using the metrics defined in Sec. 4.1, PRIMO R1 achieves the highest average MRA (82.90) and the lowest average MAE (15.52), consistently outperforming all evaluated open-source baselines.

Compared with general open-source MLLMs, PRIMO demonstrates a clear advantage. Despite using only a 7B backbone, it surpasses Qwen2.5-VL-72B by 9.10 MRA points (82.90 vs. 73.80). Against specialized reasoning and video MLLMs such as Video R1 7B and Robobrain 7B, PRIMO reduces the average MAE from approximately 27–29 to 15.52, nearly halving the estimation error.

We further compare PRIMO R1 with specialized progress estimation models, including VLAC, GVL, Robo-Dopamine, and ProgressLM, each evaluated using its recommended input configuration. PRIMO R1 still achieves the best overall performance, improving the average MRA from 78.32 to 82.90 while reducing the average MAE from 20.87 to 15.52. It consistently leads on AgiBot, Behavior, and RoboTwin across both metrics. The only exception is the Real Humanoid setting, where ProgressLM achieves higher MRA (84.24 vs. 72.32) and lower MAE (15.76 vs. 21.37), likely benefiting from its dedicated real-world input configuration. In contrast, PRIMO R1 adopts a unified input protocol across all four environments yet still achieves the best average performance, suggesting

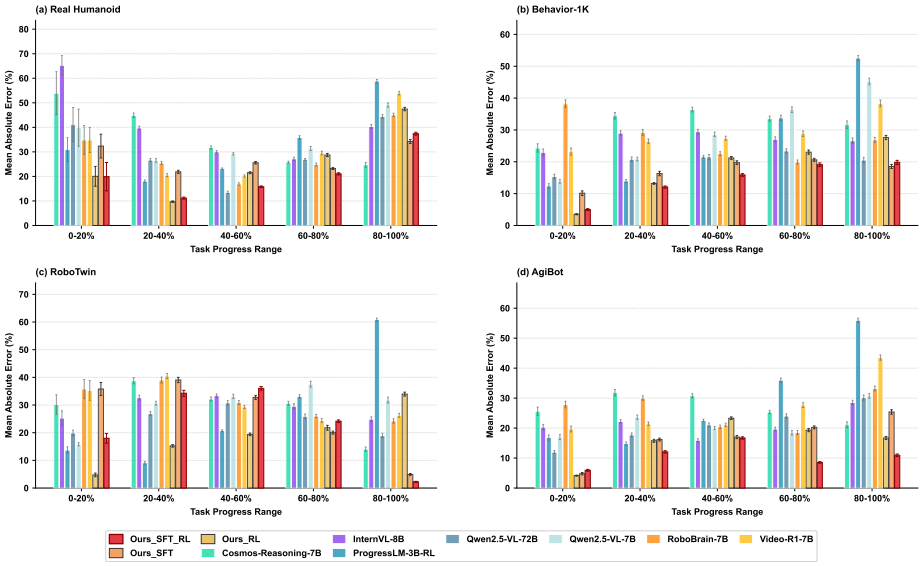

Fig. 4: Fine-Grained Error Analysis Across Task Progress Intervals. MAE evaluation across five completion stages in four environments ((a)-(d)). Compared to baselines, our RL-finetuned model (Ours_SFT_RL) maintains lower error rates, particularly mitigating severe hallucinations in the final execution stage (80−100%).

that explicit process reasoning generalizes more reliably across heterogeneous domains than configuration-specific designs.

Figure 4 further analyzes MAE across five task completion intervals. While baseline models exhibit pronounced error spikes during the final execution stage ($80-100%$), PRIMO maintains consistently low errors throughout the trajectory. This indicates that explicit process reasoning effectively mitigates premature completion hallucinations caused by superficial visual similarities near the end of a task. Notably, in the unseen and highly unstructured Real Humanoid environment, PRIMO achieves an MRA of 72.32, substantially outperforming general MLLMs (e.g., Qwen2.5-VL-7B: 56.46). This result highlights the effectiveness of generating an explicit reasoning chain before prediction for improving Sim-to-Real generalization.

Impact of RL on Generalization. To validate the effectiveness of our training pipeline, we analyze the performance evolution from the Base model to the SFT stage, and finally to the RL-finetuned stage. Table 2 details the performance on In-Domain (ID) seen tasks versus Out-of-Distribution (OOD) unseen tasks.

The Base Qwen2.5-VL-7B model exhibits a weak zero-shot capability for precise progress estimation (average MRA of 67.46). While Supervised Fine-Tuning (SFT) significantly improves overall performance to 79.35, it primarily overfits to the semantic features of the training distribution. This behavior is consistent with the observation that SFT tends to memorize training formats rather than acquire transferable reasoning, whereas RL generalizes to out-of-

Table 2: Ablation and Generalization Analysis. All reported metrics represent Mean Relative Accuracy (MRA ↑). We compare the Base model, SFT-only model, RL-only model, and our final model (SFT + RL). The results are split into In-Domain (ID) tasks and Out-of-Domain (OOD) tasks to highlight generalization capabilities.

Model	In-Domain (ID)			Out-of-Domain (OOD)				Avg.
	Seen Tasks			*Cross-Task*			*Cross-Environment*	
	Agibot	Behavior	RoboTwin	Agibot	Behavior	RoboTwin	Real Humanoid	
Qwen2.5-VL-7B (Base)	70.83	69.13	71.19	74.45	77.47	61.01	48.12	67.46
Our Model (SFT)	83.37	80.38	80.63	82.02	82.61	79.13	67.30	79.35
Our Model (RL)	86.05	85.82	73.27	82.95	81.39	75.43	52.12	76.72
PRIMO R1 (SFT+RL)	**87.83**	**89.42**	**88.15**	**87.67**	**87.08**	**84.52**	**72.32**	**85.28**

Table 3: Failure Detection Capabilities. Accuracy (%) on the RoboFail benchmark [18]. The evaluation measures the capability of models to effectively detect and quantify task execution failures. Benchmark details are provided in Appendix A.

Model	RoboFail (↑)	Model	RoboFail (↑)	Model	RoboFail (↑)
Closed-Source		*Open-Source*		*Ours*	
Gemini 2.0 Flash	**67.0**	Qwen2.5-VL-7B	57.6	PRIMO (SFT)	51.0
GPT-4o	63.0	Nemotron-H-56B	64.0	PRIMO (RL)	63.0
OpenAI o1	61.0	Cosmos-Reason1-7B	60.0	**PRIMO R1**	**67.0**
Claude-haiku-4.5	59.0	Cosmos-Reason1-56B	66.2		

distribution inputs [7]. The effect is evident in the Cross-Environment (Real Humanoid) OOD setting, where the SFT model only achieves 67.30.

Interestingly, applying RL directly without SFT (RL-only) yields suboptimal results (76.72 average MRA), as the model struggles to autonomously discover the correct output format and structural reasoning paths from scratch. However, the integration of Group Relative Policy Optimization (GRPO) after the SFT phase enables our complete PRIMO (SFT+RL) pipeline to create a powerful synergy. The RL phase pushes ID performance to near 90% (e.g., 89.42% on Behavior) and, more importantly, drastically enhances generalization capabilities. The self-correction and rigorous causal reasoning learned via RL transfer effectively to OOD settings, boosting Cross-Task performance across all simulated environments and lifting the Cross-Environment accuracy to 72.32%. This confirms that RL process supervision fundamentally shifts the model from a passive pattern-matcher to an active, generalizing critic.

4.3 Generalization Enhancement in Failure Detection

For a Vision-Language Model engaged in process supervision, tracking continuous task progress and detecting discrete execution failures represent coupled dimensions of temporal reasoning. The capability to identify physical constraints or execution errors structurally depends on an underlying representation of intended state transitions. To evaluate the zero-shot generalization of this capability, we test our model on the RoboFail benchmark (details in Appendix A), a completely unseen dataset designed to evaluate "action affordance" and "task completion verification" under complex physical constraints.

Table 4: Ablation on Input Modalities. We analyze the necessity of temporal context by varying the input information. I_{init}: Initial state image. V_{seq}: Process video clip. I_{curr}: Current state image. Results show that temporal context (V_{seq}) is crucial for reducing error (MAE).

Input Modality			**Agibot**		**Behavior**		**Robotwin**		**Avg**	
I_{init}	V_{seq}	I_{curr}	MAE ↓	Acc@10 ↑	MAE ↓	Acc@10 ↑	MAE ↓	Acc@10 ↑	MAE ↓	Acc@10 ↑
		✓	59.64	0.00	51.91	8.17	66.94	0.77	59.50	2.98
✓		✓	43.97	18.52	49.59	9.73	45.93	11.45	46.50	13.23
	✓		27.58	31.34	34.85	18.33	47.41	8.07	36.61	19.25
	✓	✓	25.04	**35.29**	27.59	29.21	**40.24**	17.37	30.96	27.29
✓	✓		**24.94**	33.98	32.55	23.01	45.37	11.97	**34.29**	22.99
✓	✓	✓	29.39	27.15	**22.73**	**31.83**	42.16	**27.69**	31.43	**28.89**

Table 3 details the quantitative performance across different model architectures. The base Qwen2.5-VL-7B model exhibits a baseline accuracy of 57.6%. Applying Supervised Fine-Tuning (SFT) alone results in a performance regression to 51.0%. We attribute this drop to format overfitting rather than a limitation of the data: because next-token prediction weights all tokens equally, SFT tends to pattern-match the chain-of-thought structure of the training distribution instead of reasoning over unseen inputs, and prior studies report the same effect, where SFT memorizes formats while RL generalizes [7,8]. The integration of Process Supervision RL with GRPO corrects this degradation, elevating the accuracy to 63.0%. By optimizing against outcome-based rewards rather than fixed token sequences, RL recovers genuine reasoning and transfers it to this unseen benchmark. The final PRIMO R1 formulation achieves an accuracy of 67.0%, matching the closed-source Gemini 2.0 Flash and outperforming larger parameter models, including GPT-4o (63.0%), OpenAI o1 (61.0%), and Cosmos-Reason1-56B (66.2%).

Prior benchmark analyses, such as those conducted for Cosmos-Reason1, indicate that standard Reinforcement Learning targeting physical AI yields limited improvements on RoboFail. The core difficulty of the benchmark stems from the prerequisite for highly observant perception and comprehensive temporal context processing, which operate as distinct variables alongside static physical common sense. The performance delta between Cosmos-Reason1-7B (60.0%) and PRIMO R1 (67.0%) establishes a specific functional relationship: optimizing a policy model for continuous progress reasoning explicitly constructs the temporal context representations necessary for failure verification. The capacity for embodied error correction structurally necessitates process reasoning capability as a parallel condition to physical common sense.

4.4 Ablation Study: The Necessity of Temporal Context

To isolate the impact of temporal context and state representations on progress estimation, we conduct an ablation study over three input variables: the initial

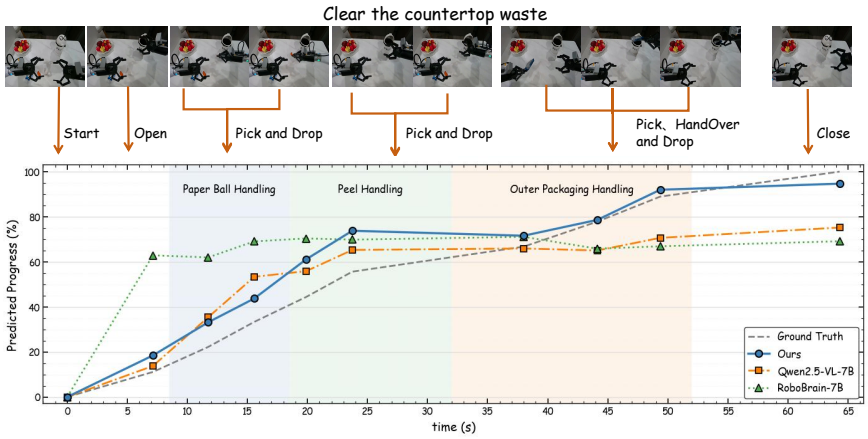

Fig. 5: Continuous Progress Estimation. Average predicted progress trajectory over 105 episodes for the "Clear the countertop waste" task, comparing temporal state alignment against baselines.

state image I_{init}, the process video sequence V_{seq}, and the current state image I_{curr}. The quantitative results are summarized in Table 4.

Using only the current state I_{curr} yields the highest estimation error, with an Average MAE of 59.50, indicating that a static snapshot lacks sufficient causal context for accurate progress estimation. Replacing it with the temporal sequence V_{seq} reduces the Average MAE to 36.61, but performance remains limited across most tasks, suggesting that temporal information alone is insufficient without explicit reference states.

Prior work [4,5] has shown the benefits of incorporating I_{curr} in video-centric planning. Consistent with this observation, adding either I_{init} or I_{curr} to V_{seq} further improves estimation performance. Specifically, $V_{seq} + I_{curr}$ achieves the lowest MAE on RoboTwin (40.24), while $I_{init} + V_{seq}$ performs best on AgiBot (24.94). Our final PRIMO R1 adopts all three modalities. Although dual-modality combinations perform competitively on shorter-horizon datasets such as AgiBot, incorporating both boundary states yields a clear advantage on the long-horizon Behavior dataset, reducing MAE to 22.73 and improving Acc@10 to 31.83. Conversely, the boundary-state paradigm ($I_{init} + I_{curr}$) used in prior work, such as SCIZOR [40] suffers without video input, increasing the Behavior MAE from 22.73 to 49.59. While the gap is smaller on RoboTwin (42.16 vs. 45.93), the substantial degradation on long-horizon tasks confirms that explicit video modeling is equally essential.

These results highlight a fundamental representation trade-off. Video features effectively capture temporal dynamics but sacrifice fine-grained spatial information needed for accurate state comparison. Explicitly incorporating both I_{init} and I_{curr} anchors the temporal evolution encoded by V_{seq}, providing complementary spatial references for robust progress estimation.

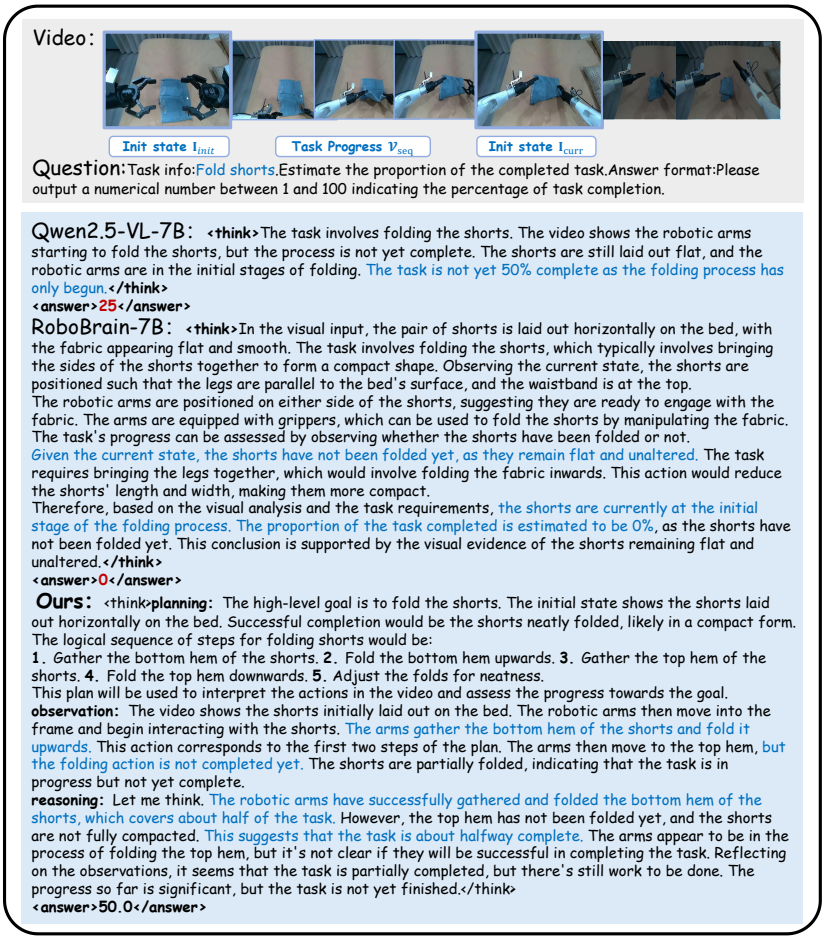

Fig. 6: Qualitative Comparison of Reasoning Processes. This case study illustrates the generated reasoning chains of Qwen2.5-VL-7B, RoboBrain-7B and our PRIMO R1 during the Fold shorts task in AgiBot environment.

4.5 Case Study

To evaluate continuous tracking capabilities in long-horizon scenarios, Figure 5 plots the predicted progress trajectory for the composite task "Clear the countertop waste", mapping the average outputs across 105 episodes. The evaluation isolates the relationship between the predicted progress variable and the ground-truth temporal execution sequence. The baseline RoboBrain-7B demonstrates a decoupling from actual physical states; its prediction overshoots during the initial 0-10s phase and subsequently loses sensitivity to further temporal advancement. Qwen2.5-VL-7B tracks the initial sub-stages accurately, but its prediction variable plateaus near 60%-70% during the latter half, failing to map visual convergence to the final task state. Our model maintains a monotonically increasing trajectory that correlates linearly with the ground truth across discrete sub-task

transitions. In the terminal phase, it accurately maps the visual state change of the final action to a progress metric approaching 100%, verifying a stable alignment between long-range temporal sequences and progress estimation.

Figure 6 details the structural decomposition of the explicit reasoning chain generated by the model for a "Fold shorts" video. Baselines struggle with fine-grained state tracking: RoboBrain-7B overlooks ongoing dynamic manipulations (0% progress), while Qwen2.5-VL-7B lacks a structured evaluation metric (25% progress). Conversely, our PRIMO R1 generates an explicit reasoning chain via three modules. The *Planning* module establishes a reference topology by breaking down the high-level semantic goal into a linear five-step execution plan (Gather bottom hem $\rightarrow$ Fold upwards $\rightarrow$ Gather top hem $\rightarrow$ Fold downwards $\rightarrow$ Adjust). The *Observation* module discretizes the continuous visual input, extracting specific dynamic variables and verifying part-level object state changes (e.g., isolating the state of the bottom hem from the top hem). Finally, the *Reasoning* module executes state alignment by mapping the extracted visual primitives against the planned execution topology. It identifies precise execution boundaries. Specifically, confirming the successful manipulation of the bottom hem while explicitly verifying the incomplete status of the top hem, which acts as a structural constraint for quantitative evaluation. The final numerical prediction (50.0%) is formulated by calculating the ratio of the verified execution steps against the complete reference plan. Since Inference latency and real-time performance are critical for robotic manipulation, we also provide an analysis and comparison of reasoning chain lengths and inference times in Appendix C.2.

5 Conclusion

In this work, we introduced PRIMO R1, a 7B framework that transforms video MLLMs into active critics for robotic process supervision via outcome-based reinforcement learning (GRPO). By explicitly anchoring temporal sequences between initial and current state images and incentivizing Chain-of-Thought generation, our approach mitigates spatial detail dilution and enables rigorous temporal reasoning. Furthermore, conditioning this reasoning process on diverse natural language task goals explicitly exploits the language generalization capabilities of foundational LLMs. Experimental results across simulation and real-world humanoid domains demonstrate that PRIMO R1 achieves state-of-the-art performance, empirically establishing that optimizing continuous progress tracking intrinsically constructs the prerequisite representations for zero-shot discrete failure detection, suggesting a pathway toward deriving reward signals essential for future autonomous policy learning in long-horizon manipulation.

Acknowledgements

This work was supported in part by the Shanghai Magnolia Talent Program Pujiang Project under Grant No. 25PJA076.

References

1. Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., et al.: Cosmos-reason1: From physical common sense to embodied reasoning. arXiv preprint arXiv:2503.15558 (2025)
2. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669 (2025)
3. Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Li, Z., Liang, Q., Lin, X., Ge, Y., Gu, Z., et al.: Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. arXiv preprint arXiv:2506.18088 (2025)
4. 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 **134**(3), 118 (2026)
5. Chen, Y., Ge, Y., Wang, R., Ge, Y., Cheng, J., Shan, Y., Liu, X.: Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. arXiv preprint arXiv:2506.16141 (2025)
6. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
7. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)
8. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
9. Ge, H., Wang, Y., Chang, K.W., Wu, H., Cai, Y.: Framemind: Frame-interleaved video reasoning via reinforcement learning. arXiv preprint arXiv:2509.24008 (2025)
10. 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)
11. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)
12. Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. arXiv preprint arXiv:2502.21257 (2025)
13. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Lingelbach, M., Sun, J., et al.: Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In: Conference on Robot Learning. pp. 80–93. PMLR (2023)
14. Li, Y., Wang, L., Wang, T., Yang, X., Luo, J., Wang, Q., Deng, Y., Wang, W., Sun, X., Li, H., et al.: Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. IEEE Trans. Pattern Anal. Mach. Intell. **47**(3), 1832–1849 (2025)
15. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)

16. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
17. Liu, Y., Liang, Z., Chen, Z., Chen, T., Hu, M., Dong, W., Xu, C., Han, Z., Qin, Y., Mu, Y.: Hycodepolicy: Hybrid language controllers for multimodal monitoring and decision in embodied agents. arXiv preprint arXiv:2508.02629 (2025)
18. Liu, Z., Bahety, A., Song, S.: Reflect: Summarizing robot experiences for failure explanation and correction. arXiv preprint arXiv:2306.15724 (2023)
19. Ma, Y.J., Hejna, J., Fu, C., Shah, D., Liang, J., Xu, Z., Kirmani, S., Xu, P., Driess, D., Xiao, T., et al.: Vision language models are in-context value learners. In: The Thirteenth International Conference on Learning Representations (2024)
20. Ma, Y.J., Kumar, V., Zhang, A., Bastani, O., Jayaraman, D.: Liv: Language-image representations and rewards for robotic control. In: International Conference on Machine Learning. pp. 23301–23320. PMLR (2023)
21. Ma, Y.J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., Zhang, A.: Vip: Towards universal visual reward and representation via value-implicit pre-training. arXiv preprint arXiv:2210.00030 (2022)
22. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
23. Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. Advances in Neural Information Processing Systems **36**, 42748–42761 (2023)
24. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323 (2024)
25. Ren, W., Yang, H., Min, J., Wei, C., Chen, W.: Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3804–3814 (2025)
26. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
27. Sermanet, P., Ding, T., Zhao, J., Xia, F., Dwibedi, D., Gopalakrishnan, K., Chan, C., Dulac-Arnold, G., Maddineni, S., Joshi, N.J., et al.: Robovqa: Multimodal long-horizon reasoning for robotics. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 645–652. IEEE (2024)
28. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
29. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18221–18232 (2024)
30. Sontakke, S., Zhang, J., Arnold, S., Pertsch, K., Bıyık, E., Sadigh, D., Finn, C., Itti, L.: Roboclip: One demonstration is enough to learn robot policies. Advances in Neural Information Processing Systems **36**, 55681–55693 (2023)

31. Tan, H., Chen, S., Xu, Y., Wang, Z., Ji, Y., Chi, C., Lyu, Y., Zhao, Z., Chen, X., Co, P., et al.: Robo-dopamine: General process reward modeling for high-precision robotic manipulation. arXiv preprint arXiv:2512.23703 (2025)
32. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025)
33. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
34. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)
35. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-chained image-language model for video localization and question answering. Advances in Neural Information Processing Systems **36**, 76749–76771 (2023)
36. Zhai, S., Zhang, Q., Zhang, T., Huang, F., Zhang, H., Zhou, M., Zhang, S., Liu, L., Lin, S., Pang, J.: A vision-language-action-critic model for robotic real-world reinforcement learning. arXiv preprint arXiv:2509.15937 (2025)
37. Zhang, J., Luo, Y., Anwar, A., Sontakke, S.A., Lim, J.J., Thomason, J., Biyik, E., Zhang, J.: Rewind: Language-guided rewards teach robot policies without new demonstrations. arXiv preprint arXiv:2505.10911 (2025)
38. Zhang, J., Qian, C., Sun, H., Lu, H., Wang, D., Xue, L., Liu, H.: Progresslm: Towards progress reasoning in vision-language models. arXiv preprint arXiv:2601.15224 (2026)
39. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024)
40. Zhang, Y., Xie, Y., Liu, H., Shah, R., Wan, M., Fan, L., Zhu, Y.: Scizor: Self-supervised data curation for large-scale imitation learning. In: IEEE International Conference on Robotics and Automation (ICRA) (2026)