

# Multimodal Reinforcement Learning with Agentic Verifier for AI Agents

Reuben Tan<sup>1</sup> Baolin Peng<sup>1</sup> Zhengyuan Yang<sup>1</sup> Hao Cheng<sup>1</sup> Oier Mees<sup>1</sup> Theodore Zhao<sup>1</sup>  
 Andrea Tupini<sup>1</sup> Isar Meijier<sup>1</sup> Qianhui Wu<sup>1</sup> Yuncong Yang<sup>2</sup> Lars Liden<sup>1</sup> Yu Gu<sup>1</sup> Sheng Zhang<sup>1</sup>  
 Xiaodong Liu<sup>1</sup> Lijuan Wang<sup>1</sup> Marc Pollefeys<sup>1,3</sup> Yong Jae Lee<sup>4</sup> Jianfeng Gao<sup>1</sup>  
<sup>1</sup>Microsoft Research, <sup>2</sup>UMass Amherst, <sup>3</sup>ETH Zurich, <sup>4</sup>UW-Madison

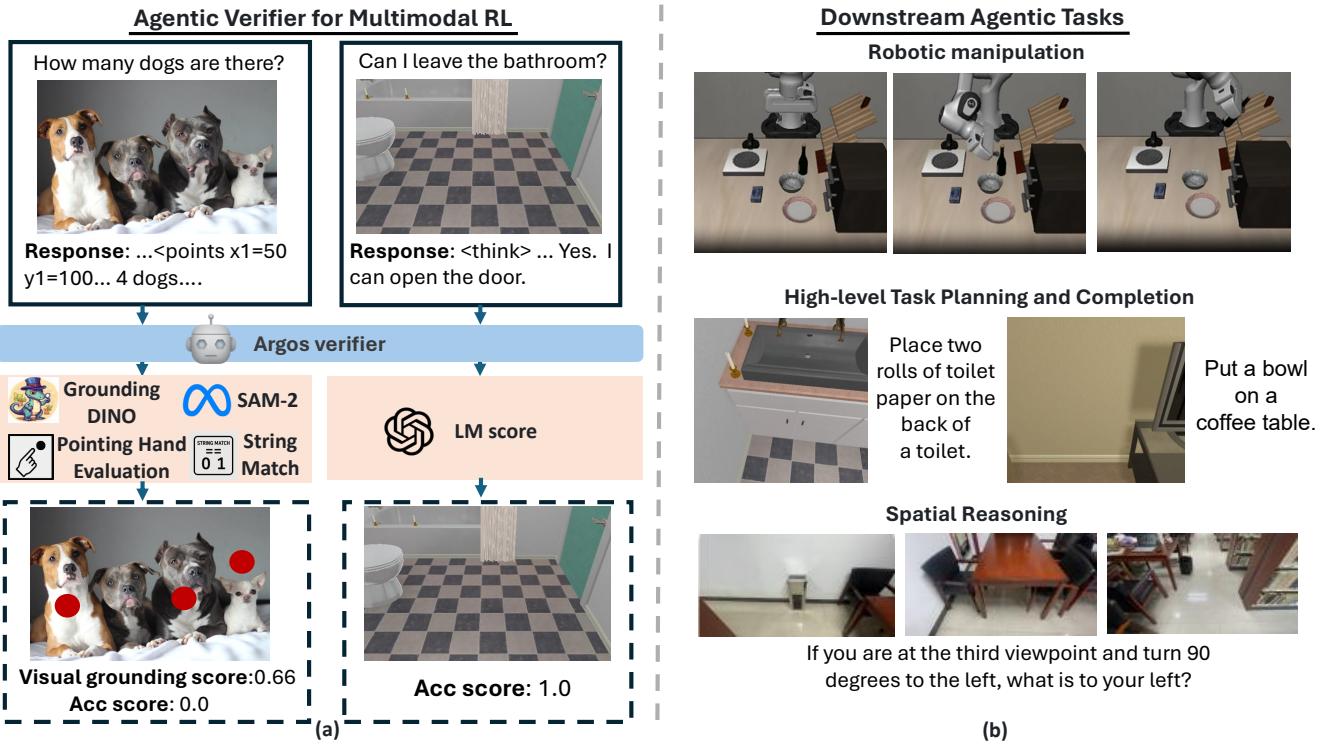


Figure 1. **Multimodal RL with our Argos agentic verifier.** We propose to train agentic foundation models using an agentic verifier Argos that adaptively selects different scoring tools based on the training sample during the RL stage. Then, we evaluate the resulting model on multiple agentic benchmarks including embodied task planning and completion as well as spatial reasoning.

## Abstract

Agentic reasoning models trained with multimodal reinforcement learning (MMRL) have become increasingly capable, yet they are almost universally optimized using sparse, outcome-based rewards computed based on the final answers. Richer rewards computed from the reasoning tokens can improve learning significantly by providing more fine-grained guidance. However, it is challenging to compute more informative rewards in MMRL beyond those based on outcomes since different samples may require different scoring functions and teacher models may provide noisy reward signals too. In this paper, we introduce the Ar-

gos (Agentic Reward for Grounded & Objective Scoring), a principled reward agent to train multimodal reasoning models for agentic tasks. For each sample, Argos selects from a pool of teacher-model derived and rule-based scoring functions to simultaneously evaluate: (i) final response accuracy, (ii) spatiotemporal localization of referred entities and actions, and (iii) the quality of the reasoning process. We find that by leveraging our agentic verifier across both SFT data curation and RL training, our model achieves state-of-the-art results across multiple agentic tasks such as spatial reasoning, visual hallucination as well as robotics and embodied AI benchmarks. Critically, we demonstrate that just relying on SFT post-training on highly curated rea-

*soning data is insufficient, as agents invariably collapse to ungrounded solutions during RL without our online verification. We also show that our agentic verifier can help to reduce reward-hacking in MMRL. Finally, we also provide a theoretical justification for the effectiveness of Argos through the concept of pareto-optimality. We will release all data, model weights and code publicly.*

## 1. Introduction

Intelligent beings seamlessly integrate perception, language, and action. With a goal in mind, they first observe a scene, interpret it in context, and then formulate and execute a plan. To emulate this ability, researchers have been shifting from static perception to agentic multimodal models that can reason about observations, plan and use tools [9, 11, 31, 34, 58]. Such agentic models for *multimodal reasoning* have wide-ranging applications, including AI agents that collaborate with humans, interactive GUI [39] and tool-using [55] assistants, and systems such as robots and self-driving cars. In particular, reinforcement learning, including the recent GRPO [52] and DAPO [67] algorithms, has been crucial in driving this progress. Verifiable outcome rewards help align such reasoning models with downstream tasks. However, using only outcome rewards provides limited guidance on the quality of the reasoning process and can cause hallucination [22]. While RL for text-only reasoning has been extensively studied, approaches for computing richer rewards in multimodal RL (MMRL) remain comparatively under-explored and introduce unique challenges, such as selecting appropriate scoring functions per sample, mitigating noisy signals from teacher models, and maintaining consistency between perception and language throughout the reasoning process.

To address the above-mentioned challenges, we introduce Agentic Reward for Grounded and Objective Scoring (Argos) verifier (Figure 1a), which adaptively selects from a set of teacher models and rule-based scoring functions like string matching to evaluate the response of each sample across spatial grounding, reasoning quality and accuracy. Our proposed verifier jointly evaluates final answer accuracy, spatiotemporal grounding and reasoning quality. Finally, we compute an aggregated final reward that is gated by correct outcomes but enriched with intermediate reward terms. Additionally, we propose an approach based on overlaying explicit 2D point coordinates on images and video frames, that leverages the OCR capability of a teacher model to generate reasoning data that are visually grounded in pixels across space and time. In addition to MMRL, we also use Argos during our data curation process to filter out low-quality rollouts from the teacher model for the SFT stage.

While some concurrent approaches have also proposed the concept of visually grounded reasoning, they place the

main focus on the curation of SFT annotations. Crucially, we observe empirically that curating grounded SFT data is not sufficient. These agentic reasoning models invariably collapse to ungrounded responses without verifying the generated 2D points during MMRL. In addition, our agentic verifier helps curb reward hacking in MMRL. Argos is also related to research on tool-augmented agents [50] but those methods generally employ tools for inference-time problem solving, leaving the intermediate reasoning and visual evidence under-verified during training. In contrast, Argos helps to convert multiple noisy reward signals into a final verifiable reward.

From a learning perspective, our verifier reframes MMRL as multi-objective optimization with multiple noisy teacher rewards. We provide a brief theoretical justification to provide an intuition on why adaptive and multi-objective reward verification can help the policy model to learn better. The modular architecture of Argos enables it to extend naturally to new modalities and objectives. As task-specific teacher models improve, our Argos has the potential to compute more informative reward signals, enabling the training of more capable and robust multimodal reasoning agents. In conclusion, we summarize our contributions as follows:

1. We propose Argos, that is used during data curation to filter out low-quality annotations and to provide aggregated and verifiable rewards during MMRL. Also, we introduce a novel data curation pipeline for generating reasoning traces that are visually grounded in space and time.
2. We demonstrate the effectiveness of Argos in achieving state-of-the-art results on multiple agentic benchmarks (Figure 1b) against similarly-sized models, including spatial intelligence reasoning, multimodal understanding, embodied task completion and robotics.
3. To the best of our knowledge, we are the first work to introduce an agentic learning framework for multimodal MMRL.

## 2. Related Work

**Multimodal understanding and reasoning.** The AI research community has seen rapid progress in large multimodal models (LMMs) that are able to process data from different modalities such as visual and audio information and generate reasonable responses. Before the advances in autoregressive language models, seminal vision-language models such as CLIP [47], ALIGN [28] and BLIP [32] are trained on web-scale image and language datasets, often with contrastive learning [44]. Building upon the advances in autoregressive language models and insights on instruction tuning [6], models including but not limited to Flamingo [1], BLIP-2/3 [33, 59], LLaVA [37] and Mini-GPT4 [75] have combined the visual capabilities of pre-trained image encoders and LLMs, leveraging the latter’s

ability for open-ended question answering, prompting and reasoning. Subsequent work broadens both scope and granularity, including region-level LMMs that operate at finer spatial resolutions [18] and video-centric LMMs designed for temporal reasoning [54, 69]. In tandem, increasingly comprehensive benchmarks have been introduced to evaluate these capabilities across tasks and modalities [38, 66]. Beyond static perception, recent works also leveraged the success of the DeepSeek-R1 [17] model with its proposed GRPO algorithm to train multimodal models that are capable of reasoning about images, videos and even audio [12, 13]. These works are highly relevant to earlier approaches that apply multimodal CoT such as (i) prompt-based strategies for zero/few-shot settings [71, 74], (ii) plan-based approaches that iteratively refine intermediate thoughts and evidence [63], and (iii) learning-based techniques that directly train models to produce rationales from paired inputs and targets [57, 64].

**Reinforcement Learning for reasoning and planning.** Multimodal planning over long horizons aims to equip AI systems with the capacity to integrate and reason over streams of multimodal inputs and observations, such as language, vision, and audio among others, across extended time horizons to complete complex, goal-driven tasks in real or simulated environments [3, 10, 26, 27, 40]. Recent models combine vision-language-action foundations with planning capabilities to execute open-ended tasks such as robotic control [29, 51] and embodied navigation [25, 61], with a particular emphasis on hierarchical planning [26, 27, 41]. A key component here is RL, which allows agents to learn robust and generalizable policies [19, 20, 43, 48, 49]. Others use RL to learn to use tools including but not limited to external APIs for computation [63] or predefined tools such as cropping and even using LMMs in the operation, as part of their multimodal reasoning loops [10, 42]. Lastly, advanced models incorporate RL fine-tuning on top of pretrained LMM backbones [5, 10, 19, 20, 27], using environment rewards to align long-term plans with task success while retaining interpretability through intermediate subgoal generation or trajectory imagination [19–21, 45, 73]. Our work builds on the idea of tool usage by using teacher models but adaptively selects them to compute multi-objective rewards instead.

### 3. Approach

We define our agentic verifier (Figure 2) as an LMM agent that selects from a set of  $K$  scoring functions to compute an aggregated reward score for each training sample. We note that the final reward computed by Argos is not a process reward as it is computed at the end of the entire response. However, our reward is more informative than the conventional outcome reward since it aggregates multiple reward

signals of intermediate reasoning steps such as their visual grounding accuracy.

For a question and visual input, we use our multimodal reasoning model, denoted as  $\pi$ , to generate a response. Each response consists of a reasoning trace and final predicted answer. The agentic verifier accepts the visual input  $v$ , question  $q$ , reasoning trace  $r$ , and predicted answer  $\hat{y}$ . To begin, it calls a parser function to extract important information from the entire response, which can include a set of 2D spatial points and temporal segments. Depending on the training sample, Argos adaptively composes a multi-objective reward process by selecting relevant tools to score the response. Finally, it leverages a gated aggregation function to compute a final reward score for the sample.

#### 3.1. Agentic Verifier

**Spatial reward.** Our key intuition behind generating reasoning thoughts that are grounded in both space and time is that it helps to alleviate the issue of hallucination of referenced objects in reasoning traces. For images, we primarily evaluate the spatial grounding accuracy. Given the set of extracted 2D spatial points mentioned in the predicted response  $P_{\text{spatial}}$ , we compute the score for each point, which is also associated with a predicted object label  $\hat{o}$ . Our spatial reward is computed in two stages. To begin, we extract a set of  $N$  generated 2D points  $P = \{(x_1, y_1, o_1), \dots, (x_N, y_N, o_N)\}$  from a rollout generated by our model, where  $x_i$ ,  $y_i$  and  $o_i$  denote the x and y coordinates, as well as the corresponding object label of the  $i$ -th point, respectively. For the  $i$ -th detected point, we use an open-vocabulary object detection model  $g_\theta$  to extract a pseudo ground-truth bounding box  $b_i^*$  based on the predicted object:  $b_i^* = g_\theta(o_i)$ . To further refine the boundaries of the object enclosed within  $b_i^*$ , we use a segmentation teacher model  $h_\phi$  to extract a fine-grained segmentation map:

$$M_i = h_\phi(b_i^*), M_i^* \in \mathbb{R}^{H \times W}, \quad (1)$$

where  $M_i$ ,  $H$ ,  $W$  denote the extracted segmentation mask, height and width of the input image, respectively. We compute the spatial grounding score for the  $i$ -th point as:

$$s_i = \mathbf{1}[M_i(x_i, y_i) = 1] \quad (2)$$

In the case of images that contain synthetic visual content such as bar charts or maps, the open-vocabulary object detection model may not work well. Thus, we use another pointing model  $f_\vartheta$  to generate 2D points before passing them into the segmentation teacher model  $h_\phi$ . Finally, we compute the spatial grounding reward  $R_{\text{spatial}}$  as follows:

$$R_{\text{spatial}} = \frac{1}{N} \sum_{i=1}^N s_i. \quad (3)$$

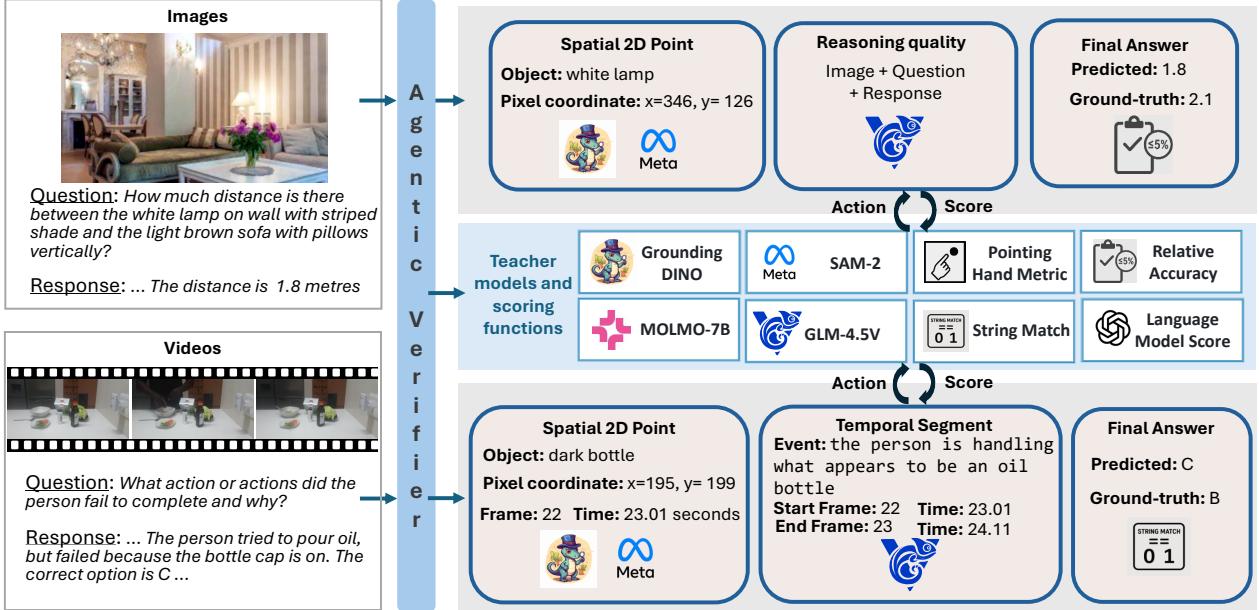


Figure 2. **Verification process.** We use the same set of scoring functions for both images and videos. Each response is first parsed to extract information about generated 2D points, temporal segments, reasoning text and answer. Then, the agentic verifier adaptively decides what scoring functions to call based on the extracted information. Finally, we aggregate the scores using a gated aggregation function.

**Temporal rewards.** The scoring functions used to compute the spatial reward term can be easily extended to videos. When a reasoning trace references an action or event in the video that span multiple frames, it is also important to verify its existence in the video. Given a video  $V$  consisting of  $N_v$  frames and a question, we use an LLM to extract both frame-level observations  $F$  and segment-level events or actions  $E$  that span multiple frames from the generated response. We provide the query prompt used for extraction in the supplemental.

We define  $F$  as a set of  $N_F$  frame-level observations  $F = \{(t_1, x_1, y_1, o_1), \dots, (t_{N_F}, x_{N_F}, y_{N_F}, o_{N_F})\}$ , where  $t_i$  can either denote the relevant frame or its corresponding timestamp. In this setting, the identified frame is analogous to an image. Thus, we leverage the spatial grounding models  $f_\theta$ ,  $g_\theta$  and  $h_\phi$  described above to compute the set of frame-level scores  $S_f$ .

We consider a set of segment-level events  $E = \{e_i\}_{i=1}^N$ , where the  $i$ -th event is represented as the tuple  $(t_i^{\text{start}}, t_i^{\text{end}}, d_i)$  with  $t_i^{\text{start}}$  and  $t_i^{\text{end}}$  denoting the start and end times (or frame indices) of the segment, and  $d_i$  the event description. Then, we query a powerful teacher reasoning model  $T$  to evaluate the visual–semantic accuracy between  $d_i$  and the corresponding video segment  $V_{t_i^{\text{start}}:t_i^{\text{end}}}$ , and return a binary score:

$$s_i = \text{video\_score}(d_i, V_{t_i^{\text{start}}:t_i^{\text{end}}}) \in \{0, 1\}. \quad (4)$$

where  $\text{video\_score}$  is the function parameterized by the reasoning teacher model. Finally, we compute the final video grounding score by computing the within-set means of the

set of event scores  $S_e$  and  $S_f$  and defining the final score as the (unweighted) average of these means.

**Reasoning quality reward.** Beyond evaluating the intermediate grounding of referenced objects and actions, we also evaluate the logical consistency between the generated reasoning trace and the final answer  $\hat{y}$ . In some cases, the policy model may generate reasonable reasoning traces but still predict the wrong answer at the end. We use a larger teacher model to compute a reasoning-quality reward as its conditional probability of the predicted response  $y$  given the question  $q$ , reasoning trace  $r$  and visual input  $v$ :

$$R_{\text{reasoning}} = P(\hat{y} | q, r, v). \quad (5)$$

Intuitively, higher values indicate stronger consistency between the reasoning and the answer in the context of the question and visual input.

**Outcome rewards.** To compute the final outcome rewards using the ground-truth answer  $y^*$ , we use a combination of a language model as well as heuristic-based functions depending on the type of the question and expected answer format.

(i) *Exact string match.* For multiple-choice questions and those that require short phrases as an answer, we compute:

$$R_{\text{acc}} = \mathbf{1}\{\hat{y} = y^*\} \quad (6)$$

(ii) *Relative numerical accuracy with 5% tolerance.* When both answers are float numbers, we compute the relative error as:

$$\text{relerr}(\hat{y}, y^*) = \frac{|\hat{y} - y^*|}{\max(|y^*|, 1)}. \quad (7)$$

$$R_{\text{acc}} = \mathbf{1}\{\text{relerr}(\hat{y}, y^*) \leq 0.05\}. \quad (8)$$

(iii) *Binary semantic accuracy via a language model.* We query a capable language model to assign  $R_{\text{acc}}$  a value of 1 if  $\hat{y}$  is semantically equivalent to  $y^*$  without any contradictions and 0 otherwise.

**Aggregation function.** Finally, we aggregate the rewards from selected scoring functions using a gated scoring function to prevent potentially noisy rewards from biasing the final answer away from the correct result. We formulate the gated function as follows:

$$R_{\text{final}} = \begin{cases} R_{\text{acc}}, & R_{\text{acc}} < \tau, \\ \frac{w_A R_{\text{acc}} + w_S R_{\text{spatial}} + w_R R_{\text{reasoning}}}{w_A + w_G + w_R}, & R_{\text{acc}} \geq \tau, \end{cases} \quad (9)$$

where  $w_A$ ,  $w_G$  and  $w_R$  denote the weight terms for the outcome, visual grounding and reasoning quality rewards, respectively.

### 3.2. GRPO training

Within each group of rollout reward values for the  $i$ -th training sample, we compute the advantage  $A_i$  over the individual  $j$ -th reward values:

$$A_i = \frac{R_i - \text{mean}(\{R_j\})}{\text{std}(\{R_j\})} \quad (10)$$

Using the aggregated reward terms, we update our policy model  $\pi_\theta$  using the GRPO formulation [52]:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{\hat{y}_i\}} \left[ \frac{1}{G} \sum_{i=1}^G \min \left( \frac{\pi_\theta(\hat{y}_i | q)}{\pi_{\theta_{\text{old}}}(\hat{y}_i | q)} A_i, \right. \right. \\ \left. \left. \text{clip} \left( \frac{\pi_\theta(\hat{y}_i | q)}{\pi_{\theta_{\text{old}}}(\hat{y}_i | q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right]. \end{aligned} \quad (11)$$

### 3.3. Theoretical analysis

To justify the intuition that combining complementary but potentially weak reward teachers may train stronger agentic models with MMRL, we provide a theoretical analysis on learning with multiple reward estimators. Inspired by [72], we show that even noisy reward signals in aggregation can guide the policy towards global Pareto-optimal solutions.

Consider any prompt/input to the model, each possible answer/action  $a \in \mathcal{A}$  has  $m$  true reward components  $R_1(a), \dots, R_m(a)$ , measuring its quality from different aspects (e.g. grounding, reasoning, etc.). The multiple rewards can arbitrarily correlate with each other, and the goal is to learn towards Pareto optimality defined as follows:

**Definition 1** ( $\delta$ -Pareto Domination). *For  $\delta > 0$ , we say  $a' \succ_\delta a$  if  $R_i(a') \geq R_i(a) + \delta$  for all  $i = 1, \dots, m$ .*

**Definition 2** ( $\delta$ -Pareto Optimality). *For  $\delta > 0$ , the set of globally  $\delta$ -Pareto-optimal actions*

$$P_\delta := \{a \in \mathcal{A} : \nexists a' \in \mathcal{A} \text{ s.t. } a' \succ_\delta a\}.$$

However, in practice we do not have access to the true rewards  $R_i(a)$ 's, but instead weak estimators  $\hat{R}_i(a) = R_i(a) + \varepsilon_i(a)$ . The estimated rewards can be inaccurate and correlated, with the only assumption as follows:

**Assumption 1.** *For any  $a$ , the noise variables  $\{\varepsilon_i(a)\}_{i=1}^m$  are independent, mean-zero, and  $\sigma$ -sub-Gaussian:*

$$\mathbb{E}[\varepsilon_i(a)] = 0, \quad \mathbb{E}[e^{\lambda \varepsilon_i(a)}] \leq e^{\sigma^2 \lambda^2 / 2}, \quad \forall \lambda \in \mathbb{R}.$$

The  $\sigma$ -sub-Gaussian assumption is a generic form constraining the error, which is easily satisfied by Hoeffding's lemma when the rewards are bounded. In each RFT step,

$$\hat{R}(a) = \sum_{i=1}^m w_i \hat{R}_i(a) = R(a) + \sum_{i=1}^m w_i \varepsilon_i(a),$$

where  $R(a) = \sum_{i=1}^m w_i R_i(a)$ , and  $w_1, \dots, w_m > 0$ . Denote  $w_{\min} = \min_i w_i$  and  $w_{\max} = \max_i w_i$ . We can show probability bound on the estimated  $\delta$ -Pareto optimality.

**Theorem 1** (Global Pareto Guarantee). *Let  $\pi$  be the sampling policy and denote  $\beta = \pi(P_\delta)$  as the probability coverage on Pareto optimal solutions. Sample  $n$  i.i.d. actions from  $\pi$  to form a group  $\mathbb{G}$  with  $\hat{a} = \arg \max_{a \in \mathbb{G}} \hat{R}(a)$ . Then*

$$\mathbb{P}(\hat{a} \in P_\delta) \geq (1 - (1 - \beta)^n) \left[ 1 - \frac{n-1}{e^{C \cdot m}} \right],$$

where  $C := \frac{\delta^2}{4\sigma^2} \cdot \frac{w_{\min}^2}{w_{\max}^2} > 0$  is a constant.

The theorem shows that as the number of rewards  $m$  increases, we can approximate global Pareto optimality even with weak estimators. We provide full proof in appendix.

## 4. Data Curation

We aim to generate training data for SFT cold-start to help the base model learn to reason. We illustrate the main stages of our curation pipeline for generating reasoning traces that are visually grounded across space and time in Figure 3. We primarily use the highly-capable GLM-4.1V [23] model to generate the reasoning traces although we also use a smaller amount of generations from a proprietary model from Gemini-2.5 Flash [15] to augment our SFT dataset due to computational constraints. We discuss the three main stages of our entire generation pipeline at a high-level in this section and provide more specific details on each step and examples of our curated SFT training samples in the supplemental.

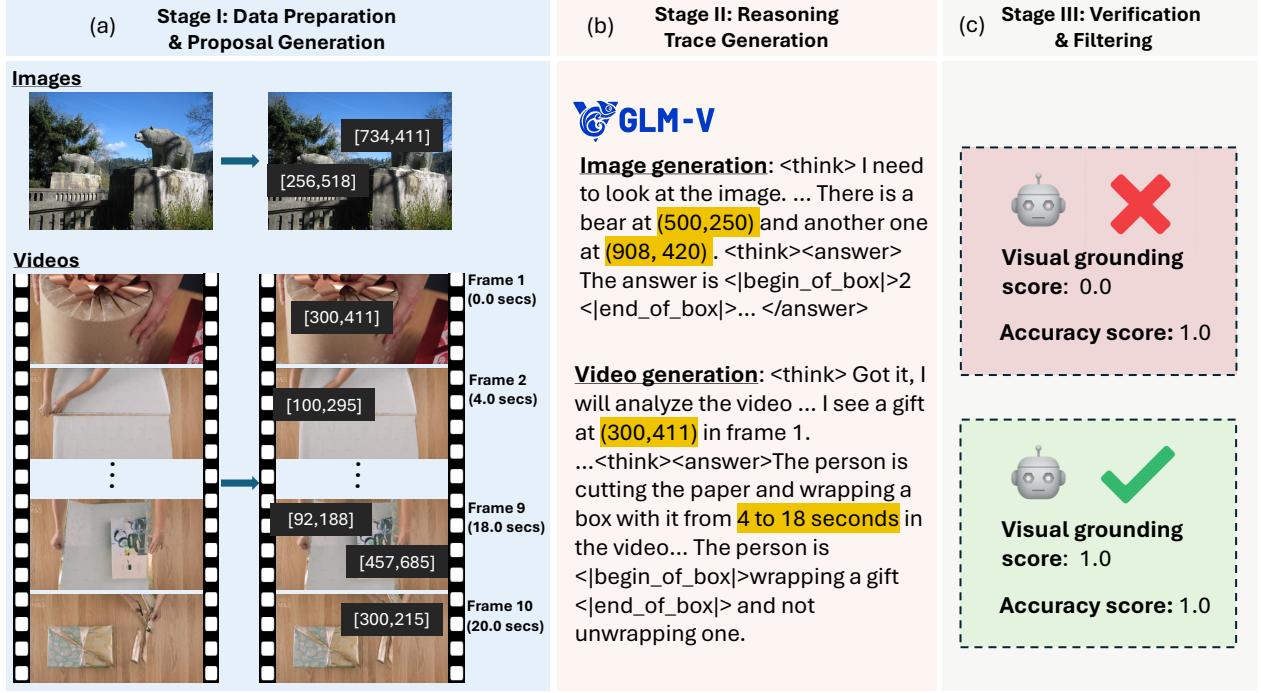


Figure 3. **Grounded reasoning generation pipeline.** (a) **Stage I:** We extract object, action and event proposals such as 2D boxes for images and video frames as well as temporal segments for videos. (b) **Stage II.** We use the overlaid images and video frames to prompt a pretrained LMM to generate grounded reasoning traces that explicitly refer to these points. For videos, we also include the frame numbers and their timestamps in the query. (c) **Stage III.** Our agentic verifier adaptively scores each trace using multi-objective rewards (e.g., visual grounding and answer accuracy) and filters out samples with low-quality generations. In the image example with the bears, the sample is filtered out due to low visual grounding accuracy despite predicting the correct answer.

#### 4.1. Data preparation and proposal generation

While the GLM-4.1V teacher model can perform object localization well, it is unable to perform grounded reasoning naturally given a question and visual input. Given the visual input, question and ground-truth answer, we begin by extracting information about relevant objects, actions and events before extracting their spatial 2D and temporal positions. Based on our observation that it performs well on the task of OCR, we first use the Molmo-7B [8] model to extract 2D points of relevant objects. As shown in Figure 3(a), we overlay the spatial 2D points on the image or sampled video frames. Additionally, for the temporal dimension in videos, we apply a similar concept by providing explicit timestamps using both frame numbers and time in seconds. Each frame is mapped to its accurate timestamp, which is computed based on its sampling FPS. We set a maximum limit of 32 frames in experiments. In our query, we provide the input overlaid frames along with their corresponding timestamps.

#### 4.2. Reasoning trace generation

After the object locations and event timestamps have been extracted, we query the GLM-4.1V model with the overlaid images and video frames (Figure 3(b)) to generate reasoning traces that contain 2D points when referring to specific

objects. In the case of images, we prompt the GLM-4.1V model to primarily use the visual information in the original image before referring to the coordinates on the overlaid image for reference to clarify ambiguous objects or referring expressions in its response. For videos, we prompt the teacher model to explicitly refer to 2D points in frames and multi-frame events in specific formats such as “(x,y) in frame F (T seconds)” as well as “from  $t_{start}$  to  $t_{seconds}$ ”. Note that we do not include the original video frames due to computational constraints. For each sample, we generate eight possible rollouts.

#### 4.3. Verification and filtering

Despite our curation pipeline for visually grounded reasoning traces, state-of-the-art reasoning models still produce unreliable rollouts with high frequency. For example, our yield rate was around 3.1%. Thus, in addition to using Argos to provide adaptive and dense rewards for multimodal reinforcement learning, we also use it to evaluate the generated rollouts for the training samples and filter out samples if the maximum score over all rollouts fall below a threshold value (Figure 3(c)). We parse the generated 2D points using regular expression and reformat the points into the format: <point x1=”x” y1=”y” alt=”object”>”object”</point>. Similarly, we also extract the templated timestamps in the

video reasoning traces and replace them with reformatted natural language phrases. This filtering step ensures that the data used for SFT consists predominantly of visually grounded and semantically accurate reasoning examples.

## 5. Experiments

In this section, we evaluate our trained model post-SFT and RL on multiple agentic benchmarks across different domains under the zero-shot setting. A key skill for multimodal AI agents to interact with their physical worlds is spatial intelligence and the ability to reason about different viewpoint perspectives and motion. We begin by evaluating on multiple vision-centric and spatial reasoning benchmarks. Next, we also examine the benefits of learning to perform grounded reasoning with Argos on reducing visual hallucination since agents have to make confident and accurate predictions. Finally, we also evaluate on fine-grained robotic manipulation and high-level task planning.

**Implementation details.** We build our approach off the publicly available Qwen2.5-VL 7B [2] model and train on our curated dataset for SFT (as discussed in Section 4) and a separate and non-overlapping subset of the same dataset for the RL training with GRPO. We provide further details about our training setup in the supplemental.

### 5.1. Spatial Reasoning Evaluations

Model	BLINK	MindCube-t	CV-Bench	CV-Bench (3D)
Qwen2.5VL 7B	54.4	34.9	77.0	77.9
Qwen2.5VL 7B (CoT)	53.5	33.1	75.6	76.6
Video-R1 (SFT)	52.7	34.2	75.1	75.6
Video-R1 (RL)	49.0	31.9	60.2	57.2
Argos (Ours)	<b>56.0</b>	<b>39.6</b>	<b>78.2</b>	<b>82.0</b>

Table 1. Results on spatial reasoning benchmarks.

We report results of Argos on multiple spatial reasoning benchmarks in Table 1. For all datasets, we use accuracy (%) as the metric.

**BLINK.** The BLINK dataset [14] contains 14 visual perception tasks that include spatial and multiview reasoning, as well as functional correspondence. Argos achieves a performance gain of over 12% over the baseline Qwen2.5-VL and even outperforms Video-R1, which was trained on around 160K more training samples than ours. These consistent improvements suggest that more accurate grounded reasoning may help close the gap with much larger models.

**MindCube.** The MindCube [65] benchmark focuses on evaluating the ability of LMMs to perform spatial reasoning by reconstructing spatial mental models using partial observations and dynamic viewpoints. We evaluate on the tiny split which contains around 1K evaluation samples. Interestingly, CoT prompting actually hurts the performance

of the base Qwen2.5VL model. In contrast, our model also outperforms the base model by over 5%.

**CV-Bench.** CV-Bench [56] is a vision-centric benchmark that assesses 2D understanding through spatial relationships and object counting, and 3D understanding through depth ordering and relative distance. Consistent with results on other datasets, our model trained with the proposed Argos gains a significant improvement over the state-of-the-art SOTA Video-R1 variants. It is worth-noting that training with visually grounded reasoning traces in 2D images enhances the resulting model’s generalization capabilities to 3D visual understanding. This appears to be corroborated by performance gains achieved by our model on embodied AI tasks, that we discuss in later sections.

Model	CounterCurate	HallusionBench	SugarCrepe
Qwen2.5VL-7B	61.4	42.4	85.2
Qwen2.5VL-7B (CoT)	60.4	38.6	83.2
Video-R1 (SFT)	60.6	40.1	83.3
Video-R1 (RL)	63.6	25.8	79.9
Argos (Ours)	<b>85.3</b>	<b>46.6</b>	<b>86.4</b>

Table 2. Results on visual hallucination benchmarks.

### 5.2. Hallucination Reasoning

To evaluate the effectiveness of performing grounded reasoning, we compare Argos against baseline approaches on three evaluation benchmarks: CounterCurate [70], HallusionBench [16] and SugarCrepe [24]. The results are summarized in Table 2. These benchmarks are aimed at evaluating the capabilities of LMMs to reduce hallucination of visual concepts in their generated responses. First, we observe that our Argos achieves a significant relative performance gain of more than 20% over the Qwen2.5VL-7B base model on CounterCurate. We also see consistent improvements obtained by SOTA multimodal reasoning models, with Video-R1 outperforming the base model by  $\sim 5\%$ . This suggests that reasoning is an important and necessary capability to reduce visual hallucination. We note that CounterCurate is considered to be an easier benchmark than HallusionBench and SugarCrepe, as it primarily evaluates on models’ ability to differentiate between left/right and top-/down.

Despite using fewer training samples during both SFT and RL stages, our Argos outperforms Video-R1 by significant margins on both HallusionBench and SugarCrepe. These results further emphasize the importance of performing visually grounded reasoning on alleviating hallucination in LMMs. In contrast, we see that Video-R1 actually performs worse than the Qwen2.5VL-7B base model. A plausible reason is that Video-R1 may have overfitted to the post-training datasets during the SFT and RL stages.

Model	Base	Common	Complex	Visual	Spatial	Long	Avg
Qwen2.5-VL-7B	4.0	3.3	2.0	0.0	0.7	1.3	1.9
Qwen2.5-VL-7B (CoT)	6.0	9.3	7.3	5.3	4.7	0.7	5.6
Video-R1	16.7	11.3	18.0	8.0	5.3	0.7	10.0
Argos (Ours)	<b>24.7</b>	<b>18.0</b>	<b>27.3</b>	<b>8.7</b>	<b>8.7</b>	<b>0.7</b>	<b>14.7</b>

Table 3. Results on EB-Alfred.

Model	Base	Common	Complex	Visual	Spatial	Long	Avg
Qwen2.5-VL-7B	23.3	5.3	10.7	6.7	8.0	0.0	9.0
Qwen2.5-VL-7B (CoT)	30.7	6.7	13.3	11.3	12.0	1.3	12.6
Video-R1	<b>46.7</b>	6.0	15.3	9.3	16.7	3.3	16.2
Argos (Ours)	45.3	<b>12.0</b>	<b>24.0</b>	<b>16.0</b>	<b>17.3</b>	<b>9.3</b>	<b>20.7</b>

Table 4. Results on EB-Habitat.

Model	Libero					
	Object	Spatial	Goal	Long	90	Avg
Qwen2.5-VL 7B	84.0	<b>93.6</b>	80.6	60.2	82.3	80.1
Qwen2.5-VL-7B-Instruct	81.6	93.4	84.4	62.4	77.3	79.8
Qwen2.5-VL 7B (SFT)	88.0	91.0	84.0	66.1	83.3	82.4
Video-R1 (SFT)	88.2	91.0	85.2	64.0	82.7	82.2
Video-R1	89.2	93.0	89.6	65.6	80.1	83.5
Argos (Ours)	<b>93.2</b>	91.2	<b>87.8</b>	<b>63.8</b>	<b>85.0</b>	<b>84.2</b>

Table 5. Success rates (%) on the LIBERO [35] continuous control robotics benchmark. When finetuned to predict continuous robot control actions, our MMRL approach outperforms baselines in terms of both performance and data efficiency.

### 5.3. Embodied AI

To determine the importance of performing grounded reasoning for agentic foundation models, we also report the results of our evaluations on EmbodiedBench [62] across high-level task planning and completion tasks in the Alfred and Habitat environments. The results are summarized in Tables 3 and 4. As shown in Table 3, the base Qwen2.5VL-7B model generalizes poorly to task planning for agentic task completion, despite its strong performance on standard visual question answering benchmarks. Notably, our Argos improves significantly on the sub-category of “complex” tasks by over 25% for task success rates to the base model. This result shows that the reason capability is particularly beneficial for planning solving complex multi-step tasks. Furthermore, our Argos is able to outperform baselines on the “visual” and “spatial” subtasks. This superior performance strongly suggest that localizing referred objects explicitly in the reasoning trace helps LMMs to leverage the visual content far more effectively.

In the Habitat evaluation environment, we observe similar trends (Table 4) as those found on the Alfred benchmark. Here, CoT prompting is beneficial on high-level task planning even for the non-reasoning base model, as evidenced by the  $\sim 3.6\%$  performance gain obtained over the base Qwen2.5-VL-7B model on average. In particular, the results demonstrate the clear benefits of adding grounded reasoning for generalizing to complex visual understanding tasks, where our model outperforms the CoT-prompted Qwen2.5-VL-7B by approximately 7%.

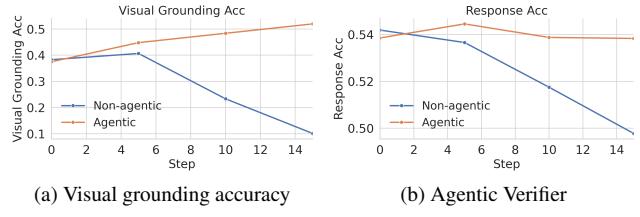


Figure 4. We run a small-scale comparison to ablate the effectiveness of Argos (*agentic*) compared to using only outcome rewards (*non-agentic*). We evaluate on a separate dataset for both variants.

### 5.4. Robotics evaluations

Embodied decision-making often requires learning complex and generalizable robotic behaviors. Such agents typically require representations informed by world knowledge for perceptual grounding, planning, and control. Given that representations learned via grounded SFT embodied chain-of-thought [4, 60, 68] have been shown to be conducive to more generalizable Vision-Language-Action models (VLAs) [7, 30, 76], we further evaluate our model on complex robotics tasks. We post-train Argos and the baselines as VLAs to assess the transferability of their representations and their ability to fit complex, multimodal action distributions across multiple simulated benchmarks.

**Libero.** We evaluate our method on the LIBERO simulation benchmark [35], which utilizes a Panda robot with delta end-effector control. Our evaluation follows two distinct protocols based on established task suites:

- **Specialized Suites:** We first evaluate on four 10-task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. For this setup, we train a single policy on the combined datasets from all four suites.
- **Diverse Suite:** We separately evaluate on the LIBERO-90 benchmark, which comprises 90 different tasks. We train a dedicated policy using only the LIBERO-90 dataset.

For all experiments, we use both third-person and wrist camera images as input. The training data is prepared by re-rendering images to 224x224, filtering unsuccessful demonstrations, and removing “no-op” actions, following Pertsch et al. [46]. We report the percentage of successful task completions in Table 5, averaged over 50 trials per task for the four specialized suites and 20 trials per task for LIBERO-90. We observe that our method outperforms the base model, the SFT models and even Video-R1, which was trained on 270k data samples, some of which contained spatial related data which should be beneficial for libero. In contrast, our multimodal reinforcement learning approach outperforms the baselines both in terms of performance as well as sample efficiency (260k vs 85k).

### 5.5. Ablation Study

We analyze the importance of using our Argos to compute more informative rewards for MMRL on a small subset of

the Pixmo-Count [8] dataset. Starting from the same SFT checkpoint finetuned with our curated data, we train two variants: one with Argos (agentic verifier) and one only using the outcome reward (non-agentic verifier). We report the evaluation results on an out-of-domain (OOD) evaluation dataset of 1.5K samples sampled from Video-R1-CoT-165k in Figure 4. We observe that, without verifiers of Argos, the visual grounding accuracy drops rapidly. Importantly, training only on the outcome reward also leads to a performance drop on the validation set. This suggests that the aggregated final reward from our agentic verifier helps reduce the likelihood of reward hacking.

## 6. Conclusion

In this work, we introduce Argos, a novel approach to perform MMRL by adaptively selecting different teacher models to compute dense rewards on a per-sample basis. It formulates MMRL into a multi-objective optimization problem by jointly rewarding outcome accuracy, spatiotemporal grounding, and reasoning quality, rather than relying on final answers alone. Importantly, we made two crucial observations. First, curating SFT grounded annotations is insufficient to train a grounded reasoning model. Second, our Argos helps the base model to avoid reward-hacking by providing dense and robust rewards. Extensive experiments show that Argos significantly outperforms prior work on a wide array of challenging agentic tasks, from spatial reasoning in images and videos to embodied AI and robotics. We hope that advances in task-specific models will improve Argos’s reward signals, enabling the training of more effective and better-grounded multimodal agents in the future.

**Acknowledgements.** We would like to thank Matt Mazola and Jianwei Yang for their valuable insights in the initial discussions.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7
- [3] Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022. 3
- [4] William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. Training strategies for efficient embodied reasoning. In *Conference on Robot Learning*, 2025. 8
- [5] William Chen, Oier Mees, Aviral Kumar, and Sergey Levine. Vision-language models provide promptable representations for reinforcement learning. *Transactions on Machine Learning Research*, 2025. 3
- [6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 2
- [7] Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishnan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Ming-tong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaaf, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yuehua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. Open X-Embodiment: Robotic learning datasets and RT-X models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, 2024. 8
- [8] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv e-prints*, pages arXiv–2409, 2024. 6, 9, 11
- [9] Sixun Dong. Tool-lmm: A large multi-modal model for tool agent learning. 2024. 2
- [10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023. 3
- [11] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent ai: Surveying the horizons of multimodal interaction. *CoRR*, 2024. 2
- [12] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayananaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025. 3
- [13] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 3, 11
- [14] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 7, 1
- [15] Google Gemini Team. Gemini: A family of highly capable multimodal models. 2023. 5
- [16] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 7
- [17] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3
- [18] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regionopt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024. 3
- [19] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3
- [20] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025. 3
- [21] Kyle Beltran Hatch, Ashwin Balakrishna, Oier Mees, Suraj Nair, Seohong Park, Blake Wulfe, Masha Itkina, Benjamin Eysenbach, Sergey Levine, Thomas Kollar, and Benjamin Burchfiel. Ghil-glue: Hierarchical control with filtered subgoal images. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Atlanta, USA, 2025. 3
- [22] Tao He, Hao Li, Jingchang Chen, Runxuan Liu, Yixin Cao, Lizi Liao, Zihao Zheng, Zheng Chu, Jiafeng Liang, Ming Liu, et al. Breaking the reasoning barrier a survey on llm complex reasoning through the lens of self-evolution. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7377–7417, 2025. 2
- [23] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Li-hang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 5
- [24] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrape: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023. 7, 1
- [25] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023. 3
- [26] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Multimodal spatial language maps for robot navigation and manipulation. *International Journal of Robotics Research (IJRR)*, 2025. 3
- [27] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 3
- [28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint*, 2021. 2
- [29] Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Fine-tuning generalist robot policies with heterogeneous sensors via language grounding. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Atlanta, USA, 2025. 3
- [30] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025. 8
- [31] Somnath Kumar, Yash Gadhia, Tanuja Ganu, and Akshay Nambi. Mrmtagent: Multi-modal critical thinking agent framework for complex visual reasoning. *arXiv preprint arXiv:2405.18358*, 2024. 2
- [32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2
- [33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [34] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025. 2
- [35] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 8, 2
- [36] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusion-bench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023. 1
- [37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2
- [38] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 3
- [39] Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025. 2
- [40] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022. 3

- [41] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023. 3
- [42] Vivek Myers, Bill Chunyuan Zheng, Oier Mees, Sergey Levine, and Kuan Fang. Policy adaptation via language optimization: Decomposing tasks for few-shot imitation. In *Conference on Robot Learning*, 2024. 3
- [43] Mitsuhiro Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. *Conference on Robot Learning (CoRL)*, 2024. 3
- [44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2
- [45] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023. 3
- [46] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, Los Angeles, USA, 2025. 8
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2
- [48] Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89):eadi9579, 2024. 3
- [49] Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task agnostic offline reinforcement learning. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, Auckland, New Zealand, 2022. 3
- [50] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023. 2
- [51] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debiddatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE, 2024. 3
- [52] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 5
- [53] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024. 1
- [54] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-lm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13581–13591, 2024. 3
- [55] Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025. 2
- [56] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 7, 1
- [57] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 3
- [58] Zhenyu Wang, Aoxue Li, Zhengu Li, and Xihui Liu. Genartist: Multimodal lm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37:128374–128395, 2024. 2
- [59] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. Blip-3: A family of open large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6124–6135, 2025. 2
- [60] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multi-modal ai agents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14203–14214, 2025. 8
- [61] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 3
- [62] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Forty-second International Conference on Machine Learning*, 2025. 8, 2
- [63] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. 3
- [64] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi,

- Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3
- [65] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025. 7, 1
- [66] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 3
- [67] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 2
- [68] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *Conference on Robot Learning*, 2024. 8
- [69] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3
- [70] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visiolinguistic compositional reasoning via counterfactual examples. *arXiv preprint arXiv:2402.13254*, 2024. 7, 1
- [71] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3
- [72] Theodore Zhao, Mu Wei, J Preston, and Hoifung Poon. Pareto optimal learning for estimating large language model errors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10513–10529, 2024. 5
- [73] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024. 3
- [74] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022. 3
- [75] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2
- [76] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 8

# Multimodal Reinforcement Learning with Agentic Verifier for AI Agents

## Supplementary Material

In this supplemental, we provide the following additional material to the main paper:

- A Implementation details for training and evaluation
- B Agentic evaluation benchmark details
  - (a) Spatial Reasoning Benchmarks
  - (b) Visual Hallucination Benchmarks
  - (c) Embodied AI and Robotics Benchmarks
- C Additional details on data curation pipeline
  - (a) Image reasoning trace examples
  - (b) Video reasoning trace examples
- D Agentic Verifier Scoring functions
  - (a) Spatial Grounding Reward for Images
  - (b) Spatiotemporal Grounding Reward for Videos
- E Full theoretical proof of Pareto-optimal solutions
- F Additional ablation experiments
  - (a) Ablation of training reward signals
  - (b) Ablation over training hyperparameters
- G Qualitative visualizations

### A. Additional implementation details for training and evaluation

We train our SFT and RL model variants using the AdamW optimizer with a learning rate of  $1e^{-5}$ , and 256 batch size for SFT and 56 for RL. We observe that training converges in around 1000 steps for SFT and 80 steps for RL. We implement our two-stage training pipeline using PyTorch with a combination of 8xH100 and 8xA100 40GB GPUs based on the Easy-R1 training framework [53]. During evaluation, we use a maximum of 6144 or 8192 new tokens and a temperature value of 0.6.

### B. Agentic evaluation benchmarks

In our experiment section, we mainly evaluate on several benchmarks that evaluate the agentic reasoning capabilities of our model on multiple tasks in diverse domains including visual hallucination, spatial intelligence and embodied AI. Here, we dive deeper into the setup and composition of these aforementioned benchmarks.

**Visual Hallucination.** Visual hallucination [36, 70] has been a critical limitation of using reasoning and non-reasoning based LMMs as AI agents. When a LMM hallucinate visual details such as non-existent objects or spurious relationships, it reduces the confidence of each subsequent decision since there might be limited relevance between perception, reasoning and planned actions. This is especially problematic in interactive settings where agents

have to follow instructions to complete physical tasks or navigate GUI elements. Consequently, robustness to visual hallucination is a prerequisite for the deployment of multimodal AI agents. Thus, we evaluate the effectiveness of our model trained with the agentic verifier to remain grounded in the visual scene. We use the accuracy metric for all visual hallucination benchmarks.

The first benchmark is **CounterCurate** [70], which contains hard positive and negative image-caption pairs for understanding ambiguous spatial relations. It is curated from the Flickr30K dataset and uses the GPT-4V and DALLE 3 models to generate semantically plausible but incorrect captions and images. We also evaluate on a visual-context **HallusionBench** [36], which is aimed at evaluating the tendency of LMMs to be affected by language hallucination and visual illusion. It contains 346 images and 1129 questions that are authored by humans. Additionally, the questions are further categorized into those that are *visual-dependent* and *visual-supplement*. Finally, the **Sugarcrape** [24] benchmark is introduced to evaluate LMMs' capability to understand fine-grained vision-language compositionality. It contains 7512 questions to systematically cover 7 types of hard negatives including the replacement and addition of objects.

**Spatial Reasoning.** Besides robustness to visual hallucination, spatial reasoning is a core capability for multimodal agents to interact effectively in the physical world. AI agents have to understand where objects are and the relationships between them in space to generate feasible and accurate responses. We report performance using the accuracy metric since all benchmarks are based on multiple choice-questions.

First, we begin by evaluating on the **BLINK** [14] benchmark, which contains around 3.8K questions that are generated from 7.3K images. This benchmark sources questions from 14 classic computer vision tasks including depth estimation and visual similarity. Second, we use the tiny version of the MindCube [65] dataset, which is one of the most recent spatial reasoning benchmarks. It is intended to test the ability of LMMs to construct internal spatial mental models from a few visual perspectives. The full benchmark contains around 21K questions, which are categorized into the “Around”, “Among” and “Rotation” types. Due to the scale of the data, we evaluate on the tiny version, which is a smaller but representative subset spanning the same spatial settings. Finally, we also evaluate on the Cambrian Vision-Centric Benchmark (CV-Bench) [56], which consists of about 2.6K manually-inspected evaluation sam-

## Image Instruction Template

### You are given:

1. The original question: {question}
2. The associated image (if any).
3. A model-generated answer: {generated\_text}

### TASK

Extract a concise set of referenced **entities, objects, and interactions** that are **visibly present in the image**. Focus on the **main objects** and their key visible attributes. When descriptive expressions are used, prefer **expressive but non-redundant terms** that capture the important visible features (e.g., “seedling with roots” instead of just “seedling”).

### ENTITY RULES

- Entities should be **specific and expressive**, not overly vague.
- Avoid redundant variants that describe the same object with slightly different wording (e.g., do not include both “seedling with roots” and “early growth with roots and leaves”; pick the clearest and most representative one).
- Merge synonyms/aliases into one expressive label.
- Include distinct parts (like “roots”, “leaves”, “soil”) only if they are visually important on their own, not just mentioned as part of a longer phrase.
- If more than 10 entities are visible, include only the **10 most salient** ones (salient = important, central, or repeatedly emphasized).

### INTERACTION RULES

- Only include interactions that are visually supported (e.g., arrows, measurable processes, progressive changes).
- Express them concisely but clearly (e.g., “growth progression”, “roots growing in soil”).
- Merge duplicates or semantically equivalent variants.

### OUTPUT FORMAT

Return **only** this JSON object (no explanations, no extra text):

```
{  
    "entities": [up to 10 expressive, non-redundant entities/objects visible in the image],  
    "interactions": [unique interactions visually supported by the image, if any]  
}
```

Figure 5. Instruction template for extracting entities and interactions from generated image-level rollouts.

ples. These questions evaluate LMMs’ 2D and 3D spatial understanding capability.

**Embodied AI and Robotics.** Last but not least, we also evaluate on agentic task planning and completion using two popular benchmarks. The first benchmark, Embodied-Bench [62], is a dataset that spans multiple environments such as AI2-THOR and Habitat 2.0. In our experiments, we mainly focus on the high-level task completion setting. The EB-Alfred sub-task evaluates an agent’s ability to compose and execute high-level skills like pick up, open, slice, and find objects over diverse household tasks. It contains approximately 300 test samples evenly split across the categories of base, commonsense, complex instructions, visual appearance, spatial awareness and long-horizon. In contrast, the EB-Habitat evaluation task focuses on high-level object rearrangement based on language queries. Finally, we also evaluate on the Libero evaluation suite [35], which is a dataset for language-based robot manipulation. It mainly consists of Libero-Spatial, Libero-Object, Libero-Goal and Libero-100. The Libero-100 subset is further split into Libero-90 which contains 90 short-horizon tasks and

10 long-horizon tasks in Libero-Long.

## C. Additional details on data curation

In this section, we present more specific details about our entire data curation pipeline starting from rollout generation to the final verification process.

### C.1. Data preparation

Our data preparation process consists of multiple steps. In Section 3, we briefly describe our process to extract relevant objects, interactions and events in images and videos. In this section, we augment that description by providing the specific prompts that we used to query GPT-4o to extract the relevant meta-information from the question, ground-truth response and original generated rollouts. The prompt template used for image samples (Figure 5) is almost identical to that used for video samples (Figure 6), except that we also extract information about actions that span multiple frames. In both cases, the LM will return a JSON dictionary where the extracted information is stored in the relevant key-value pairs. Based on the extracted objects and interactions, we use the Molmo-7B model to generate their

## Video Instruction Template

### You are given:

1. The original question: {question}
2. The associated video frames.
3. A model-generated answer: {generated\_text}

### FRAME TIMELINE

- You are given {num\_frames} frames with timestamps (seconds) in chronological order:  
 $\{ \text{formatted\_frame\_timestamps} \}$
- Use timestamps to infer ordering and durations. When reporting actions/interactions, include a time range like  $t_{\text{start}} \rightarrow t_{\text{end}}$  based on visible evidence.
- If frames are discontinuous or sparsely sampled, reason only over the provided times.

### TASK

From the VIDEO FRAMES, extract a concise set of referenced **entities**, **objects**, **actions**, and **interactions** that are **visibly supported in the video** and **relevant to the question and answer**. Emphasize **spatiotemporal evidence**: movements, changes over time, ordering (before→after), and persistent states across frames. If no video is present, fall back to static visual evidence.

### ENTITY RULES

- Entities should be **specific and expressive**, not overly vague.
- Avoid redundant variants describing the same object; merge synonyms/aliases into one expressive label.
- Include distinct parts (e.g., “roots”, “leaves”, “soil”) only if **visually important** on their own or central to the Q/A.
- If >10 entities are visible, include only the **10 most salient** (important, central, or repeatedly emphasized for the Q/A).

### ACTION & INTERACTION RULES (SPATIOTEMPORAL)

- Prefer **motion/change verbs** and **state transitions** (e.g., “seedling grows”, “cup tilts”, “person picks up book”).
- Capture **temporal structure** when visible: start/end states, increasing/decreasing, repetition (“x3”), or ordering (“A→B”).
- Distinguish **object motion vs. camera motion**; do not attribute camera pans/zooms to objects.
- Only include interactions/actions that are **visually grounded across frames** and **support the Q/A**.
- Express concisely (e.g., “roots growing in soil”, “ball rolls left→right and stops”); merge duplicates or equivalents.

### RELEVANCE FILTER

- Include only entities/actions/interactions that **help answer the question or support the given answer**; omit irrelevant background details.

### OUTPUT FORMAT

Return **only** this JSON object (no explanations, no extra text):

```
{
  "entities": [up to 10 expressive, non-redundant entities/objects visible in the video],
  "actions": [concise motion/change events grounded across frames, if any.
              Encode their time ranges as (start_time=Xs, end_time=Ys)],
  "interactions": [unique spatiotemporal relations expressed in concise natural English,
                  if any]
}
```

Figure 6. Instruction template for extracting entities and interactions from generated video-level rollouts.

2D point coordinates. Note that Molmo-7B normalizes the 2D coordinates between 0 and 100 while the base Qwen2.5-VL 7B model represents pixel coordinates in absolute units. After rescaling the generated pixel coordinates, we overlay these coordinates on images and sampled video frames.

### C.2. Reasoning trace generation

To generate rollouts, we primarily use the GLM-4.1V 9B reasoning model. As mentioned in the main paper, we prompt the model to ground its reasoning in the original visual input, while using the 2D point coordinates to disambiguate objects in images and to specify frame-level or

multi-frame events in videos. For each sample, we generate 8 rollouts with a temperature of 0.6 and a maximum of 6144 new tokens. We use the prompt templates in Figure 8 and 9 to extract generated 2D points, timestamps and their corresponding descriptions in images and videos, respectively.

### C.3. Verification and filtering

Although the curation pipeline produces grounded reasoning traces, many rollouts remain unreliable, with only  $\sim 3.1\%$  being usable. To address this, Argos is also used to score and filter rollouts, discarding samples whose best score falls below a threshold while still enforcing outcome

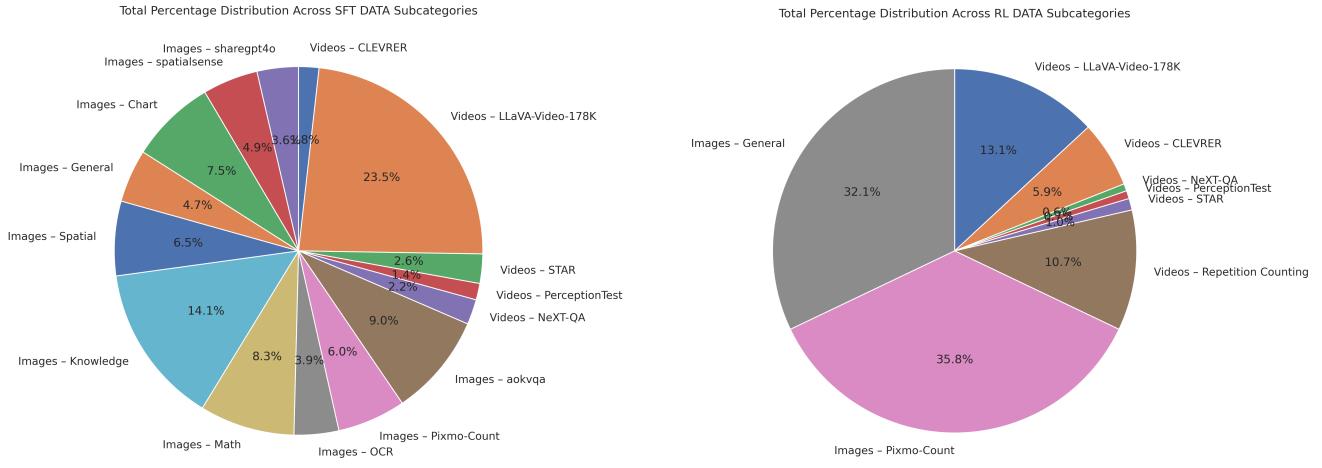


Figure 7. **Training data mixtures.** We use around 85K and 4.6K SFT and RL training samples, respectively.

### Point–Object Extraction Prompt

You will be given a multimodal reasoning trace that includes text with references to visual observations and 2D points in  $(x, y)$  format.

Your task is to extract all 2D point coordinates in the text, and for each one, identify the most semantically relevant noun phrase, object mention, or referring expression that is closely associated with that point. This may include specific items seen in the image, categories, or localized objects mentioned in the same or nearby sentences.

Return your output as a list of mappings in the following format:

```
[{"point": [x, y], "object": "most relevant noun phrase or object"},  
 ...]
```

Do not include any other text.

#### Guidelines:

- Only consider 2D points mentioned explicitly in the text (e.g., (190, 247)).
- Prefer noun phrases that refer to concrete objects or regions localized in the image (e.g., “notebooks”, “mechanical pencils”).
- If multiple possible objects are nearby in the text, pick the one most likely referred to by the point based on sentence structure, context, or positional cues.
- Do not hallucinate new coordinates or object names not mentioned in the original text.
- For each point, ensure the associated object is specific and descriptive, avoiding vague terms like “thing” or “item”. It should never be null.

Here is the reasoning trace: {text}

Figure 8. We use this prompt template to query a LM to extract the generated 2D points and their corresponding object or noun phrase.

accuracy. In practice, we use a threshold value of 0.7. We further clean and standardize extracted 2D points and temporal phrases. This filtering ensures that the SFT dataset

contains primarily high-quality, visually grounded, and semantically accurate reasoning examples.

#### C.4. SFT and RL training data mixtures

We plot the distributions of our final curated training data mixtures in Figure 7. The final set of SFT training samples numbers approximately 85K images and videos . During the RL stage, we use a much smaller subset of around 4.5K images and videos from the Video-R1-260K full split. Our curated datasets contains both video and image samples. The selected videos consist of diverse open-domain clips depicting everyday scenarios, aimed at improving temporal understanding and event reasoning. On the image side, we include general visual QA for basic perception, alongside more task-specific skills such as interpreting scientific figures for quantitative reasoning, reading embedded text via OCR tasks and performing visual commonsense reasoning. Finally, we incorporate some spatial reasoning samples to help the model learn to reason not just about what is present, but where and how it is arranged.

#### D. Agentic verifier scoring functions

The evaluation processes of the visual grounding accuracy of generated responses for images and videos differ slightly. For images, we observe that using a simple regex expression is sufficient to extract 2D points and their corresponding objects. In video reasoning traces, we prompt GPT-4o to extract 2D points as well as timestamps using the template in Figure 10. For each trace, the returned dictionary contains the following lists:

1. spatiotemporal points with 2D point coordinates and relevant frames
2. frame-level descriptions with relevant frames
3. segment-level descriptions with the start and end frames.

The generated spatiotemporal points are scored using the same process as images, where we describe how we use the visual grounding teacher models in Section 3. In the case of a frame-level description, we first encode its corresponding frame before querying the GLM-4.5V model to assign a binary score using the template in Figure 11. Similarly, we query the GLM-4.5V model to evaluate the visual-semantic accuracy between the encoded set of frames and the generated description using the template in Figure 12. Finally, we compute a final visual grounding accuracy score by first averaging the scores within each category before averaging over them.

## Video Point-Event Extraction Prompt

You will be given a multimodal **video** reasoning trace that may include:

- Explicit 2D points written as `(x, y)` or `[x, y]`
- Temporal anchors like `[F<id> @ t=<seconds>s]`
- Temporal lists like `[F1 @ t=0.00s, F2 @ t=1.66s, F3 @ t=3.31s, ...]`
- Temporal spans like `[F<start>-F<end> @ t=<start>-<end>s]`
- Mentions of objects (nouns/referring expressions), actions/events, and single-frame observations

### YOUR TASK

Return **STRICT JSON** with three top-level arrays: `"observations"`, `"events"`, `"points"`. Map extracted items as follows (this mapping is REQUIRED):

- **"points"**: ONLY items that are a point+anchor **as an exact substring** in the text → `"(x, y)"` in `[Fk @ t=Ts]` or `"[Fk @ t=Ts]"` at `(x, y)`.
- **"events"**: ONLY multi-frame spans `"[Fstart-Fend @ t=Tstart-Tend]"` or temporal lists `"[F1 @ t=T1, F2 @ t=T2, ...]"`.
- **"observations"**: ONLY standalone single anchors `"[Fk @ t=Ts]"` (no point in same substring) or standalone points `"(x, y)"` / `"[x, y]"` (no timestamp in same substring).

### ANCHOR TEXT — MUST BE AN EXACT SUBSTRING (VERBATIM)

For every record, return `"anchor_text"` as a single string that is an exact substring copied verbatim from the provided text. Do *not* rewrite or synthesize connectors.

Allowed formats for `"anchor_text"` (choose exactly one per record):

- `"(x, y)"` in `[Fk @ t=Ts]` or `"[Fk @ t=Ts]"` at `(x, y)` (goes to `"points"`)
- `"[Fstart-Fend @ t=Tstart-Tend]"` (goes to `"events"`)
- `"[F1 @ t=T1, F2 @ t=T2, F3 @ t=T3, ...]"` (goes to `"events"`)
- `"[Fk @ t=Ts]"` (goes to `"observations"`)
- `"(x, y)"` or `"[x, y]"` (goes to `"observations"`)

### CRITICAL EXAMPLE (exactness requirement)

Source text: For example, “In `[F17 @ t=19.21s]`, pallbearers at coordinates `(140, 679)`, `(290, 554)`, and `(654, 544)` maintain a steady stance.”

Valid `"anchor_text"`:

- `"[F17 @ t=19.21s]"` (observation)
- `"(140, 679)"` (observation)
- `"(290, 554)"` (observation)
- `"(654, 544)"` (observation)

### STRUCTURES TO RETURN

#### 1) OBSERVATIONS (ONLY Format 4 or 5)

```
{  
    "anchor_text": "<exact substring (Format 4 or 5)>",  
    "frame": "Fk_or_null",  
    "time_s": "Ts_or_null",  
    "description": "<short phrase nearest to anchor_text describing what is visible/occurring>",  
    "object": "<noun phrase/object or null>",  
    "points": "[[x, y], ...]  
}
```

#### 2) EVENTS (ONLY Format 2 or 3)

```
{  
    "anchor_text": "<exact substring (Format 2 or 3)>",  
    "event": "<action/event phrase>",  
    "frames": [F_start, F_end],  
    "times_s": [t_start, t_end],  
    "points": "[[x, y], ...]  
}
```

#### 3) POINTS (ONLY Format 1)

```
{  
    "anchor_text": "<exact substring (Format 1)>",  
    "point": "[x, y]",  
    "object": "<most relevant noun phrase or object>",  
    "frame": "Fk",  
    "time_s": "Ts  
}
```

**RETURN FORMAT (STRICT JSON ONLY; NO EXTRA TEXT)** and **GUIDELINES** can be added in the same way with paragraphs + itemize.

Here is the reasoning trace: {text}

Figure 9. We use this prompt template to query a LM to extract the generated 2D points and their corresponding object or noun phrase, as well as timestamps for video reasoning traces.

## Video Scoring Prompt

You are an information extraction model. Read the reasoning text and extract three categories of temporally grounded facts. Follow the rules exactly.

### DEFINITIONS

#### 1) spatiotemporal

- Target: every 2D point tag of the form <points ...>...</points>.
- For each <points> tag, produce ONE item with:
  - "frame": the nearest explicitly stated single frame number (case-insensitive "frame N") in the same sentence or immediately preceding clause, if present; else null.
  - "time": the nearest explicitly stated single absolute time in seconds (e.g., "36.31 seconds", "t=36.31s", "36.31s") in the same sentence or immediately preceding clause, if present; else null.
  - "object": the object name from the tag's inner text if non-empty, otherwise from its alt attribute. This has to be an object and not timestamps or frame numbers.
  - "x": the x coordinate from the tag.
  - "y": the y coordinate from the tag.
- Do NOT infer values; only use explicitly stated numbers. If both a frame and a time are given, include both; otherwise set the missing field to null.
- If multiple <points> tags are present near the same mention, output one item per tag.

#### 2) frame-level temporal

- Target: any mention tied to a SINGLE frame (e.g., "frame 6") and/or a SINGLE time (e.g., "4.47 seconds"). IMPORTANT: This is similar to the spatiotemporal case but used when no <points> tag is present. No 2D coordinates are involved here and there should not be any overlap with the extracted spatiotemporal points.
- Output an item with:
  - "frame": that single frame number, or null if none.
  - "time": that single absolute time in seconds, or null if none.
  - "description": an EXACT substring copied verbatim from the input text that describes what happens at that frame/time. Do NOT paraphrase or invent wording; choose the minimal contiguous snippet that fully states the observation.
- Do NOT include <points> tags themselves inside "description" unless they are part of the original wording you copy.

#### 3) segment-level temporal

- Target: any mention that spans MULTIPLE frames (e.g., "frames 1–6", "frame 7 through frame 20") and/or a TIME RANGE (e.g., "4.47 to 16.11 seconds").
- Output an item with:
  - "start\_frame": first frame number if given, else null.
  - "end\_frame": last frame number if given, else null.
  - "start\_time": start time in seconds if given, else null.
  - "end\_time": end time in seconds if given, else null.
  - "description": an EXACT substring copied verbatim from the input that describes the segment-level event. Do NOT paraphrase.
- Open-ended ranges like "frame 25 onwards" are allowed: fill known start\_\* and set unknown end\_\* fields to null.

### NORMALIZATION & MATCHING RULES

- Frames: integers only. Match case-insensitively for the word "frame" (e.g., "Frame 6", "frame 6").
- Times: floats in seconds; accept forms like "36.31 seconds", "t=36.31s", "36.31s". Output them as numbers (no units) rounded to two decimals.
- Ranges: recognize hyphen/en dash ("1-6", "1–6"), "from ... to ...", and "through". For "onwards", only the start\_\* is known.
- Association for spatiotemporal points: prefer the closest explicit single frame/time in the same sentence or immediately preceding clause. If only a range is present and no single value is explicitly tied to the point, set both "frame" and "time" to null (do NOT invent).
- Deduplicate identical items across lists.
- Use null where a field cannot be filled **without explicit evidence**.
- Do NOT hallucinate any text. For the "description" fields in (2) and (3), the value MUST be an exact substring found in the input text.

### OUTPUT FORMAT (STRICT)

Return ONLY a JSON object with exactly these keys:

- "spatiotemporal": a list of objects of the form "frame": <int|null>, "time": <float|null>, "object": <string>, "x": <string>, "y": <string>
- "frame\_level\_temporal": a list of objects of the form "frame": <int|null>, "time": <float|null>, "description": <string>
- "segment\_level\_temporal": a list of objects of the form "start\_frame": <int|null>, "end\_frame": <int|null>, "start\_time": <float|null>, "end\_time": <float|null>, "description": <string>

If a category has no items, return an empty list for that key. TEXT

{reasoning\_text}

## Video frame-level accuracy

You will receive ONE image (a single video frame) and a short description.

### TASK

Score how well the description matches ONLY the visible content of the image. Ignore unverifiable timing info (timestamps/frame indices) in the text; treat them as neutral metadata. Focus on objects, attributes, and spatial relations that can be visually verified in this single frame.

### SCORING (0.0 or 1.0)

- **1.0** All key claims are clearly supported; no contradictions.
- **0.0** Contradicted or refers to things not visible in the frame.

### RULES

- Judge ONLY this single image (no assumptions from outside the frame).
- If a relation like “near the closet door” is claimed, require both entities to be visible and the relation plausible in the image.
- Be conservative when evidence is unclear; do not infer beyond what is visible.

### OUTPUT (JSON only, no extra text)

```
{"score": <float 0 - 1>}
```

### DESCRIPTION

”{description}”

Figure 11. We use this template to prompt the teacher model to assign a binary score based on the relevance of the generated frame-level description and the visual content in the corresponding video frame.

## Video segment-level accuracy

You will receive an ORDERED sequence of images (consecutive video frames) and a short event/action description.

### TASK

Score how well the description matches ONLY what is visually supported across the entire frame sequence. Evaluate both the action semantics and the temporal extent (start→middle→end). Assume that the provided frames correspond to the mentioned timestamps and rely primarily on visible evidence and ordering of frames.

### INPUTS

- **FRAMES:** An ordered list of frames F1..FN. Assume they are evenly spaced.
- **DESCRIPTION:** “{description}”

### EVALUATION GUIDELINES

- **Visual Evidence:** Are the required entities/objects present? Is the claimed action (e.g., “cutting the cake”) visibly happening?
- **Temporal Progression:** Do frames show a plausible sequence for that action (setup → interaction → outcome)? Look for state changes (e.g., knife touches cake, slices appear).
- **Extent & Alignment:** Does the action substantially occur within the claimed start/end span? Small off-by-one frame or slight timing drift is minor; clear mismatches are penalized.
- **Consistency:** Penalize claims contradicted by any clear frame (e.g., no knife ever shown, or subject doing a different action).
- **Single-Sequence Scope:** Judge ONLY these frames; don’t assume anything outside the provided sequence.

### SCORING (0.0 or 1.0)

- **1.0 – Fully supported:** the action is clearly visible and progresses as described, with onset/offset aligned within a small tolerance; no contradictions.
- **0.0 – Contradicted:** action not happening, key objects missing, or description clearly mismatched.

### OUTPUT (JSON only, no extra text)

```
{"score": <float 0 - 1>}
```

Figure 12. We use this template to prompt the teacher model to assign a binary score based on the relevance of the generated segment-level description and the visual content in the corresponding video segment.

## E. Full theoretical justification

Consider any prompt/input to the model, each possible answer/action  $a \in \mathcal{A}$  has  $m$  true reward components  $R_1(a), \dots, R_m(a)$ , measuring its quality from different aspects (e.g. grounding, reasoning, etc.). The multiple rewards can arbitrarily correlate with each other, and the goal is to learn towards Pareto optimality defined as follows:

**Definition 1** ( $\delta$ -Pareto Domination). *For  $\delta > 0$ , we say  $a' \succ_\delta a$  if  $R_i(a') \geq R_i(a) + \delta$  for all  $i = 1, \dots, m$ .*

**Definition 2** ( $\delta$ -Pareto Optimality). *For  $\delta > 0$ , the set of globally  $\delta$ -Pareto-optimal actions*

$$P_\delta := \{a \in \mathcal{A} : \nexists a' \in \mathcal{A} \text{ s.t. } a' \succ_\delta a\}.$$

However, in practice we do not have access to the true rewards  $R_i(a)$ 's, but instead weak estimators  $\hat{R}_i(a) = R_i(a) + \varepsilon_i(a)$ . The estimated rewards can be inaccurate and correlated, with the only assumption as follows:

**Assumption 1.** *For any  $a$ , the noise variables  $\{\varepsilon_i(a)\}_{i=1}^m$  are independent, mean-zero, and  $\sigma$ -sub-Gaussian:*

$$\mathbb{E}[\varepsilon_i(a)] = 0, \quad \mathbb{E}[e^{\lambda\varepsilon_i(a)}] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

The  $\sigma$ -sub-Gaussian assumption is a generic form constraining the error, which is easily satisfied by Hoeffding's lemma when the rewards are bounded. In each RFT step,

$$\hat{R}(a) = \sum_{i=1}^m w_i \hat{R}_i(a) = R(a) + \sum_{i=1}^m w_i \varepsilon_i(a),$$

where  $R(a) = \sum_{i=1}^m w_i R_i(a)$ , and  $w_1, \dots, w_m > 0$ . Denote  $w_{\min} = \min_i w_i$  and  $w_{\max} = \max_i w_i$ . We can show probability bound on the estimated  $\delta$ -Pareto optimality.

**Lemma 1.** *Fix two actions  $a, a'$  with  $a' \succ_\delta a$ , then we have*

$$\mathbb{P}(\hat{R}(a) \geq \hat{R}(a')) \leq \exp\left(-\frac{\delta^2}{4\sigma^2} m \frac{w_{\min}^2}{w_{\max}^2}\right).$$

*Proof.* If  $a' \succ_\delta a$  then  $R(a') - R(a) \geq \delta \sum_i w_i \geq \delta m w_{\min}$ . Let  $Z = \sum_i w_i (\varepsilon_i(a') - \varepsilon_i(a))$ , then we have

$$\mathbb{E}[Z] = \sum_i^m w_i (\mathbb{E}[\varepsilon_i(a')] - \mathbb{E}[\varepsilon_i(a)]) = 0,$$

and  $\forall \lambda \in \mathbb{R}$ ,

$$\begin{aligned} \mathbb{E}[e^{\lambda Z}] &= \mathbb{E}[e^{\lambda \sum_i w_i (\varepsilon_i(a') - \varepsilon_i(a))}] \\ &= \prod_i^m \mathbb{E}[e^{\lambda w_i \varepsilon_i(a)}] \mathbb{E}[e^{\lambda w_i \varepsilon_i(a')}] \\ &\leq e^{m w_{\max}^2 \sigma^2 \lambda^2}. \end{aligned}$$

Thus  $Z$  is sub-Gaussian with proxy variance bounded by  $2m\sigma^2 w_{\max}^2$ . Now since

$$\hat{R}(a') - \hat{R}(a) = (R(a') - R(a)) + Z \geq \delta m w_{\min} + Z,$$

we can bound

$$\begin{aligned} \mathbb{P}(\hat{R}(a) \geq \hat{R}(a')) &\leq \mathbb{P}(Z \leq -\delta m w_{\min}) \\ &\leq \exp\left(-\frac{\delta^2}{4\sigma^2} m \frac{w_{\min}^2}{w_{\max}^2}\right). \end{aligned}$$

□

Next, we study optimization of  $\hat{R}(a)$  over a finite sample group  $\mathbb{G} = \{a_1, \dots, a_n\}$  drawn from policy  $\pi$ .

**Lemma 2** (Batch-Level Approximate Pareto Preservation). *Let  $\mathbb{G} = \{a_1, \dots, a_n\}$  and  $\hat{a} = \arg \max_{a \in \mathbb{G}} \hat{R}(a)$ , we have*

$$\mathbb{P}(\exists a' \in \mathbb{G} \text{ with } a' \succ_\delta \hat{a}) \leq (n-1) \exp\left(-\frac{\delta^2}{4\sigma^2} m \frac{w_{\min}^2}{w_{\max}^2}\right).$$

*Proof.* By Lemma 1, each pair  $(\hat{a}, a')$  for  $a' \neq \hat{a}$  violates the ordering with probability less than  $\exp\left(-\frac{\delta^2}{4\sigma^2} m \frac{w_{\min}^2}{w_{\max}^2}\right)$ . Simply applying a union bound over the  $n-1$  other samples yields the probability bound

$$(n-1) \exp\left(-\frac{\delta^2}{4\sigma^2} m \frac{w_{\min}^2}{w_{\max}^2}\right)$$

□

With the above lemmas, we can guarantee global  $\delta$ -Pareto optimality over the entire possible action space.

**Theorem 1** (Global Pareto Guarantee). *Let  $\pi$  be the sampling policy and denote  $\beta = \pi(P_\delta)$  as the probability coverage on Pareto optimal solutions. Sample  $n$  i.i.d. actions from  $\pi$  to form a group  $\mathbb{G}$  with  $\hat{a} = \arg \max_{a \in \mathbb{G}} \hat{R}(a)$ . Then*

$$\mathbb{P}(\hat{a} \in P_\delta) \geq (1 - (1 - \beta)^n) \left[1 - \frac{n-1}{e^{C \cdot m}}\right],$$

where  $C := \frac{\delta^2}{4\sigma^2} \cdot \frac{w_{\min}^2}{w_{\max}^2} > 0$  is a constant.

*Proof.* With probability  $1 - (1 - \beta)^n$ , the batch contains at least one element of  $P_\delta$ . Conditioned on that event, Lemma 2 bounds the probability that the contained optimal action is not selected. Multiplying the two probabilities yields the stated bound. □

The theorem shows that as the number of rewards  $m$  increases, we can approximate global Pareto optimality even with weak estimators.

**Corollary 1** (Additional Correctness Reward). *Let  $R_0(x)$  denote a dominant outcome correctness reward and  $R_1(a), \dots, R_m(a)$  the reasoning rewards. For a threshold  $\tau$  and margin  $\gamma > 0$ , define the gated scalarization*

$$\hat{R}_\tau(a) = \begin{cases} \hat{R}_0(a), & \hat{R}_0(a) < \tau, \\ w_0 \hat{R}_0(a) + \sum_{i=1}^m w_i \hat{R}_i(a), & \hat{R}_0(a) \geq \tau, \end{cases}$$

where all weights  $w_i > 0$ , including  $w_0$  for outcome in the high-correctness regime.

Assume that all actions satisfy  $R_0(a) \notin (\tau - \gamma, \tau + \gamma)$ . Let  $P_{\delta,\tau}$  be the set of actions that are (i) correct ( $R_0(a) \geq \tau + \gamma$ ) and (ii)  $\delta$ -Pareto-optimal in the reasoning rewards. Then, when optimizing  $\hat{R}_\tau(a)$  over a batch  $\mathbb{G}$  of size  $m$ ,

$$\mathbb{P}(\hat{a} \in P_{\delta,\tau}) \geq (1 - (1 - \beta)^n) \left(1 - ne^{-\frac{\gamma^2}{2\sigma^2}}\right) \left[1 - \frac{n-1}{e^{C \cdot (m+1)}}\right],$$

where  $\beta = \pi(P_{\delta,\tau})$  is the policy mass on truly  $\delta$ -Pareto-optimal, high-correctness actions.

*Proof.* We perform similar analysis by conditioning the joint events. As in Theorem 1, the probability that  $\mathbb{G}$  contains at least one action  $a^* \in P_{\delta,\tau}$  is  $1 - (1 - \beta)^n$ .

Then we bound the probability of ‘‘gating error,’’ defined as the event where noise  $\varepsilon_0(a)$  causes an action to cross the threshold  $\tau$ . Given the gap assumption  $|R_0(a) - \tau| > \gamma$ , a classification error occurs only if  $|\varepsilon_0(a)| > \gamma$ . By the sub-Gaussian assumption:

$$\mathbb{P}(|\varepsilon_0(a)| > \gamma) \leq e^{-\frac{\gamma^2}{2\sigma^2}}.$$

Applying a union bound over all  $n$  actions in the batch, the probability that all actions are correctly gated is at least

$$1 - ne^{-\frac{\gamma^2}{2\sigma^2}}.$$

Conditioned on correct gating, the optimization for high-correctness actions is determined by the weighted sum of  $m+1$  rewards (including  $R_0$ ). For any  $a' \in \mathbb{G}$  that is dominated by  $a^* \in P_{\delta,\tau}$  in the  $(m+1)$ -dimensional sense, we apply Lemma 2 adapted for  $m+1$  dimensions. The probability that the approximate optimizer  $\hat{a}$  is not dominated is bounded using the constant  $C$  with dimension  $m+1$ :

$$1 - \frac{n-1}{e^{C(m+1)}}.$$

Combined together, we obtain the lower bound:

$$(1 - (1 - \beta)^n) \left(1 - ne^{-\frac{\gamma^2}{2\sigma^2}}\right) \left[1 - \frac{n-1}{e^{C \cdot (m+1)}}\right].$$

□

**Remark 1** (GRPO Training). *In GRPO, the normalized advantage  $A_i = (R_i - \text{mean}(\{R_j\})) / \text{std}(\{R_j\})$  preserves the same ordering as the aggregated rewards  $R_i$ . Hence, the Pareto-optimality and ordering guarantees established above apply directly to GRPO updates. The groupwise mean and standard deviation serve only as variance-reduction terms, ensuring stable gradients without affecting which samples are reinforced. Consequently, Pareto-optimal and high-correctness responses receive positive advantages with high probability.*

**Remark 2** (Optimal Sample Size). *The lower bound in Theorem 1,*

$$H(n) = \underbrace{(1 - (1 - \beta)^n)}_{\text{exploration}} \cdot \underbrace{\left[1 - \frac{n-1}{e^{Cm}}\right]}_{\text{exploitation}}$$

shows an inherent exploration–exploitation tradeoff: larger  $n$  improves policy coverage (exploration) through  $1 - (1 - \beta)^n$ , but also increases the chance of misranking due to the term  $1 - \frac{n-1}{e^{Cm}}$ . Maximizing  $H(n)$  yields an approximate optimal batch size

$$n^* \approx \frac{1}{\beta} \log(1 + \beta e^{Cm}),$$

which balances exploration and estimation reliability. The optimal batch size  $n^*$  decreases with  $\beta$  (the probability of generating a Pareto-optimal action), which encourages larger sample size when the model is not good enough (exploration), and relatively smaller sample size as the model is concentrated on good responses (exploitation).

Model	<b>BLINK</b>	<b>MindCube-t</b>	<b>CV-Bench</b>	<b>CV-Bench (3D)</b>
Argos (Ours)	<b>57.6</b>	<b>37.2</b>	<b>79.5</b>	<b>82.3</b>
- RQ	56.9	36.3	76.9	78.5
- VG	55.8	36.3	77.5	80.5

Table 6. Ablation results on spatial reasoning benchmarks.

Model	CounterCurate	HallusionBench*	SugarCrepe
Argos (Ours)	81.9	<b>49.1</b>	<b>88.0</b>
- RQ	81.9	48.7	87.5
- VG	<b>82.5</b>	48.0	86.7

Table 7. Ablation results on visual hallucination benchmarks. \*We note that HallusionBench typically uses the GPT-4 Turbo as a grader but this model got deprecated after the main paper deadline. As such, we run this ablation by replacing GPT-4 Turbo with GPT-4.1 as the grader and may lead to higher performances on average than before.

Model	Base	Common	Complex	Visual	Spatial	Long	Avg
Argos (ours)	<b>42.0</b>	10.7	<b>26.0</b>	<b>15.3</b>	14.7	4.0	<b>18.8</b>
- RQ	36.0	10.7	<b>26.0</b>	12.7	<b>15.3</b>	6.0	17.8
- VG	41.3	10.7	25.3	11.3	<b>15.3</b>	<b>6.7</b>	18.4

Table 8. Ablation results on EB-Habitat.

## F. Additional ablation experiments

In this section, we further ablate the benefits of adding the different reward terms during the RL training stage in the paper across the spatial reasoning (Table 6), visual hallucination (Table 7) and embodied AI benchmarks (Table 8). We conduct a much smaller scale ablation with fewer numbers of steps to analyze the reasoning quality (RQ) and visual grounding (VG) reward terms for MMRL on a small subset of the Pixmo-Count [8] dataset combined with another subset of the Video-R1 [13] image split. In total, our ablation RL training set numbers about 1.5K samples. Starting from the same SFT checkpoint finetuned with our curated data, we train three variants by removing the reasoning quality reward term and visual grounding reward terms in sequence. In general, we observe that adding the VG reward term is beneficial for improving performance across visual hallucination, spatial reasoning and embodied tasks. Additionally, we also see that adding the RQ reward term can also be helpful to further improve performance on such agentic benchmarks. Interestingly, using only the outcome reward term actually leads to the best performance on the CounterCurate benchmark in this ablation study. However, CounterCurate only evaluates on up/down and left/right questions is much less complex than HallusionBench and SugarCrepe. On the latter datasets which evaluates multiple axes of visual hallucination including removal and ad-

dition of objects, the results demonstrate that adding RQ and VG reward terms can be helpful.

## G. Qualitative visualizations

In this section, we provide some qualitative visualizations of our curated SFT reasoning traces. For image samples, we highlight any generated 2D points in the reasoning traces and also overlay them as red circles onto the images for visibility. We provide both original and the overlaid images for comparison but we note that we do not actually use the overlaid images during the SFT stage.

### Image Example 1

**Question:**

Is there a vehicle with a spare tire visible on its exterior? A. No. B. Yes.

**Response:**

<think>Let me analyze the image carefully while identifying the vehicle with a spare tire.

Scanning the scene, I need to look for vehicles that show a spare tire on their exterior. I'll move systematically through the image looking at each vehicle.

I notice a dark-colored jeep or utility vehicle on the right side of the image, which appears to have a spare tire mounted on its exterior. Let me focus on the exact location of this vehicle.

Looking at the right side of the image, I can see a vehicle that appears to be a jeep or utility vehicle. Upon closer inspection, I can confirm there's a spare tire mounted on the back of this vehicle, which is clearly visible. This appears to be the vehicle with the spare tire on its exterior.

I observe a jeep type vehicle at approximately <points x1="408" y1="179" alt="jeep or utility vehicle">jeep or utility vehicle</points> with a spare tire visible at its rear. The spare tire appears to be mounted on the exterior of this vehicle, making it clearly visible.

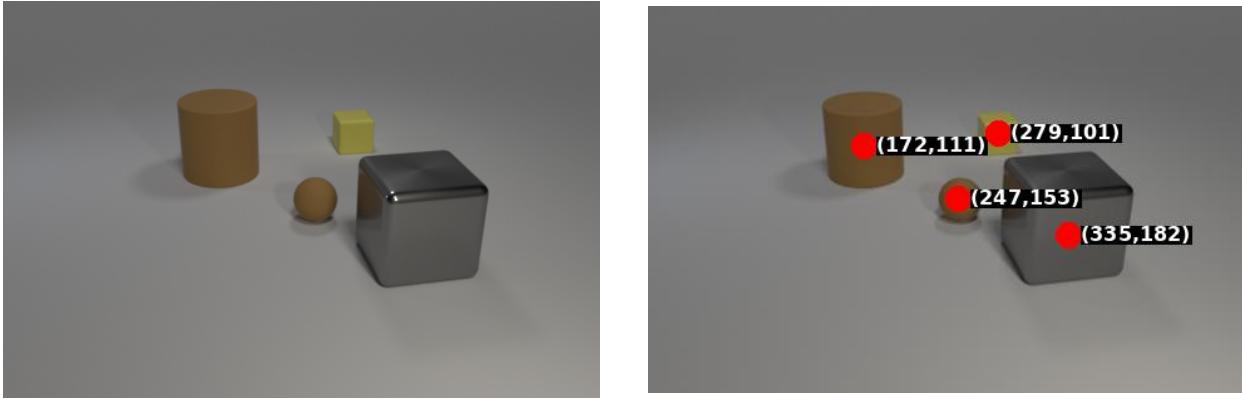
Therefore, the answer should be B. Yes. There is a vehicle with a spare tire visible on its exterior.</think><answer>I examine the scene systematically from left to right.

As I scan the right side of the image toward the middle-right area, I notice a dark-colored utility vehicle that appears to be a jeep type. On its exterior, I can clearly see a spare tire mounted on the back of the vehicle. This spare tire is visible and prominently displayed on the side of the vehicle's exterior structure. The vehicle is positioned near the right edge of the scene, traveling on the right side of the road. I confirm this is indeed a vehicle with a spare tire visible on its exterior.

<|begin\_of\_box|>B. Yes.<|end\_of\_box|>

Figure 13. Image example 1

## Image Example 2



### Question:

How many objects are there in total?

### Response:

<think>Looking at the original image, I need to count and identify all objects. Starting from the left side, I see a large brown cylinder positioned at around the left-center area. Moving rightward, there is a small brown sphere near the center of the image. Continuing further right, there is a gray metallic cube. At the top right area, I see a small yellow cube.

To confirm with spatial awareness, I scan from left to right: the large brown cylinder is at <points x1="172" y1="111" alt="large brown cylinder">large brown cylinder</points> area, the small brown sphere is positioned closer to the center around <points x2="247" y2="153" alt="small brown sphere">small brown sphere</points>, the large gray metallic cube is at <points x3="335" y3="182" alt="large gray metallic cube">large gray metallic cube</points>, and the small yellow cube is positioned at the top right area at about <points x4="279" y4="101" alt="small yellow cube">small yellow cube</points>.

Counting these objects, there are 4 distinct objects total. I observe a large brown cylinder at <points x1="172" y1="111" alt="large brown cylinder">large brown cylinder</points>, a small brown sphere at <points x2="247" y2="153" alt="small brown sphere">small brown sphere</points>, a large gray metallic cube at <points x3="335" y3="182" alt="large gray metallic cube">large gray metallic cube</points>, and a small yellow cube at <points x4="279" y4="101" alt="small yellow cube">small yellow cube</points>. Therefore, there are 4 objects in total.</think><answer>Looking at the scene, I observe a large brown cylinder at position <points x1="172" y1="111" alt="large brown cylinder">large brown cylinder</points> on the left side. Moving rightward toward the center, I notice a small brown sphere at <points x2="247" y2="153" alt="small brown sphere">small brown sphere</points>. Continuing further right, there is a large gray metallic cube at position <points x3="335" y3="182" alt="large gray metallic cube">large gray metallic cube</points>. In the upper right area of the image, I see a small yellow cube at position <points x4="279" y4="101" alt="small yellow cube">small yellow cube</points>. Counting all these distinct objects, there are 4 objects in total. </begin\_of\_box|>4<|end\_of\_box|></answer>

Figure 14. Image example 2

### Image Example 3



#### Question:

What is at the top of this structure in the middle of the city square? A. church tower  
B. advertisement. C. latitude. D. optiplex.

#### Response:

<think>Let me analyze the image focusing on the structure in the middle of the city square. I need to identify what's at the top of this structure.

I observe a prominent building in the center of the image which appears to be the structure in question. Looking at the top of this building:

- There's a spire-like structure at the top
- There's a clock visible on what appears to be a tower
- The building has architectural features typical of a church tower or bell tower

Looking more carefully at the top of this central building specifically, I can see that there's a tall spire or tower element which has a distinctive shape typical of church architecture. The building appears to be historic with ornate architectural details around what would typically be a church tower or clock tower.

Now let me consider the answer choices:

- A. church tower - This seems most likely given the architectural style with the spire, the historic nature of the building, and the prominent tower
- B. advertisement - There doesn't appear to be an advertisement at the top, rather architectural elements
- C. latitude - This is a directional coordinate measurement, not something you'd find at the top of a structure
- D. optiplex - This appears to be related to computer equipment, not something that would be at the top of a building

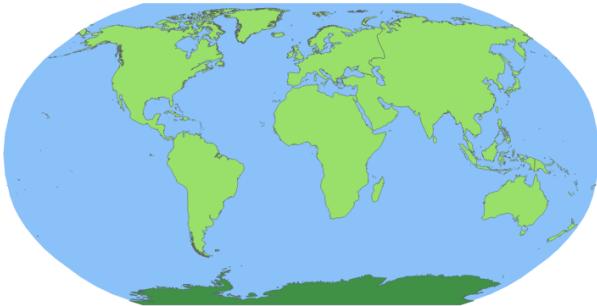
Based on the visual evidence of the spire/tower structure with architectural details like what appears to be a clock, this fits the description of a church tower rather than an advertisement, latitude (which is not applicable), or optiplex (which is irrelevant).

I observe a distinctive spire and tower at the top of the central building around `<points x1="165" y1="156" alt="church tower">church tower</points>` area, which has architectural characteristics typical of a church tower .</think><answer>Looking at the central building in the city square, I can see it has a distinctive tower structure. Examining the top portion of this central structure, I observe a prominent spire-like element at the very top, approximately at the middle-upper section of the beige building with the clock. The architectural style includes ornate details and the tower appears to be an elevated structure with a pointed or tapered top that's characteristic of religious or civic architecture. Given the building's historic appearance with its detailed facade, corner tower, and clock, the structure at the top is most likely a church tower. None of the other options (advertisement, latitude, or optiplex) would naturally be found as a prominent feature at the top of this architectural structure.

My answer is <|begin\_of\_box|>A<|end\_of\_box|>4.</answer>

Figure 15. Image example 3

#### Image Example 4

**Question:**

Which continent is being indicated in the highlighted area? A. Australia. B. Antarctica. C. Africa.

**Response:**

<think>Let me analyze the original image carefully. I'm looking for the highlighted area that indicates a specific continent.

Looking at the original image (first image), I can see a world map with blue oceans and green land masses. There's a distinct green landmass at the bottom of the image that appears to be completely surrounded by water. This is typically Antarctica on a world map.

Now I need to focus on that highlighted area specifically. Looking at the bottom portion of the original image, there is indeed a green landmass at the southernmost part of the map. It appears completely isolated by water, which is characteristic of Antarctica.

Looking at the second image (with coordinate overlays), I can see a coordinate at `<points x1="427" y1="346" alt="green landmass at the southernmost part of the map">`green landmass at the southernmost part of the map`</points>` which appears to be pointing to the green area at the very bottom of the map, confirming that it is the southernmost continent.

Based on the position at the bottom of the world map, completely separated by water from all other continents, I can identify this as Antarctica. This matches with the characteristics of Antarctica being located at the South Pole region.

The answer should be B. Antarctica.</think><answer>I examine the original world map, focusing on the lower portion where there is a distinct green landmass completely surrounded by water at the southernmost region of the map. Looking at this highlighted area, I note that it's positioned at the bottom of the map, separated by ocean from all other continents. This corresponds to the continent located at the South Pole. This geographic characteristic - isolated at the southernmost point of the map, completely surrounded by water - clearly indicates Antarctica. The distinctive positioning and separation from other landmasses confirms this is the proper identification.

<|begin\_of\_box|>B.<|end\_of\_box|></answer>

Figure 16. Image example 4

### Video Example 1

**Question:**

What does the person do after topping the sushi roll with the orange fish mixture?

**Response:**

<think>Got it, let's try to figure out what the person does after topping the sushi roll with the orange fish mixture. First, I need to look at the sequence of frames. The question is about what happens immediately after the topping step.

Looking at the frames, after adding the orange fish mixture (which seems to be done in frames where the topping is visible), the next action would be wrapping it. Let's check the frames where the person is handling the sushi roll. After the topping, the person likely wraps it in plastic wrap or uses a bamboo mat.

Looking at the frames, after the topping (which is done in frames like frame 2, frame 3, frame 4, frame 5, etc.), the next step would be wrapping with plastic wrap. Let's check frame 24 where the person is placing plastic wrap over the sushi roll. Wait, let's check the sequence. The person first tops with the orange mixture, then wraps in plastic wrap, then rolls with bamboo mat. So the action after topping would be wrapping the sushi roll in plastic wrap. Let me confirm the frames. After the topping is done (frames where the orange mixture is on top), the next action is wrapping in plastic wrap, then rolling with the bamboo mat. So the spatiotemporal reasoning would involve tracking the sushi roll's state. After being topped with the orange fish mixture, the person wraps the sushi roll in plastic wrap. Then uses the bamboo mat to roll it.

Now, to make sure I'm citing the correct frames. Let's check the frames where the plastic wrap is used. Frame 24 shows the person placing plastic wrap over the sushi roll, and then frames 25 to 27 show wrapping with plastic wrap. Then frames 28 to 30 show using the bamboo mat to roll the sushi. So the action after topping is wrapping the sushi roll in plastic wrap.</think><answer>To determine what the person does after topping the sushi roll with the orange fish mixture, we analyze the sequence of frames. After applying the orange fish mixture (visible in multiple frames, such as frame 2 at 0.81 seconds where the mixture is at <points x1="355" y1="189" alt="orange fish mixture">orange fish mixture</points> and frame 5 at 3.25 seconds where it's at <points x2="339" y2="169" alt="orange fish mixture">orange fish mixture</points>), the next action involves wrapping the sushi roll. In frame 24 at 18.71 seconds, the person places clear plastic wrap over the sushi roll (at coordinates around <points x3="342" y3="169" alt="plastic wrap">plastic wrap</points>), and subsequent frames from frame 25 at 19.52 seconds to frame 27 show the person continuing to wrap the sushi roll in plastic. This is followed by using a bamboo rolling mat from frame 28 at 21.15 seconds to frame 30 at 23.59 seconds to shape the sushi. Therefore, after topping with the orange fish mixture, the person wraps the sushi roll in plastic wrap. The answer is <begin\_of\_box>Wraps the sushi roll in plastic wrap<end\_of\_box>.</answer>

Figure 17. Video example 1

### Video Example 2

**Question:**

How does the camera move when showing the frying pan on the stove? A. It tilts upwards . B. It pans to the left. C. It zooms in closer to the frying pan. D. It moves away from the frying pan.

**Response:**

<think>Got it, let's analyze the camera movement when showing the frying pan. First, I need to check how the camera's perspective changes. Let's look at the frames with the frying pan.

Looking at the sequence, when the camera focuses on the frying pan, does it zoom in? Let's check the frames. For example, when the oil is poured, the camera might zoom in closer. Wait, let's check the frames where the frying pan is the main focus. The camera seems to move closer, increasing the size of the frying pan in the frame.

Let's check the coordinates. In earlier frames, the frying pan might be smaller, then as the camera zooms in, it becomes larger. For example, comparing **frame 1 at 0.00 seconds** and **frame 4 at 2.53 seconds**, the frying pan's size relative to the frame changes. If it's zooming in, the frying pan would appear larger, meaning the camera is moving closer. So, the camera zooms in closer to the frying pan. Let's confirm. The other options: tilting upwards would change the angle vertically, panning left would move the camera horizontally, moving away would make the pan smaller. The evidence shows the pan becomes larger in the frame, indicating zooming in.

Therefore, the camera moves by zooming in closer to the frying pan.</think><answer>To determine how the camera moves when showing the frying pan on the stove, we analyze the spatiotemporal changes. When focusing on the frying pan, the camera's perspective shifts such that the frying pan appears larger in successive frames. For example, comparing **frame 1 at 0.00 seconds** and **frame 4 at 2.53 seconds**, the frying pan at <points x1="356" y1="273" alt="frying pan">frying pan</points> in frame 1 appears smaller relative to the frame than the frying pan at <points x2="356" y2="262" alt="frying pan">frying pan</points> in frame 4. This indicates the camera is moving closer to the frying pan. Additionally, as the sequence progresses, the frying pan's size increases in the frame, confirming that the camera zooms in closer rather than tilting, panning, or moving away. Therefore, the camera moves by zooming in closer to the frying pan. <|begin\_of\_box|>C<|end\_of\_box|>.</answer>

Figure 18. Video example 2