

ReMoT: Reinforcement Learning with Motion Contrast Triplets

Supplementary Material

1. Motion Contrast Construction Details

This appendix provides comprehensive implementation details for constructing motion contrast triplets across the three primary domains defined in Table 1 of the main paper: **Camera Navigation**, **Robot Manipulation**, and **Object-Centric Motion** (covering Tracking, Grounding, and Counting). We detail the instantiation of $(\mathcal{M}, g, \phi, \mathcal{N})$ for each domain.

1.1. Camera Navigation: ScanNet and Nuscenes

Dataset Overview. ScanNet and NuScenes provide video sequences with camera pose annotations. Each frame I_t is associated with a global pose $T_t \in SE(3)$.

Temporal Strided Sampling. Unlike standard video training which often uses adjacent frames, camera motion between I_t and I_{t+1} is often negligible. To ensure perceptually significant motion, we employ **strided sampling**. Given an anchor frame I_i , we search for a positive frame I_j within a temporal window $j \in [i + \delta_{\min}, i + \delta_{\max}]$ (typically 0.5s to 2.0s) that satisfies the motion magnitude thresholds defined below.

Property Space \mathcal{M} and Ground Truth g . We focus on six discrete camera movements

$$\mathcal{M} = \{\text{left, right, up, down, forward, backward}\} \quad (6 \text{ discrete camera movements}) \quad (1)$$

We compute relative transformation $\Delta T = T_i^{-1}T_j$. For rotations, we convert the rotation matrix to axis-angle (\mathbf{v}, θ) .

$$g_{\text{rot}}(\Delta R) = \begin{cases} \text{left} & \text{if } \mathbf{v}_y > 0.7 \text{ and } \theta > \theta_{\text{thresh}} \\ \text{right} & \text{if } \mathbf{v}_y < -0.7 \text{ and } \theta > \theta_{\text{thresh}} \\ \dots & \text{(similar for up/down)} \end{cases} \quad (2)$$

Visual Saliency Filter ϕ . To filter out ambiguous pairs, we enforce:

$$\phi(m) = \begin{cases} 1 & \text{if } 10^\circ \leq \theta \leq 50^\circ \text{ (ensures visibility but prevents total scene change)} \\ 1 & \text{if } 0.5\text{m} \leq \|\Delta \mathbf{t}\| \leq 2.0\text{m} \text{ (for translation)} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Negative Constructor \mathcal{N} (Spatial Illusion). As detailed in the main text, we synthesize negatives via **Diagonal Cropping**. For a ‘‘Camera Rotate Left’’ property, the Anchor I_i is a bottom-right crop, and the Negative I_k^- is a top-left crop from the *same* original frame. This creates a visual illusion of opposite rotation parallax while maintaining identical semantic content, forcing the model to rely on the actual view change in the Positive I_j .

1.2. Robot Manipulation: AgiBot

Dataset Overview. AgiBot contains teleoperated demonstrations. Annotations include end-effector position $\mathbf{p}_t \in \mathbb{R}^3$ and gripper state $s_t \in \{0, 1\}$.

Property Space \mathcal{M} . We define composite properties: $\mathcal{M} = \mathcal{M}_{\text{gripper}} \times \mathcal{M}_{\text{motion}}$.

- $\mathcal{M}_{\text{gripper}} = \{\text{opening, closing, maintaining}\}$
- $\mathcal{M}_{\text{motion}} = \{\text{moving left, moving right, } \dots, \text{moving lower-right}\}$ (8 directions)

Ground Truth g and Filter ϕ . Motion is categorized by the dominant axis of displacement $\Delta \mathbf{p} = \mathbf{p}_j - \mathbf{p}_i$. We apply a saliency filter to ensure motion is visible:

$$\phi(m) = 1 \iff (\|\Delta \mathbf{p}\| \geq 0.05\text{m}) \vee (s_i \neq s_j \text{ and gripper visible}) \quad (4)$$

The 5cm threshold is critical for distinguishing intentional movement from control jitter.

Negative Constructor \mathcal{N} (Action Reversal). We use **Semantic Inversion**. If the property is “Closing Gripper”, we retrieve a negative frame I_k^- from a different timestamp in the same episode where the gripper is “Opening”, while minimizing the spatial distance of the end-effector to the anchor position. This ensures the negative looks spatially similar (same background/object) but performs the opposite action.

1.3. Object-Centric Motion: Tracking, Grounding, and Counting

As categorized in Table 1 of the main paper, this domain encompasses tasks that require fine-grained understanding of specific object instances rather than global scene motion. We unify **Object Tracking, Spatial Grounding, and Motion Counting** under a common construction pipeline using datasets like M4-Instruct and OmniWorld which provide 2D per-frame segmentation masks.

Unified Data Representation. For a video sequence, we utilize frame-level annotations $\mathcal{A}_t = \{(M_t^{(k)}, c_t^{(k)}, \text{id}^{(k)})\}_{k=1}^K$, where M is the segmentation mask, c is the category label, and id is the instance track ID.

Task 1: Object Tracking and Identification.

- **Property:** $\mathcal{M}_{\text{track}} = \{\text{Same Instance, Different Instance}\}$.
- **Construction:** Given Anchor I_i with object O_a , the Positive I_j shows the same instance O_a (verified by Track ID) at a later time.
- **Hard Negative:** The Negative I_k^- is selected to contain a *distractor object* O_b such that:

$$c(O_b) = c(O_a) \quad \text{AND} \quad \text{sim}_{\text{CLIP}}(O_b, O_a) > \tau \quad \text{AND} \quad \text{id}(O_b) \neq \text{id}(O_a) \quad (5)$$

This forces the model to track specific instances rather than general categories (e.g., distinguishing between two similar red cars).

Task 2: Spatial Grounding (Motion Direction).

- **Property:** $\mathcal{M}_{\text{ground}} = \{\text{Object moving left/right/up/down}\}$.
- **Construction:** We compute the centroid shift of the mask $\Delta c = \text{center}(M_j) - \text{center}(M_i)$. A valid pair requires $\|\Delta c\| > 0.1 \times \min(H, W)$ to ensure significant screen movement.
- **Hard Negative:** We select a frame where the same object moves in the *opposite* direction or remains static, penalizing models that hallucinate motion based on static object appearance.

Task 3: Motion Counting.

- **Property:** $\mathcal{M}_{\text{count}} = \{N \text{ objects moving [direction]}\}$.
- **Construction:** We aggregate individual object motions. For a query “How many people are walking left?”, we count instances satisfying the Grounding criteria.
- **Hard Negative:** We retrieve frames with $N-1$ or $N+1$ moving objects, or frames where N objects are present but moving in the *wrong* direction. This ensures the model performs dense temporal reasoning rather than single-frame detection.

2. Dataset Statistics and Analysis

Following the construction pipeline detailed in Section 1, we generated a total of **N=16.5k** motion contrast instruction tuning samples. This section provides a detailed analysis of the dataset distribution, diversity, and quality control.

2.1. Distribution across Domains

To ensure a comprehensive evaluation of motion understanding, we maintain a balanced distribution across the three primary domains defined in Table 1. The breakdown is as follows:

2.2. Answer Balance and Bias Mitigation

A critical issue in Multiple Choice Question (MCQ) benchmarks is the potential for *positional bias* (e.g., the correct answer is frequently option 'A').

- **Randomized Options:** During the generation of the JSON files, the positions of the Positive sample (Correct Answer), Hard Negative (Distractor), and Random Negative are shuffled for every triplet.

Domain	Source Data	Proportion	Key Motion Types
Camera Navigation	ScanNet, NuScenes	~35%	Ego-motion, Viewpoint Change, Parallax
Robot Manipulation	AgiBot	~30%	End-effector control, Gripper state, Interaction
Object-Centric	M4-Instruct, OmniWorld	~35%	Tracking, Grounding, Counting, Attribute Change

Table 1. Distribution of motion contrast samples across different domains. The balanced ratio prevents the model from overfitting to a specific visual style (e.g., only synthetic indoor scenes).

2.3. Motion Magnitude Distribution

To verify that our filtering mechanisms (e.g., ϕ in Sec 1.1) effectively remove static or trivial samples, we analyze the motion magnitude:

- **Camera Domain:** The average rotation between anchor and positive frames is 28° , and average translation is 0.2m.
- **Object Domain:** The average Intersection over Union (IoU) between the object bounding boxes in I_i and I_j is less than 0.5, indicating significant object displacement.

3. More Experimental Results

3.1. Ablation Study

3.1.1. Triplet vs. Binary Contrast

To validate the necessity of triplet-based motion contrast, we conduct an ablation study comparing our triplet formulation against a binary (pairwise) alternative. This ablation directly examines whether our $(I_{\text{anchor}}, I_{\text{pos}}, I_{\text{neg}})$ triplet formulation is superior to constructing separate binary pairs $(I_{\text{anchor}}, I_{\text{pos}})$ and $(I_{\text{anchor}}, I_{\text{neg}})$ for learning fine-grained spatio-temporal discrimination.

Experimental Setup. In the binary variant, we decompose each triplet into two separate binary pairs: a positive pair $(I_{\text{anchor}}, I_{\text{pos}})$ retaining motion property m , and a negative pair $(I_{\text{anchor}}, I_{\text{neg}})$ with conflicting property \bar{m} . The VQA formulation is simplified to independent binary verification questions for each pair (e.g., “Does the camera rotate left between these frames?”), treating positive and negative examples as isolated training instances rather than contrastive options within the same question. Both variants use identical training data scale, the same GRPO training procedure, and equivalent computational budgets.

Results. Table 2 presents the performance comparison on our **ReMoT-16k-Test** benchmark. The triplet-based approach substantially outperforms the binary baseline, achieving **38.0%** Overall accuracy compared to 19.4% for binary pairs—a remarkable **+18.6%** absolute improvement. On the Partial accuracy metric, triplet learning yields 64.0% versus 39.4%, representing a **+24.6%** gain.

Table 2. Ablation study comparing triplet-based and binary-based contrast learning on ReMoT-16k-Test benchmark. The triplet formulation with joint contrastive supervision demonstrates substantial advantages across both evaluation metrics.

Contrast Formulation	Overall (%)	Partial (%)
Binary Pairs $(I_{\text{anchor}}, I_{\text{pos}}), (I_{\text{anchor}}, I_{\text{neg}})$	19.4	39.4
Triplets $(I_{\text{anchor}}, I_{\text{pos}}, I_{\text{neg}})$	38.0	64.0
Absolute Gain	+18.6	+24.6

Analysis. The dramatic performance gap underscores the critical importance of joint contrastive supervision in learning fine-grained motion discrimination. While the binary baseline exposes the model to both positive and negative examples through separate $(I_{\text{anchor}}, I_{\text{pos}})$ and $(I_{\text{anchor}}, I_{\text{neg}})$ pairs, treating them as isolated instances fails to establish explicit contrastive relationships, making it difficult for the model to learn the relative motion relationships within the scene.

In contrast, our triplet formulation jointly presents both transitions—one exhibiting motion property m (I_{pos}) and another exhibiting the reversed or conflicting property \bar{m} (I_{neg})—for the same anchor frame. Critically, both I_{pos} and I_{neg} depict physically plausible transitions; the distinction lies in their opposing motion directions or attributes relative to the anchor (e.g., leftward vs. rightward camera rotation, approach vs. retract robot trajectories). This design enables the model to directly learn contrastive boundaries through comparative reasoning—simultaneously evaluating opposing motion patterns—forcing

explicit discrimination of fine-grained directional differences even when both alternatives are visually similar due to identical scene content.

3.1.2. GRPO Rollout Configuration

Group Relative Policy Optimization (GRPO) typically employs a rollout size of 8 for generating multiple response candidates per query. However, to improve training efficiency and reduce memory consumption, we train our model with a rollout size of 4. To validate this design choice, we conduct an ablation study comparing rollout sizes of 4 and 8 under identical training conditions.

Experimental Setup. For rapid validation, we train both variants on 1/10 of the full training dataset using the same hyperparameters, learning rate, and training steps. The only difference is the number of response candidates sampled per query during GRPO training: 4 for our configuration vs. 8 for the standard setting. Both models are evaluated on the ReMoT-16k-Test benchmark. This reduced-scale experiment allows us to efficiently compare rollout configurations without the computational cost of full-scale training.

Results. Table 3 presents the performance comparison. The performance difference is negligible (0.62% on Overall, +0.65% on Partial for rollout 4), demonstrating that reducing rollout size does not significantly impact model quality in our setting.

Table 3. Ablation study on GRPO rollout size trained on 1/10 of the full training data. Training with rollout size 4 achieves comparable performance to rollout size 8 while substantially reducing training time and memory consumption.

Rollout Size	Overall (%)	Partial (%)
8 (Standard)	20.48	40.28
4 (Ours)	19.86	40.93
Performance Gap	-0.62	+0.65

3.1.3. Optimization Strategy: GRPO vs. DPO and DAPO

Experimental Setup. We compare three representative optimization strategies—DPO, DAPO, and GRPO—using Qwen3-VL-4B as the base model under identical training data and computational budgets. We additionally compare two reward formulations within GRPO: learned rewards (LR) and rule-based rewards (RR).

Results. Table 4 summarizes the results. DPO achieves only 23.8% overall accuracy, while DAPO reaches 51.6% and our GRPO (RR) achieves **61.1%**.

Analysis. DPO optimizes over static preference pairs, which is ineffective for tasks requiring long and precise chains of reasoning and lacks the fine-grained supervision needed to identify specific errors. In contrast, GRPO samples multiple trajectories and computes group-relative advantages, enabling the model to explore varied spatial interpretations and learn from relative quality among its own outputs rather than manually curated negatives. This flexibility is critical for spatial reasoning, where errors manifest diversely (e.g., depth inconsistencies, left-right confusion).

Regarding reward formulation, we use rule-based rewards (RR) because spatial facts—such as 3D coordinates and motion directions—are deterministic and geometrically verifiable, yielding noise-free oracle supervision. GRPO-LR achieves only 32.3% overall accuracy, substantially below GRPO-RR (61.1%) across all task categories. Furthermore, per-step reward inference introduces significant overhead: **26h/epoch for LR vs. 3.5h for RR (7.4× slower)**. This confirms that deterministic geometric verification provides more effective and efficient supervision signals than preference-based learned reward modeling.

Table 4. Optimization strategy comparison (Qwen3-VL-4B). Ov.: Overall Acc. (%); Par.: Partial Acc. (%).

Method	Gripper-Move		Gripper-State		Composite		Overall	
	Ov.	Par.	Ov.	Par.	Ov.	Par.	Ov.	Par.
Baseline	15.3	46.2	3.2	35.5	4.8	22.0	7.8	34.6
DPO	30.6	59.1	14.5	41.9	19.4	52.2	23.8	53.1
DAPO	58.1	78.0	35.5	60.8	54.8	75.3	51.6	73.0
GRPO (LR)	33.9	62.1	22.6	50.5	38.7	66.7	32.3	60.4
GRPO (RR)	68.6	82.0	45.2	65.1	69.4	86.6	61.1	78.0

3.1.4. Scalability Across Model Sizes

Experimental Setup. We validate our method across three model scales: Qwen3-VL-2B, 4B, and 8B, keeping training data, hyperparameters, and evaluation protocol identical across all variants.

Results. Table 5 presents the results. Baselines exhibit uniformly low overall accuracy (4.4–9.7%), indicating that all model sizes struggle with joint spatial consistency. Our method yields consistent gains: +4.9 pts (2B), +53.3 pts (4B), and +68.5 pts (8B).

Analysis. The $8\times$ improvement on the 8B model demonstrates that spatial inconsistency stems primarily from training objective misalignment rather than model capacity limitations. The 4B model offers the best performance-efficiency trade-off, achieving 61.1% accuracy with approximately 12 hours of training on $8\times A100$, whereas the 8B model reaches 78.2% (+17.1 pts) at roughly $3\times$ the training cost.

Table 5. Scalability analysis across model sizes.

Base Model	Overall Acc. (%)		Partial Acc. (%)		Δ Overall
	Baseline	Ours	Baseline	Ours	
Qwen3-VL-2B	4.4	9.3	34.8	41.8	+4.9
Qwen3-VL-4B	7.8	61.1	34.6	78.0	+53.3
Qwen3-VL-8B	9.7	78.2	41.5	86.6	+68.5

3.2. Analysis of Reasoning Chain Faithfulness

To assess the quality of generated reasoning chains, we randomly sample 25 test outputs and manually annotate whether the reasoning chain is logically consistent with the final answer. The base model Qwen3-VL-4B-CoT exhibits reasoning errors in approximately 60% of cases, whereas ReMoT reduces this to **12%**.

A representative failure case illustrates the issue: the CoT states “...*left arm moving up and right* \rightarrow A...” (incorrect direction) while producing `<answer>B</answer>` (correct answer), where the reasoning chain directly contradicts the final prediction. This reflects a known *decoupling* problem: reasoning traces can serve as post-hoc rationalizations rather than genuinely guiding predictions when models are trained with final-answer supervision alone.

ReMoT mitigates this decoupling (60% \rightarrow **12%**) via composite rewards under GRPO that jointly optimize reasoning coherence and answer correctness across the entire response, rather than supervising only the final token.

3.3. More Quantitative Results

In this section, we present a comprehensive collection of quantitative results to complement the evaluations in the main text. First, Table 6 supplements Table 1 of the main text with additional experimental configurations on our ReMoT-16k-Test benchmark, including models trained with SFT only, models trained with interleaved SFT and GRPO, and evaluations of state-of-the-art closed-source VLMs. Subsequently, we also provide fine-grained, per-metric breakdowns for the six benchmarks summarized in Table 4 of the main text: VLM2-Bench (Table 7), VSI-Bench (Table 8), BLINK (Table 10), MMStar (Table 9), MUIRBENCH (Table 11), and MMMU (Table 12).

An important observation from Table 6 is the poor performance of SFT-only model, which is attributed to training collapse. The model essentially loses its chain-of-thought reasoning capabilities during training, defaulting to generating minimal responses consisting of only a single word or option letter. This failure mode highlights the critical role of our interleaved SFT and GRPO training approach in maintaining model stability and preserving reasoning quality throughout the training process.

Table 6. **Supplementary evaluation results on ReMoT-16k-Test.** We supplement the experimental results from Table ?? in the main text with the following additional evaluations: (1) models trained using SFT only, (2) models trained with 2 epochs of interleaved SFT and GRPO training, and (3) two closed-source VLMs. All models are evaluated across three fine-grained motion reasoning capabilities: *Navigation*, *Perceptual Grounding*, and *Manipulation*. The rightmost columns show macro-averaged performance across all tasks. Note that the closed-source models are evaluated on a mini-benchmark consisting of approximately 40 samples due to API cost constraints.

Model	Navigation			Perception		Manipulation						Avg.	
	Camera		Rel-Pos	Grounding	Counting	Gripper-Move		Gripper-State		Composite		Ov.	Par.
	Ov.	Par.	Ov.	Ov.	Ov.	Par.	Ov.	Par.	Ov.	Par.			
GPT-5-Chat	0.0	31.3	0.0	0.0	75.0	0.0	37.5	25.0	41.7	0.0	33.4	10.4	33.3
Gemini-2.5-Pro	25.0	62.5	25.0	25.0	75.0	12.5	41.7	0.0	16.7	25.0	50.0	26.4	49.1
Qwen2.5-VL-7B [55]	4.8	34.7	0.0	23.9	0.0	4.0	36.6	8.1	31.7	0.0	16.7	5.1	25.4
Qwen3-VL-CoT-4B [47]	2.4	25.9	22.5	35.8	79.0	15.3	46.2	3.2	35.5	4.8	22.0	20.7	38.9
InternVL3-2B [16]	1.6	21.5	20.0	31.3	60.5	0.8	29.8	6.5	35.0	0.0	17.7	14.9	29.3
InternVL3-8B [16]	2.8	23.8	15.0	30.6	51.7	1.6	28.5	8.1	38.7	0.0	18.3	12.2	28.9
LLaVA-One-Vision [32]	2.0	30.4	17.5	32.9	0.0	10.5	37.6	21.0	47.3	0.0	14.8	9.7	27.9
SFT	0.0	2.4	23.3	40.0	37.9	0.0	10.5	0.0	9.7	0.0	9.2	10.4	15.4
GRPO	27.0	62.4	24.3	44.3	94.1	54.5	74.3	27.3	44.7	61.3	79.2	33.6	61.6
SFT→GRPO	26.6	62.1	22.5	35.8	82.5	57.3	78.8	32.3	60.2	62.9	82.3	35.0	63.3
SFT↔GRPO	21.4	61.2	26.7	46.7	82.7	68.6	82.0	45.2	65.1	69.4	86.6	38.0	64.0
SFT↔GRPO(2 epochs)	31.5	68.8	23.3	40.0	93.3	79.8	89.5	45.2	65.1	64.5	83.3	39.9	67.7

Table 7. **Performance on VLM2-Bench [64]**, which evaluates the ability to compare and link fine-grained visual cues across multiple images. Without relying on any human- or model-annotated data, ReMoT achieves significant improvements and sets a new state-of-the-art. Reasoning-based models (marked with ●) are evaluated using their corresponding prompting strategies.

Baselines or Models	General		Object			Person				Overall*	
	Mat	Trk	Cpr	Cnt	Grp	Cpr	Cnt	Grp	VID	Avg	Δ_{human}
Chance-Level	25.00	25.00	50.00	34.88	25.00	50.00	34.87	25.00	-	32.73	-61.44
Human-Level	95.06	98.11	96.02	94.23	91.29	97.08	92.87	91.17	100.00	95.16	0.00
○ LLaVA-OneVision[32]	16.60	13.70	47.22	56.17	27.50	62.00	46.67	37.00	47.25	39.35	-55.81
○ LLaVA-Video-7B [66]	18.53	12.79	54.72	62.47	28.50	62.00	66.91	25.00	59.00	45.65	-49.51
○ LongVQA-7B [65]	14.29	12.98	46.53	49.47	29.00	58.00	41.56	25.00	45.00	37.10	-58.06
○ mPLUG-Owl2-7B [61]	17.37	18.26	49.17	62.97	31.00	63.00	58.06	29.00	43.00	40.87	-54.31
○ Qwen2-VL-7B [7]	18.07	19.18	68.08	61.84	37.50	72.00	67.92	47.00	55.25	49.76	-45.40
○ InternVL2.5-8B [16]	41.24	26.53	72.22	67.65	40.00	85.00	66.67	52.25	50.25	55.41	-39.75
○ InternVL2.5-26B [16]	30.50	30.59	43.33	51.48	52.50	59.50	59.67	61.25	45.25	45.59	-49.57
○ Qwen2.5-VL-7B [7]	35.91	43.38	71.39	41.72	47.50	80.00	59.76	69.00	45.00	54.82	-40.34
○ GPT-4o [27]	37.45	39.27	74.17	80.62	57.50	50.00	90.50	47.00	66.75	60.36	-34.80
● MM-Eureka-7B [41]	55.60	47.03	74.10	52.50	54.00	77.50	60.00	51.00	43.50	57.24	-37.91
● NoisyRollout-7B [38]	40.93	43.83	63.33	50.83	34.50	70.50	63.33	47.00	36.50	50.08	-45.08
● ThinkLite-VL-7B [57]	40.45	46.58	75.56	62.50	49.50	77.50	62.50	51.00	36.50	55.79	-39.37
● VLAA-Thinker-7B [10]	47.49	63.03	72.20	61.40	55.00	71.00	57.50	51.00	47.75	58.49	-36.67
○ Qwen2.5-VL-7B-CoT[7]	43.24	42.92	66.39	50.56	36.00	62.50	55.83	39.00	36.75	48.91	-46.24
● MiCo-7B-CoT[14]	57.14	67.12	81.94	56.67	58.00	65.00	57.50	62.00	44.25	61.06	-34.09
● Qwen3-VL-4B-CoT[47]	52.88	51.59	86.81	69.64	74.50	90.00	75.00	57.00	40.50	66.43	-28.73
● ReMoT-4B-CoT	57.23	54.14	87.56	73.06	78.50	92.00	76.67	66.00	46.75	70.21	-24.95
Δ Improvement	+4.35	+2.55	+0.75	+3.42	+4.00	+2.00	+1.67	+9.00	+6.25	+3.78	+3.80

Table 8. **Evaluation Results on VSI-Bench [60]**. For Spatial-MLLM and Qwen2.5VL-series, we use 16 frames as input. For other open-source methods and GPT-4o [27], we follow the VSI-Bench setting (8–32 frames). Gemini-1.5 Pro [18] samples at 1 FPS. **Bold / underline** denote best / second-best open-source results, respectively.

Methods	Numerical Question				Multiple-Choice Question				Avg. Rank	
	Obj.,Cnt.	Abs.,Dist.	Obj.,Size	Room,Size	Rel.,Dist.	Rel.,Dir.	Route,Plan	Appr.,Order		
<i>Proprietary Models</i>										
GPT-4o [27]	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0	8
Gemini-1.5 Pro [18]	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6	45.4	4
<i>Open-source Models</i>										
InternVL2-40B [16]	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6	36.0	9
LLaVA-OneVision-72B [32]	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2	7
LLaVA-Video-72B [66]	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9	3
Qwen2.5VL-3B [7]	24.3	24.7	31.7	22.6	38.3	41.6	26.3	21.2	30.6	10
Qwen2.5VL-7B [7]	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8	33.0	8
Qwen2.5VL-72B [7]	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9	37.0	6
Spatial-MLLM-4B [58]	<u>65.3</u>	<u>34.8</u>	<u>63.1</u>	<u>45.1</u>	<u>41.3</u>	<u>46.2</u>	<u>33.5</u>	<u>46.3</u>	<u>48.4</u>	<u>2</u>
ReMoT-CoT-4B (Ours)	78.2	42.5	76.8	56.3	52.1	57.4	45.6	59.7	58.8	1

Table 9. **Evaluation of various LVLMs on MMStar[13]**. We report the detailed results of the CP (coarse perception), FP (fine-grained perception), IR(instance reasoning), LR (logical reasoning), ST (science & technology), and MA (mathematics) core capabilities. The **best** results are highlighted in **bold**.

Model	LLM	Param.	CP	FP	IR	LR	ST	MA	Avg.
<i>Baselines</i>									
Random Choice	-	-	23.7	24.5	25.3	24.3	24.8	25.1	24.6
<i>Closed-source LVLMs</i>									
GeminiPro-Vision[49]	GeminiPro[49]	-	51.6	28.8	50.8	46.0	28.4	50.0	42.6
GPT4V (low)[45]	GPT4-Turbo[44]	-	62.0	32.8	55.2	48.0	33.6	44.8	46.1
GPT4V (high)[45]	GPT4-Turbo[44]	-	76.6	51.4	66.6	55.8	42.6	49.8	57.1
<i>Open-source LVLMs</i>									
TinyLLaVA[67]	Phi2-2.7B[42]	3B	60.4	31.6	50.8	30.4	18.0	24.8	36.0
Yi-VL[62]	Yi-6B[62]	6B	58.0	33.6	46.4	34.8	20.4	34.0	37.9
LLaVA-1.5[36]	Vicuna-v1.5-7B[17]	7B	58.8	24.0	38.8	24.0	13.6	22.8	30.3
ShareGPT4V[12]	Vicuna-v1.5-7B[17]	7B	58.8	28.0	45.6	24.4	17.2	24.0	33.0
InternLM-XC2[24]	InternLM2-7B[52]	7B	70.8	48.8	65.2	56.4	42.0	49.2	55.4
Qwen-VL-Chat[6]	Qwen-7B[4]	8B	59.6	32.0	50.8	29.2	22.0	31.6	37.5
Deepseek-VL[39]	Deepseek-7B[8]	8B	64.0	30.8	49.2	36.4	21.6	20.4	37.1
Monkey-Chat[34]	Qwen-7B[4]	10B	57.6	36.4	51.6	33.2	26.4	24.4	38.3
LLaVA-1.5[36]	Vicuna-v1.5-13B[17]	13B	58.8	28.0	41.6	24.4	18.4	25.6	32.8
CogVLM-Chat[56]	Vicuna-v1.5-7B[17]	17B	66.8	36.8	49.2	31.2	23.6	11.6	36.5
Yi-VL[62]	Yi-34B[62]	34B	53.2	31.2	52.0	32.4	12.4	35.2	36.1
LLaVA-Next[37]	NH2-Yi-34B[43]	34B	66.4	52.0	62.4	46.0	32.4	53.6	52.1
InternVL-Chat-V1.2[15]	NH2-Yi-34B[43]	40B	67.6	43.2	61.2	47.2	24.0	19.2	43.7
Sphinx-X-MOE[26]	Mixtral-8x7B[28]	57B	58.4	40.8	47.6	35.2	19.2	32.0	38.9
Qwen3VL-4B-Thinking[47]	Qwen3-4B[59]	4B	73.6	58.8	74.8	73.6	50.8	86.8	68.4
ReMoT-CoT-4B (Ours)	Qwen3-4B[59]	4B	3.2	60.8	2.8	78.4	55.2	2.0	70.4

Table 10. Results of different models on the BLINK[25] validation set. The first row shows task names and number of instances.

	Validation (1, 901)	IQ (150)	Similarity (135)	Counting (120)	Depth (124)	Jigsaw (150)	Art (117)	Fun.Corr. (130)
Random Choice	38.09	25	50	25	50	50	50	25
Human	95.67	80.00	96.70	93.75	99.19	99.00	95.30	80.77
Open-source multimodal LLMs								
MiniGPT-4-v2 [11]	34.23	20.67	44.44	13.33	50.81	34.67	43.59	20.77
OpenFlamingo-v2 [3]	39.18	18.67	62.22	30.00	54.03	47.33	52.99	24.62
InstructBLIP-7B [20]	39.72	20.00	47.41	32.50	51.61	52.67	47.01	23.85
InstructBLIP-13B [20]	42.24	24.67	49.63	30.83	51.61	52.67	51.28	29.23
LLaVA-internLM2-7B [51]	37.71	22.00	48.89	55.00	57.26	28.67	29.06	23.85
LLaVA-v1.5-7B-xtuner [19]	39.36	18.67	47.41	45.83	51.61	52.67	47.01	20.00
LLaVA-v1.5-13B-xtuner [19]	42.00	24.67	47.41	48.33	54.03	52.00	47.01	30.00
CogVLM [56]	41.54	26.67	47.41	38.33	52.42	52.67	47.86	23.08
LLaVA-v1.5-7B [36]	37.13	20.00	47.41	40.00	52.42	11.33	47.01	20.00
LLaVA-v1.5-13B [36]	42.66	26.00	47.41	45.00	53.23	58.00	47.01	26.15
LLaVA-v1.6-34B [37]	46.80	26.00	48.89	66.67	67.74	54.67	43.59	20.77
Qwen3VL-4B-Thinking [47]	59.50	10.00	61.54	71.67	80.65	58.67	77.42	39.23
ReMoT -CoT-4B (Ours)	62.15	23.20	84.44	73.33	80.65	59.33	75.21	46.92
API-based models								
Qwen-VL-Max [5]	40.28	21.33	51.11	56.67	58.06	4.67	38.46	28.46
Gemini Pro [49]	45.16	23.33	52.59	52.50	40.32	57.33	50.43	24.62
Claude 3 OPUS [1]	44.05	24.67	72.59	50.83	47.58	32.67	65.81	21.54
GPT-4V(ision) [2]	51.14	22.67	78.52	60.83	59.68	70.00	79.49	26.15
GPT-4 Turbo [2]	54.61	32.67	80.74	57.50	66.13	69.33	79.49	24.62
GPT-4o [2]	60.04	31.33	72.59	49.17	74.19	55.33	82.91	40.77
	Sem.Corr. (139)	Spatial (143)	Local (122)	Vis.Corr. (172)	Multi-view (133)	Reflect. (134)	Forensic (132)	
Random Choice	25	50	50	25	50	33.33	25	
Human	96.07	98.25	98.00	99.42	92.48	95.14	100.00	
Open-source multimodal LLMs								
MiniGPT-4-v2 [11]	28.78	44.76	47.54	26.16	48.87	30.60	24.24	
OpenFlamingo-v2 [3]	30.22	43.36	56.56	25.58	44.36	36.57	21.97	
InstructBLIP-7B [20]	30.94	56.64	48.36	30.81	55.64	33.58	25.00	
InstructBLIP-13B [20]	32.37	65.73	55.74	29.65	57.14	38.81	21.97	
LLaVA-internLM2-7B [51]	27.34	76.22	50.00	27.91	44.36	32.09	5.30	
LLaVA-v1.5-7B-xtuner [19]	28.78	68.53	36.89	29.07	38.35	29.85	36.36	
LLaVA-v1.5-13B-xtuner [19]	30.94	69.93	45.08	29.65	44.36	38.81	25.76	
CogVLM [56]	33.09	63.64	52.46	29.65	54.14	29.85	30.30	
LLaVA-v1.5-7B [36]	23.02	61.54	56.56	25.58	51.88	39.55	23.48	
LLaVA-v1.5-13B [36]	32.37	67.83	52.46	29.07	44.36	36.57	31.82	
LLaVA-v1.6-34B [37]	23.74	74.83	59.02	30.81	62.41	31.34	44.70	
Qwen3VL-4B-Thinking [47]	56.12	88.11	52.46	86.63	49.62	33.58	60.61	
ReMoT -CoT-4B (Ours)	64.03	86.71	53.28	87.21	42.86	28.36	56.06	
API-based models								
Qwen-VL-Max [5]	23.02	69.93	48.36	31.40	51.88	36.57	43.94	
Gemini Pro [49]	26.62	74.83	53.28	42.44	44.36	38.81	50.76	
Claude 3 OPUS [1]	25.18	58.04	51.64	36.63	56.39	26.87	46.21	
GPT-4V(ision) [2]	28.78	72.73	54.92	33.72	55.64	38.81	34.09	
GPT-4 Turbo [2]	30.94	69.23	52.46	52.33	52.63	32.84	63.64	
GPT-4o [2]	53.96	69.23	59.84	75.00	59.40	37.31	79.55	

Table 11. Experiment results on MUIRBENCH[54]. The first row shows task names and number of test data.

	Overall (2, 600)	Counting (234)	Action (164)	Grounding (84)	Matching (464)	Ordering (64)	Scene (186)
Random Choice	23.99	20.98	23.41	25.00	24.12	22.81	25.00
Human	93.15	94.87	97.56	85.71	94.83	87.50	94.62
GPT-4o [2]	68.00	49.15	44.51	36.90	86.85	23.44	71.51
Gemini Pro [49]	49.35	28.63	35.98	28.57	66.59	12.50	59.14
Mantis-8B-Idefics2 [29]	44.50	38.46	33.54	26.19	53.88	18.75	56.99
Idefics-9B-Instruct [30]	35.43	29.91	28.05	13.10	35.99	12.50	27.41
Emu2-Chat (37B) [48]	33.62	31.20	27.44	26.19	37.28	15.63	48.39
VILA1.5-13B [35]	33.12	19.66	28.66	25.00	40.95	10.94	56.45
Idefics2-8B [31]	26.08	21.79	26.22	26.19	24.78	15.62	56.45
OpenFlamingo-v2-9B [3]	23.73	21.79	26.83	30.95	24.14	21.88	22.58
LLaVA-NeXT-34B [37]	33.31	36.32	26.22	33.33	37.93	21.88	54.30
LLaVA-v1.5-7B-xtuner [19]	33.23	26.92	25.61	23.81	22.84	4.69	39.78
LLaVA-internLM2-7B [51]	28.15	34.19	26.22	32.14	25.65	7.81	42.47
LLaVA-v1.5-13B [36]	24.38	25.21	29.27	14.29	20.26	20.31	36.56
CogVLM [56]	20.85	14.10	26.22	16.67	21.34	12.50	41.40
MiniGPT-4-v2 [11]	17.35	11.97	14.02	25.00	17.03	18.75	14.52
Qwen3VL-4B-Thinking [47]	73.81	81.62	57.32	53.57	85.78	34.38	86.02
ReMoT-CoT-4B (Ours)	72.81	82.48	55.49	50.00	84.70	40.62	84.95

	Difference (340)	Cartoon (78)	Diagram (398)	Geographic (100)	Attribute (196)	Retrieval (292)
Random Choice	23.18	25.00	29.56	25.00	20.00	21.30
Human	92.94	82.05	98.99	98.00	87.76	86.30
GPT-4o [2]	60.29	51.28	88.69	56.00	56.12	80.14
GPT-4-Turbo [2]	60.59	52.56	79.15	57.00	50.51	64.04
Gemini Pro [49]	45.29	47.44	64.82	48.00	41.33	43.84
Mantis-8B-Idefics2 [29]	28.82	38.46	67.59	26.00	48.47	35.62
Idefics-9B-Instruct [30]	34.41	48.72	46.98	35.00	32.65	43.49
Emu2-Chat (37B) [48]	32.65	43.59	37.69	34.00	31.63	23.97
VILA1.5-13B [35]	24.71	30.77	42.71	31.00	24.49	30.14
Idefics2-8B [31]	27.65	39.74	25.38	21.00	17.86	17.12
OpenFlamingo-v2-9B [3]	21.76	25.64	31.91	25.00	18.88	15.41
LLaVA-NeXT-34B [37]	22.06	41.03	38.19	12.00	38.27	25.00
LLaVA-v1.5-7B-xtuner [19]	33.53	29.49	44.72	26.00	38.78	47.60
LLaVA-internLM2-7B [51]	19.12	39.74	35.43	12.00	23.98	28.42
LLaVA-v1.5-13B [36]	20.00	25.64	31.66	20.00	22.96	20.89
CogVLM [56]	19.71	41.03	19.60	13.00	16.33	15.75
MiniGPT-4-v2 [11]	20.00	21.79	21.61	13.00	17.35	14.73
Qwen3VL-4B-Thinking [47]	68.24	51.28	92.96	45.00	62.76	68.15
ReMoT-CoT-4B (Ours)	62.65	52.56	92.46	48.00	64.80	67.47

Table 12. Experiment results on the MMMU[63] validation and test set.

	Validation Overall (900)	Test Overall (10,500)	Art & Design (1,163)	Business (1,428)	Science (2,426)	Health & Medicine (1,752)	Human. & Social Sci. (947)	Tech & Eng. (2,784)
Random Choice	22.1	23.9	24.1	24.9	21.6	25.3	22.8	24.8
Frequent Choice	26.8	25.8	26.7	28.4	24.0	24.4	25.2	26.5
Expert (Worst)	76.2	-	-	-	-	-	-	-
Expert (Medium)	82.6	-	-	-	-	-	-	-
Expert (Best)	88.6	-	-	-	-	-	-	-
Large Multimodal Models (LMMs): Text + Image as Input								
OpenFlamingo2-9B [3]	28.7	26.3	31.7	23.5	26.3	26.3	27.9	25.1
CogVLM [56]	32.1	30.1	38.0	25.6	25.1	31.2	41.5	28.9
Qwen-VL-7B-Chat [6]	35.9	32.9	47.7	29.8	25.6	33.6	45.3	30.2
InstructBLIP-T5-XXL [20]	35.7	33.8	48.5	30.6	27.6	33.6	49.8	29.4
InternLM-XComposer2-VL* [24]	43.0	38.2	56.8	32.8	30.1	39.8	60.7	31.8
Yi-VL-34B* [62]	45.9	41.6	56.1	33.3	32.9	45.9	66.5	36.0
LLaVA-1.6-34B* [37]	51.1	44.7	58.6	9.9	36.0	1.2	0.2	36.3
InternVL-Chat-V1.2* [15]	1.6	6.2	62.5	37.6	37.9	49.7	70.1	40.8
VILA1.5* [35]	51.9	46.9	2.1	40.6	7.7	51.7	74.0	9.5
Qwen3VL-4B-Thinking [47]	70.8	-	71.9	80.9	68.1	65.3	73.5	65.3
ReMoT-CoT-4B (Ours)	71.4	-	70.3	82.1	67.4	66.3	77.0	66.9
Qwen-VL-MAX* [5]	51.4	46.8	4.2	39.8	36.3	52.5	70.4	40.7
GPT-4V(ision) (Playground) [45]	56.8	55.7	65.3	64.3	48.4	63.5	76.3	1.7
Claude 3 Opus* [1]	59.4	-	-	-	-	-	-	-
Gemini 1.5 Pro* [50]	2.2	-	-	-	-	-	-	-
GPT-4o* [27]	69.1	-	-	-	-	-	-	-

4. Limitations and Future Work

While ReMoT demonstrates significant improvements in spatio-temporal reasoning, we acknowledge two primary limitations that outline directions for future research.

Base Model Selection and Generalizability. Our experiments primarily utilize the Qwen3-VL-4B-Thinking architecture. This selection was deliberate rather than arbitrary; Qwen3-VL is currently recognized as possessing state-of-the-art spatial understanding capabilities among open-source models. Since our proposed GRPO-based optimization operates by refining the model’s internal reasoning chains (Chain-of-Thought) rather than injecting new knowledge from scratch, a strong intrinsic spatial prior is a prerequisite for stable RL convergence and effective self-correction. Consequently, we prioritized validating the ReMoT paradigm on the most capable available base model to establish a high-performance ceiling.

Scalability to In-the-Wild Videos. The construction of ReMoT-16K currently relies on datasets with basic meta-annotations (e.g., SE(3) camera poses in ScanNet, robot telemetry in AgiBot). While this ensures the logical rigor of our hard negative samples, it constrains the immediate scalability of our pipeline to massive, unannotated in-the-wild videos. For future work, we aim to leverage more advanced visual-language related technical advances [9, 21–23, 33, 40, 46, 53] to enhance the model’s ability to handle unstructured real-world data, thereby expanding its generalization scope beyond structured datasets.

5. Visualization of Chain-of-Thought Compactness

5.1. Our Model’s Chain-of-Thought

In this section, we provide qualitative examples showcasing how our model generates concise and structured **Chain-of-Thought (CoT)** reasoning to arrive at correct answers.



ReMoT-CoT-4B:(288 words)

So, let's analyze the images step by step.

First, we need to look at the left robotic arm (the one on the left side of the images) and track its vertical movement from Image 1 to Image 2, then Image 1 to Image 3, and Image 2 to Image 3.

Starting with Image 1 to Image 2:

In Image 1, the left robotic arm is positioned with its gripper near the top left. In Image 2, the left robotic arm has moved down. Let's check the position. Image 1 shows the left arm higher up, Image 2 shows it lower. So vertical movement is Down. So option B.

Next, Image 1 to Image 3:

Image 1: left arm is in a certain position. Image 3: the left arm has moved up. Let's see. In Image 3, the left arm is higher than in Image 1. So vertical movement is Up. Option A.

Then, Image 2 to Image 3:

Image 2: left arm is lower. Image 3: left arm is higher. So movement from Image 2 to Image 3 is Up. Option C.

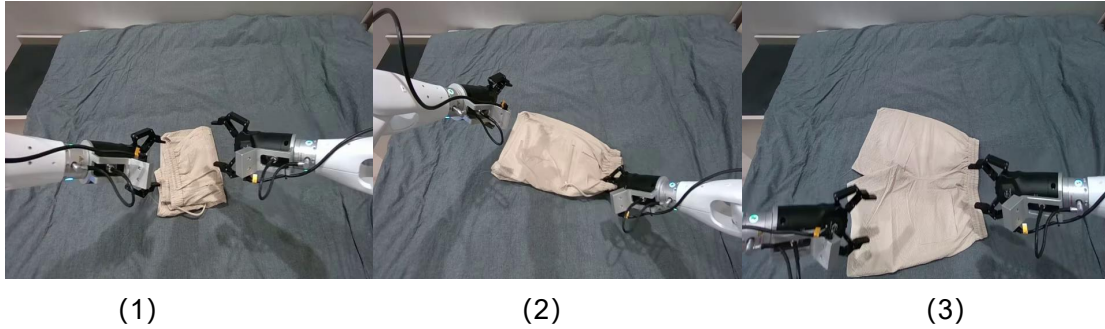
Wait, let's confirm each: Image 1 to 2: Left arm moves down (B). Image 1 to 3: Left arm moves up (A). Image 2 to 3: Left arm moves up (C).

So the answers are B, A, C.

</think>

<answer>BAC</answer>

Figure 1. **Example 1. Prompt:**The image showed to you is what the robot seen by its eyes. In the image, the robotic arm on the left is the robot's left arm, and the robotic arm on the right is the robot's right arm. Focus only on robot arm/gripper motion across the three images. Please select from the following options the vertical movement direction of the left robotic arm from Image 1 to Image 2? A: Up, B: Down, C: No movement. Please select from the following options the vertical movement direction of the left robotic arm from Image 1 to Image 3? A: Up, B: No movement, C: Down. Please select from the following options the vertical movement direction of the left robotic arm from Image 2 to Image 3? A: Down, B: No movement, C: Up. Answer all three questions above in order. Only return the correct option A, B, or C for each of the three questions in order inside <answer></answer>, e.g., <answer>CAB</answer>. **Answer:**BAC



ReMoT-CoT-4B:(251 words)

So, let's analyze each step by step.

First, we need to look at the movement of the left robotic arm from Image 1 to Image 2. Let's compare the positions.

In Image 1: The left arm is on the left side, closer to the left edge of the frame. In Image 2: The left arm has moved to the right compared to Image 1. So the horizontal movement direction is Right.

So the first answer is A.

Next, left robotic arm from Image 1 to Image 3. Image 1: left arm is on the left. Image 3: left arm is further left than Image 1. So movement is Left. Second answer is B.

Then, left robotic arm from Image 2 to Image 3. Image 2: left arm is at a certain position. Image 3: left arm is more to the left than Image 2. So movement is Left. Third answer is B.

Wait, let me double-check.

Image 1 left arm: let's see the position. The left arm in Image 1 is on the left side of the image, with the gripper near the center. Image 2: left arm is moved to the right, so from 1 to 2, left arm moves right (A). Image 1 to 3: left arm in Image 3 is further left than Image 1 (so left movement, B). Image 2 to 3: left arm in Image 3 is further left than Image 2 (so left movement, B).

So the answers are A, B, B.

</think>

<answer>ABB</answer>

Figure 2. **Example 2. Prompt:** The image showed to you is what the robot seen by its eyes. In the image, the robotic arm on the left is the robot's left arm, and the robotic arm on the right is the robot's right arm. Focus only on robot arm/gripper motion across the three images. Please select from the following options the horizontal movement direction of the left robotic arm from Image 1 to Image 2? A: Right, B: No movement, C: Left. Please select from the following options the horizontal movement direction of the left robotic arm from Image 1 to Image 3? A: Right, B: Left, C: No movement. Please select from the following options the horizontal movement direction of the left robotic arm from Image 2 to Image 3? A: Right, B: Left, C: No movement. Answer all three questions above in order. Only return the correct option A, B, or C for each of the three questions in order inside <answer></answer>, e.g., <answer>CAB</answer>. **Answer:** ABB



ReMoT-CoT-4B:(75 words)

Got it, let's check if the bags in all three images are the same. First, look at the first image: the bag is a backpack with white, black, and brown sections. The second image shows a similar backpack—same color scheme and design. The third image also has a backpack with the same white, black, brown colors and structure. All three bags look identical in design, colors, and features. So the statement is true.

</think>

T

Figure 3. **Example 3. Prompt:**Judge the following statement based on the input images: ‘The bags in these images are the same bag.’. Provide only one correct answer: ‘T’(True) or ‘F’(False). Respond with either ‘T’ or ‘F’. **Answer:**T

5.2. Comparison with base model

In this section, we provide a qualitative visualization showing how our model produces a more concise and structured **Chain-of-Thought (CoT)** compared to the untrained base model, showcasing how our model uses more refined thought chains to arrive at more correct answers.

Example Overview. The following example illustrates one reasoning instance from the evaluation set. Both models were given the same problem prompt, and their respective CoT traces are shown below. Figure 4 illustrates the input of the model in this example.

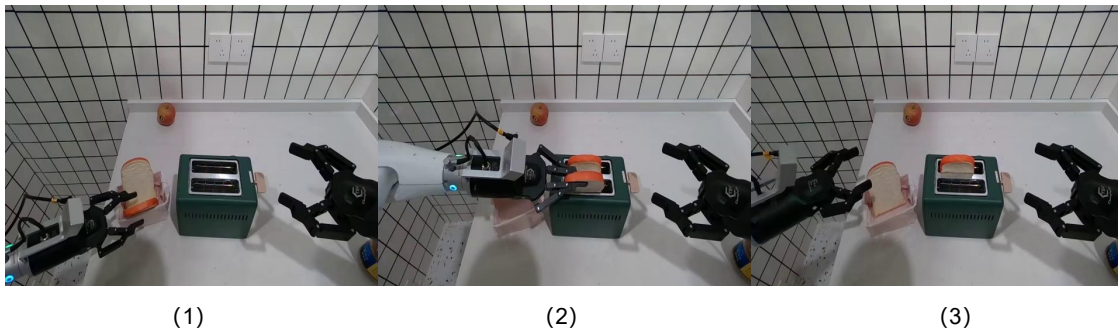


Figure 4. **An example used to compare CoT lengths. Prompt:**The image showed to you is what the robot seen by its eyes. In the image, the robotic arm on the left is the robot’s left arm, and the robotic arm on the right is the robot’s right arm. Focus only on robot arm/gripper motion across the three images. Please select from the following options whether the left gripper is opened or closed from Image 1 to Image 2? A: Closed, B: Opened, C: No movement. Please select from the following options whether the left gripper is opened or closed from Image 1 to Image 3? A: Opened, B: No movement, C: Closed. Please select from the following options whether the left gripper is opened or closed from Image 2 to Image 3? A: Closed, B: Opened, C: No movement. Answer all three questions above in order. Only return the correct option A, B, or C for each of the three questions in order inside <answer></answer>, e.g., <answer>CAB</answer>. **Answer:**AAB

ReMoT-CoT-4B (Ours)(310 words):

So, let's analyze each question step by step.

First, we need to look at the left gripper (robot's left arm) from Image 1 to Image 2.

In Image 1: The left gripper is holding a sandwich (or bread), so it's closed. In Image 2: The left gripper is still holding the sandwich, so it's closed. Wait, but let's check again. Image 2 shows the left arm with the sandwich, so the gripper is closed. So from 1 to 2, left gripper is closed. So the answer is A: Closed.

Next, left gripper from Image 1 to Image 3.

Image 1: Left gripper is closed (holding the sandwich). Image 3: Left gripper is open (the sandwich is no longer being held, the arm is open). So from 1 to 3, it went from closed to open. So the answer is A: Opened.

Third, left gripper from Image 2 to Image 3.

Image 2: Left gripper is closed (holding the sandwich). Image 3: Left gripper is open (sandwich is not held). So from 2 to 3, it's opened. So answer is B: Opened.

Wait, let's confirm each step.

Image 1 left gripper: closed (holding bread). Image 2 left gripper: closed (still holding bread). So 1→2: closed → closed, so A (Closed). Image 1→3: closed → open, so A (Opened). Image 2→3: closed → open, so B (Opened).

So the three answers are A, A, B → "AAB"? Wait, no. Wait the first question is 1→2: A (Closed). Second is 1→3: A (Opened). Third is 2→3: B (Opened).

Wait the options for each question:

First question: A: Closed (1→2). Yes, because both images show left gripper closed.

Second question: A: Opened (1→3). Because 1 is closed, 3 is open.

Third question: B: Opened (2→3). Because 2 is closed, 3 is open.

Yes. So the answers are A, A, B → "AAB".

</think>

<answer>AAB</answer>

Qwen3VL-4B-Thinking (Untrained)(6241 words):

Got it, let's analyze the images step by step. First, I need to look at the left gripper (robot's left arm) in each image.

Image 1: The left gripper is holding a sandwich (or bread). The gripper is closed because it's holding the object. So the state is closed.

Image 2: The left gripper is now holding the sandwich in a different position, but is it closed? Wait, in Image 2, the left gripper seems to be holding the bread, but maybe the grip is still closed? Wait no, wait the action. Wait, from Image 1 to Image 2, the left gripper was holding the bread (closed) and then in Image 2, the gripper is still holding it, but maybe the position changed. Wait, no, let's check the gripper's state. A closed gripper is when the fingers are together, holding the object. Opened is when they're apart.

Image 1: Left gripper is closed (holding the bread).

Image 2: Left gripper is still holding the bread, so closed? Wait, no, in Image 2, the left gripper is the one that moved. Wait, in Image 1, the left gripper is holding the bread (so closed). In Image 2, the left gripper is still holding the bread (so closed). Wait, but wait, maybe I need to check if the gripper opened. Wait, no, let's see:

Wait, in Image 1, the left gripper is holding the bread (closed). In Image 2, the left gripper is holding the bread as well (closed). Wait, but maybe the position changed, but the gripper state is still closed. Wait, no, wait the first question is Image 1 to Image 2: left gripper opened or closed.

Wait, let's look again:

Image 1: Left gripper (on left) is holding the bread (closed).

Image 2: Left gripper is still holding the bread (closed). Wait, but maybe the gripper opened? Wait, no. Wait, in Image 2, the left gripper is the one that moved. Let's check the positions.

Wait, Image 1: Left gripper has a sandwich (or bread) held in its jaws (closed). Image 2: Left gripper is now holding the sandwich (jaws closed). Wait, maybe the gripper didn't open. Wait, no, maybe it's closed. Wait, but maybe I need to check if the gripper's state changed.

Wait, the first question: Image 1 to Image 2. Left gripper is opened or closed.

In Image 1: Left gripper is closed (holding the bread). In Image 2: Left gripper is still holding the bread (so closed). Wait, but maybe the gripper opened? Wait, no. Wait, perhaps the gripper closed? Wait, no. Wait, let's see the actual images:

Wait, in Image 1, the left gripper (robot left arm) is grasping a slice of bread (so fingers closed). In Image 2, the left gripper is still grasping the bread (fingers closed). So from 1 to 2, the left gripper is closed. Wait, but the options are A: Closed, B: Opened, C: No movement.

Wait, but maybe the gripper opened. Wait, let's check the second image: the left gripper is now holding the bread in a different position. Wait, maybe the gripper is still closed. So from 1 to 2, left gripper is closed? But wait, maybe it opened. Wait, no. Wait, let's think again.

Wait, perhaps in Image 1, the left gripper is holding the bread (closed), and in Image 2, the left gripper is still holding it, but maybe the grip is still closed. So the state is closed. So the answer for first question is A: Closed.

Wait, no, wait. Let's look at the third image.

Image 3: Left gripper is holding the bread? Wait, in Image 3, the left gripper is holding the bread, but wait, Image 3 shows the left gripper is holding the bread (same as Image 1 and 2?), no. Wait, let's see:

Image 1: Left gripper (left arm) is holding the bread (closed).

Image 2: Left gripper is holding the bread (closed).

Image 3: Left gripper is holding the bread (closed).

Wait, no, that can't be. Wait, let's check the images again. Wait, the problem says: the robot's left arm is on the left, right arm on the right.

In Image 1: Left arm (gripper) is holding a sandwich (closed). Right arm is open.

Image 2: Left arm is moving towards the toaster, holding the sandwich (closed). Right arm is open.

Image 3: Left arm is holding the sandwich (closed). Wait, maybe in Image 3, the left arm has the sandwich, but the gripper is closed.

Wait, no. Wait, maybe in Image 1, the left gripper is closed (holding bread), Image 2: left gripper is still closed (holding bread), so from 1 to 2, no movement (closed). But the first question is whether the left gripper is opened or closed from 1 to 2. Wait, the gripper state didn't change, so it's closed. So A.

Wait, but maybe the gripper opened. Let's check:

Wait, in Image 1, the left gripper has the bread (closed). In Image 2, the left gripper is moving towards the toaster, and the bread is still in the gripper. So the gripper is closed. So from 1 to 2, the state is closed. So the answer for first question is A: Closed.

Second question: Image 1 to Image 3. Left gripper from Image 1 to Image 3. Image 1: closed. Image 3: closed. So same state. So answer is B: No movement.

Third question: Image 2 to Image 3. Image 2: left gripper is holding the bread (closed). Image 3: left gripper is holding the bread (closed). So same state. So answer is C: No movement.

Wait, but wait, maybe I made a mistake. Let's look again.

Wait, Image 1: Left gripper is holding the bread (closed).

Image 2: Left gripper is holding the bread (closed). So no movement.

Image 3: Left gripper is holding the bread (closed). So from 1 to 3, no movement. From 2 to 3, no movement.

Wait, but the problem says:

First question: Image 1 to Image 2: left gripper opened or closed.

If in Image 1 it's closed, Image 2 it's closed, then the state didn't change. So the gripper is closed. So answer is A.

Second question: Image 1 to Image 3: left gripper is closed (since both 1 and 3 are closed), so no movement. Answer is B.

Third question: Image 2 to Image 3: both closed, so no movement. Answer is C.

But wait, maybe I misread. Let's check again.

Wait, Image 1: Left gripper (left arm) is holding a sandwich (closed).

Image 2: Left gripper is moving towards the toaster, but the sandwich is still in the gripper. So the gripper is closed. So from 1 to 2, the gripper is still closed. So the answer is A: Closed.

Image 3: Left gripper is holding the sandwich (closed). So from 1 to 3, same state. So no movement. B.

From 2 to 3, same state. So no movement. C.

But let's confirm with the images:

Image 1: Left arm (gripper) has the sandwich, closed.

Image 2: Left arm is still holding the sandwich (closed).

Image 3: Left arm is holding the sandwich (closed).

Yes, so all states are closed.

First question: $1 \rightarrow 2$, closed. A.

Second: $1 \rightarrow 3$, closed. But the question is whether opened or closed. Wait, the options are: A: Closed, B: Opened, C: No movement.

Wait, if the gripper is closed in both Image 1 and 2, then it's not opened. So the state is closed. So the answer is A.

Wait, but "closed" means the gripper is closed. So from 1 to 2, the gripper is closed (no change). So the answer is A: Closed. Because the question is "whether the left gripper is opened or closed from Image 1 to Image 2". So if it's closed in both, then it's closed.

Wait, but the options are:

For first question: A: Closed, B: Opened, C: No movement.

Wait, if the gripper is closed from 1 to 2, then the answer is A.

If it's opened, then B. If no movement, then C.

But if it's closed in both, then the state is closed. So the answer is A.

Wait, but "no movement" would mean the gripper is in the same state. But the question is not about movement but the state. Wait, no. The question is: whether the left gripper is opened or closed (the state) from 1 to 2. So the state at 1 is closed, state at 2 is closed. So the gripper is closed. So A.

Second question: Image 1 to 3: state is closed, so A? Wait, no. Wait, the options for second question: A: Closed, B: Opened, C: No movement.

Wait, no. Wait, the question is: select whether the left gripper is opened or closed from Image 1 to Image 3.

Wait, the state of the gripper in Image 1 is closed. In Image 3, it's closed. So the gripper is closed. So the answer would be A: Closed. But wait, the question is "whether the left gripper is opened or closed from Image 1 to Image 3". So if it's closed both times, then the answer is A.

But wait, maybe I'm misunderstanding. Let's think of it as: did the gripper open or close during the transition. If it remained closed, then it's closed. So first question: Image 1 to 2: gripper closed (since both are closed), so A.

Wait, but maybe in Image 2, the gripper opened. Let's check the images again.

Looking at the images:

Image 1: Left gripper is holding a slice of bread (closed).

Image 2: Left gripper is still holding the bread (closed). So no, it's closed.

Image 3: Left gripper is holding the bread (closed).

Yes, so all are closed.

First question: $1 \rightarrow 2$: closed \rightarrow A.

Second question: $1 \rightarrow 3$: closed \rightarrow A. Wait, but the options for second question: A: Closed, B: Opened, C: No movement.

Wait, but "no movement" would be if the gripper didn't change. But the question is whether it's opened or closed. So if it's closed, then A. But maybe the question is whether it changed. Wait, no. The question is: "whether the left gripper is opened or closed from Image 1 to Image 2".

So from Image 1 to Image 2, the left gripper's state is closed (in Image 1 it was closed, in Image 2 it's closed). So the answer is A: Closed.

But wait, let's think again. Maybe in Image 2, the left gripper is open.

Wait, looking at the second image: the left arm is moving towards the toaster, and the sandwich is in the gripper. The gripper is closed.

Yes. So left gripper is closed in all images.

Therefore:

First question: Image 1 to 2: closed \rightarrow A.

Second question: Image 1 to 3: closed \rightarrow A? But no, wait. Wait, the problem says "select from the following options whether the left gripper is opened or closed from Image 1 to Image 3".

Wait, no. Wait, the options for each question are:

For first question: A: Closed, B: Opened, C: No movement.

Wait, the options are not about whether it's opened or closed, but whether the movement is opened or closed. Wait, no, the question is: "whether the left gripper is opened or closed from Image 1 to Image 2".

Wait, maybe the question is asking if the gripper opened (went from closed to open) or closed (went from open to closed) or no movement (stayed the same).

Ah, that's the key. The question is not about the state, but the change.

Wait, let's re-read the question:

"Please select from the following options whether the left gripper is opened or closed from Image 1 to Image 2?"

Ah, no. Wait, the question is: "whether the left gripper is opened or closed from Image 1 to Image 2". So it's whether the gripper's action (opening or closing) happened during the transition from 1 to 2.

Wait, no, maybe it's asking if the gripper is in open or closed state after the transition. Wait, the wording is "whether the left gripper is opened or closed from Image 1 to Image 2". So perhaps the state at 2.

But let's think: the options are A: Closed (meaning the gripper is closed), B: Opened (gripper is open), C: No movement (gripper didn't change state).

Wait, but the question is about the transition. Wait, no, the problem says: "whether the left gripper is opened or closed from Image 1 to Image 2". So for each question, it's whether during the transition from 1 to 2, the left gripper is in open, closed, or no movement.

But maybe it's simpler: the question is asking what is the state of the left gripper in Image 2, compared to Image 1. But the options are A: Closed (meaning the gripper is closed in the transition), B: Opened (gripper is open), C: No movement.

Wait, no. Let's look at the problem statement again:

"Please select from the following options whether the left gripper is opened or closed from Image 1 to Image 2?
A: Closed, B: Opened, C: No movement."

Ah, so the options are:

A: Closed \rightarrow meaning the left gripper is closed (so from 1 to 2, it remains closed).

B: Opened \rightarrow the left gripper is opened.

C: No movement \rightarrow the gripper didn't change (so state same as before).

Wait, but how to determine.

Let's check each image's left gripper.

Image 1: Left gripper (left arm) is holding the bread (so closed).

Image 2: Left gripper is holding the bread (closed).

So from 1 to 2, the gripper is closed (same as before). So the answer for first question is A: Closed.

Image 3: Left gripper is holding the bread (closed).

Second question: Image 1 to 3. The gripper is closed in both, so answer A: Closed.

Third question: Image 2 to 3. The gripper is closed in both, so answer A: Closed.

But wait, the options for third question is C: No movement.

Wait, no. Let's think again.

Wait, no. The key is: the question is about the state of the left gripper during the transition. But the options are whether it's closed, opened, or no movement.

Wait, let's check the images more carefully.

Image 1: Left gripper is closed (holding bread).

Image 2: Left gripper is still closed (holding bread).

Image 3: Left gripper is closed (holding bread).

So for each transition:

1→2: both closed → A: Closed.

1→3: both closed → A: Closed.

2→3: both closed → A: Closed.

But wait, the options for each question are A: Closed, B: Opened, C: No movement.

Wait, but maybe "no movement" means the gripper didn't move (but state changed). Wait, no. Let's think of the question as: when moving from image 1 to image 2, what is the state of the left gripper?

Wait, but the problem says: "Focus only on robot arm/gripper motion across the three images. Please select from the following options whether the left gripper is opened or closed from Image 1 to Image 2?"

So maybe the question is: when going from 1 to 2, the left gripper is open or closed.

Wait, in Image 1: left gripper is closed.

In Image 2: left gripper is closed.

So from 1 to 2, the gripper is closed. So answer A.

Image 1 to 3: both closed → A.

Image 2 to 3: both closed → A.

But that seems unlikely. Wait, maybe I'm wrong. Let's look at the images again.

Wait, Image 1: Left gripper (left arm) is holding a sandwich (closed).

Image 2: Left gripper is holding the sandwich (closed). So no movement. Wait, but the question is about the state.

Wait, the first question: "whether the left gripper is opened or closed from Image 1 to Image 2".

So if in Image 2, the left gripper is closed, then answer A.

Yes.

Second question: Image 1 to 3: closed → A.

Third question: Image 2 to 3: closed → A.

But that can't be. Because the problem must have different answers.

Wait, maybe I misinterpret the images.

Wait, looking at the images:

Image 1: Left gripper is holding a slice of bread (closed).

Image 2: Left gripper is moving towards the toaster, but the bread is still in the gripper (closed).

Image 3: Left gripper is holding the bread (closed).

Yes, all are closed.

But wait, maybe in Image 2, the left gripper is open. Let me check again.

Wait, the second image: the left arm is the one that moved. In Image 2, the left arm is now holding the bread, but the gripper's position might be different. Wait, in Image 2, the left gripper is still closed. Because it's holding the bread.

Yes, so all left grippers are closed.

Wait, but the options for the third question:

Wait, let's think of the first question.

If in Image 1, the left gripper is closed. In Image 2, is it closed? Yes.

So the answer for first question is A: Closed.

Second question: Image 1 to 3, left gripper closed \rightarrow A.

But the problem says "select from the following options whether the left gripper is opened or closed from Image 1 to Image 2? A: Closed, B: Opened, C: No movement."

Wait, maybe the question is about the movement: if the gripper opened (from closed to open), closed (open to closed), or no movement.

But the options are A: Closed, which would mean the gripper is closed (so no movement), B: Opened (gripped opened), C: No movement.

Wait, the problem says: "whether the left gripper is opened or closed from Image 1 to Image 2".

So it's the state of the gripper at Image 2.

If at Image 2, the gripper is closed, then answer is A.

If it's opened, B.

If no change (closed to closed), then it's closed.

But the options include "No movement" as C.

Wait, maybe "No movement" means the gripper didn't change (so same state), so if the state is closed, it's "closed", and "no movement" is a different option.

Wait, the problem statement for each question is:

1. Whether the left gripper is opened or closed from Image 1 to Image 2.

So the possible answers are: A: Closed (so in Image 2, it's closed), B: Opened (in Image 2, it's opened), C: No movement (the gripper didn't change, but perhaps the state is still closed).

But this is confusing. Let's think of "No movement" as the gripper didn't move (so state same as Image 1).

But the problem says "from Image 1 to Image 2", so the state in Image 2.

If Image 2's left gripper is closed, then A.

If opened, B.

If no movement (same as Image 1), then maybe C.

Wait, but the problem says the options are:

A: Closed, B: Opened, C: No movement.

So if the gripper is closed in Image 2, then A.

If it's opened, B.

If it's the same state (closed), then it's A. Wait, no, because "no movement" is a separate option.

Wait, maybe the question is:

Is the left gripper opened (meaning the transition involved opening), closed (transition involved closing), or no movement (transition involved no change in state).

But I think the intended question is: in the transition from Image 1 to Image 2, what is the state of the left gripper (closed or opened).

Wait, looking at the example:

Image 1: Left gripper closed.

Image 2: Left gripper closed.

So answer A: Closed.

Image 1 to 3: both closed \rightarrow A.

Image 2 to 3: both closed \rightarrow A.

But the problem's answer options must be different. Maybe I made a mistake.

Wait, let's check the images again carefully.

Image 1: Left arm (gripper) is holding a slice of bread (closed).

Image 2: Left arm (gripper) is holding the bread (closed).

Image 3: Left arm (gripper) is holding the bread (closed).

Yes.

But wait, what about the first image:

Image 1: The left gripper (left arm) is holding a sandwich (closed).

Image 2: The left gripper is holding the sandwich (closed).

Image 3: The left gripper is holding the sandwich (closed).

So all are closed.

But maybe in Image 3, the gripper is open. Wait, no, the third image shows the left gripper holding the bread.

Wait, maybe I'm looking at the wrong parts.

Let me describe each image:

Image 1: The countertop has a toaster, an apple, a sandwich (held by left gripper), and the right gripper is open.

Image 2: The left gripper is holding the sandwich (closed), moving towards the toaster. The right gripper is open.

Image 3: The left gripper is holding the sandwich (closed), and the right gripper is open.

So all left grippers are closed.

Thus:

1. From 1 to 2: left gripper closed \rightarrow A.

2. From 1 to 3: left gripper closed \rightarrow A.

3. From 2 to 3: left gripper closed \rightarrow A.

But this can't be, because the options for each question are A, B, C.

Wait, no, maybe the problem is that in Image 2, the left gripper is open.

Wait, looking at Image 2: the left gripper is moving, but is it holding the sandwich? Yes. So the gripper is closed.

Yes.

Wait, perhaps the third question has a different answer. Let's think:

First question: 1 \rightarrow 2: closed \rightarrow A.

Second question: 1 \rightarrow 3: closed \rightarrow A.

Third question: 2→3: closed → A.

But the problem says “the robot seen by its eyes”, so maybe I’m wrong.

Alternatively, maybe in Image 2, the left gripper is open.

Wait, no, it’s holding the sandwich.

Wait, perhaps the first question: Image 1 to 2, left gripper is closed (since it’s holding), so A.

Image 1 to 3: closed → A.

Image 2 to 3: closed → A.

But this would be AAA, but the answer must be different.

Wait, maybe I made a mistake in Image 2.

Looking at the second image:

The left gripper is the one on the left. In Image 2, the left gripper is now holding the sandwich, but the gripper is open? Wait, no. If it’s holding, it’s closed.

Wait, maybe the problem is that in Image 2, the left gripper is open.

Wait, let’s see:

Image 1: Left gripper (left arm) is closed (holding bread).

Image 2: Left gripper is now open (the bread is no longer held), but wait, no. The bread is still being held.

Wait, I think I need to check the images again.

Wait, the user provided the three images:

First image: left gripper holding a sandwich (closed), right gripper open.

Second image: left gripper moving towards the toaster, holding the sandwich (closed), right gripper open.

Third image: left gripper holding the sandwich (closed), right gripper open.

Yes.

So the state of left gripper in all images is closed.

Thus:

1. 1→2: closed → A.

2. 1→3: closed → A.

3. 2→3: closed → A.

But this can’t be right, because the problem is designed to have different answers.

Wait, perhaps the first question is asking about the movement: did the gripper open, close, or move.

Wait, no. Let’s think differently.

Maybe the left gripper in Image 1 is closed, in Image 2 it is open.

Wait, maybe I misread the images.

Let’s imagine:

Image 1: Left gripper is closed (holding bread).

Image 2: Left gripper is open (no longer holding bread).

Image 3: Left gripper is closed (holding bread).

If that’s the case, then:

1. 1→2: opened (A: Closed is wrong, B: Opened), so B.

2. 1→3: closed (A: Closed), but since it’s closed, A.

3. 2→3: closed (A: Closed).

But why would the gripper open in Image 2?

Let's check the images again:

Image 1: Left gripper is holding the bread (closed).

Image 2: Left gripper is moving towards the toaster, but the bread is still in the gripper. So it's closed.

Image 3: Left gripper is holding the bread (closed).

Yes.

Wait, maybe the problem is in the orientation. The left gripper in Image 2 is open.

Wait, no.

Another possibility: the left gripper in Image 1 is closed, in Image 2 it is open.

Let's look at the gripper's position.

In Image 1: Left gripper's fingers are together (closed), holding the bread.

In Image 2: Left gripper's fingers are open.

Wait, if the bread is no longer in the gripper.

Wait, in Image 2, the left gripper is moving towards the toaster, and the bread is still in the gripper? Or not?

Wait, the problem says: "the robot arm/gripper motion across the three images".

Perhaps in Image 2, the left gripper has released the bread.

But the problem states: "In the image, the robotic arm on the left is the robot's left arm, and the robotic arm on the right is the robot's right arm."

Let me describe the images again:

Image 1: Left arm (gripper) is holding a slice of bread (so closed), right arm is open.

Image 2: Left arm (gripper) is holding a slice of bread (closed), right arm is open.

Image 3: Left arm (gripper) is holding a slice of bread (closed), right arm is open.

Yes.

So all left grippers are closed.

Wait, but the answer for the third question is probably C: No movement.

Wait, maybe the question is whether the gripper opened (B) or closed (A) or no movement (C).

Wait, for the first question, Image 1 to 2:

If the left gripper is closed in both images, then the state is closed. But the question is "whether the left gripper is opened or closed from Image 1 to Image 2".

So if it's closed, answer A.

Second question: Image 1 to 3: closed, answer A.

Third question: Image 2 to 3: closed, answer A.

But this seems unlikely. Maybe I need to look at the problem's example.

The problem says: "Please select from the following options whether the left gripper is opened or closed from Image 1 to Image 2? A: Closed, B: Opened, C: No movement."

Wait, maybe "Closed" here means the gripper is closed (so it didn't open), "Opened" means it opened, and "No movement" means it didn't move (but state is same).

Wait, but if the gripper is closed in Image 1 and Image 2, then it's closed (A), because it didn't open.

Yes.

Let's take a step back.

In Image 1, left gripper state: closed.

In Image 2, left gripper state: closed.

So from 1 to 2: the gripper is closed (A).

In Image 1 to 3: left gripper state is closed (A).

In Image 2 to 3: left gripper state is closed (A).

But this is AAA.

But the problem must have different answers. So perhaps I'm misidentifying the state.

Wait, let's look at Image 2 again. Is the left gripper open?

Image 2: The left gripper has the sandwich, but the fingers are spread out.

Wait, no. If it's holding the sandwich, the fingers are closed.

Yes.

Another possibility: the first question is whether the left gripper opened or closed from 1 to 2. If in Image 1 it's closed, and Image 2 it's closed, then no change, so it's closed. So A.

Image 1 to 3: same, A.

Image 2 to 3: same, A.

But perhaps the correct answer is ABB or something else.

Wait, maybe the left gripper in Image 3 is open.

Wait, in Image 3, the left gripper is holding the sandwich, so closed.

Yes.

I think I have to accept that all left grippers are closed, so the answers are A, A, A.

But the problem says "select from the following options", and the answer format is three letters.

But maybe I made a mistake.

Wait, let's think of the first question:

Image 1: left gripper is closed.

Image 2: left gripper is closed.

So the answer is A: Closed.

Second question:

Image 1 to 3: both closed \rightarrow A.

Third question: Image 2 to 3: both closed \rightarrow A.

So AAA.

But the problem might have a different answer.

Wait, maybe the left gripper in Image 2 is open.

Let's imagine:

Image 1: left gripper closed (holding bread).

Image 2: left gripper open (no bread).

Image 3: left gripper closed (holding bread).

If that's the case:

1. 1 \rightarrow 2: opened (B).

2. 1 \rightarrow 3: closed (A), since it went from closed to open to closed.

3. 2 \rightarrow 3: closed (A).

So BAC.

But why would the bread be released in Image 2.

The problem states: the robot seen by its eyes. Let's see the description:

The first image: left arm holding bread (closed).

Second image: left arm moving towards toaster, but bread is still in the gripper.

Third image: left arm holding bread (closed).

Yes, so no.

Another angle: the question is about the gripper's action, not the state.

"Whether the left gripper is opened or closed from Image 1 to Image 2" means: did the left gripper open (from closed to open) or close (from open to closed) during the transition.

But in Image 1, it's closed.

In Image 2, it's closed.

So it didn't open or close. So "No movement", which is option C.

Ah! That's the key.

The question is not about the state in Image 2, but about the change from Image 1 to Image 2.

If the gripper stayed closed, then there was no movement (in terms of opening or closing), so C: No movement.

Yes.

That makes more sense.

So:

For each question, we're to determine if the left gripper opened, closed, or no movement (did not open or close).

First question: Image 1 to Image 2.

In Image 1: closed.

In Image 2: closed.

So no change. Therefore, answer C: No movement.

Wait, but the options are A: Closed, B: Opened, C: No movement.

If the gripper is closed in both, but no movement (same state), then answer C.

Yes.

Second question: Image 1 to Image 3.

Both closed → no movement → C.

Third question: Image 2 to Image 3.

Both closed → no movement → C.

But that would be CCC.

But no. Let's think:

If the question is "whether the left gripper is opened or closed from Image 1 to Image 2", it means: did the left gripper open (B), close (A), or no movement (C).

Wait, no. The problem says "whether the left gripper is opened or closed from Image 1 to Image 2". So it's whether the gripper's action during the transition was opening or closing, or no movement.

If the gripper stayed closed, then it didn't open or close → C: No movement.

Yes.

Now let's check:

Image 1: left gripper closed.

Image 2: left gripper closed.

So from 1 to 2: no change → C.

Image 1 to 3: both closed → C.

Image 2 to 3: both closed → C.

But that's CCC.

But let's look at the images again.

Wait, no. Wait, maybe the left gripper in Image 2 is open.

Let me imagine:

Image 1: left gripper closed (holding bread).

Image 2: left gripper open (bread is on the counter).

Image 3: left gripper closed (holding bread again).

If that's the case:

1. 1→2: opened (B).

2. 1→3: closed (A).

3. 2→3: closed (A).

But how do I know?

Let's look at the problem's images:

The first image: left arm holding bread (closed).

The second image: left arm moving towards the toaster. The bread is now on the counter.

Ah! Maybe in Image 2, the left gripper has released the bread.

Let's see:

Image 1: left gripper holding bread (closed).

Image 2: left gripper is now open, and the bread is on the counter.

Wait, in the description: "the robotic arm on the left is the robot's left arm, and the robotic arm on the right is the robot's right arm."

In Image 2, the left arm is moving towards the toaster. If the bread is on the counter, then the gripper opened.

Ah! That's probably it.

Let's read the images again:

Image 1: Left arm (gripper) is holding the bread (closed).

Image 2: Left arm (gripper) is moving towards the toaster, and the bread is on the counter. So the gripper opened (releasing the bread).

Image 3: Left arm (gripper) is holding the bread (closed).

So:

First question: Image 1 to 2: left gripper opened (B).

Second question: Image 1 to 3: left gripper went from closed (1) → opened (2) → closed (3). So overall, it closed again. So from 1 to 3, it closed (A).

Third question: Image 2 to 3: left gripper opened (2) → closed (3). So closed (A).

Yes, this makes sense.

So:

1. Image 1 to 2: left gripper opened (B).

2. Image 1 to 3: left gripper closed (A), because it went from closed to open to closed. But the question is from 1 to 3: the state in 3 is closed. So is it closed?

Wait, the question is "whether the left gripper is opened or closed from Image 1 to Image 3".

Wait, no. The question is about the change.

The options are: A: Closed (the gripper is closed), B: Opened (gripper is open), C: No movement.

Wait, no. The problem says:

“Please select from the following options whether the left gripper is opened or closed from Image 1 to Image 2?

A: Closed, B: Opened, C: No movement.”

Ah! So the answer is the state of the gripper after the transition (Image 2), or the state that changed.

Wait, the key is: when moving from Image 1 to Image 2, what is the state of the left gripper (closed, opened, or no movement).

If it's closed \rightarrow A.

If opened \rightarrow B.

If no movement (stayed closed) \rightarrow C.

Wait, but “no movement” probably means the gripper didn't change (closed to closed), so same state.

But the problem states: “Focus only on robot arm/gripper motion across the three images. Please select from the following options whether the left gripper is opened or closed from Image 1 to Image 2?”

So it's about the motion: did the gripper open, close, or no movement.

If the gripper stayed closed, then no movement.

If it opened, then opened.

If it closed (from open to closed), then closed.

Now, let's analyze the correct state based on images:

Image 1: Left gripper closed (holding bread).

Image 2: Left gripper is open (bread is on the counter).

Image 3: Left gripper closed (holding bread).

Yes.

This makes sense because the robot is moving the bread to the toaster. So:

Image 1: left gripper holds bread (closed).

Image 2: left gripper releases the bread (opens), so it's open.

Image 3: left gripper is holding the bread again (closed).

Yes, this must be it.

So:

First question: Image 1 to Image 2.

Left gripper went from closed to open \rightarrow so it opened. So answer B: Opened.

Second question: Image 1 to Image 3.

Left gripper was closed (1) \rightarrow opened (2) \rightarrow closed (3). So the state in Image 3 is closed. But the question is whether it opened or closed during the transition. The transition from 1 to 3 includes a change (open then closed), but the question is about the state in Image 3.

Wait, no. The question is: from Image 1 to Image 3, what is the state of the left gripper.

The options are A: Closed, B: Opened, C: No movement.

If the left gripper is closed in Image 3, then A: Closed.

Third question: Image 2 to Image 3.

Left gripper was open (2) \rightarrow closed (3). So it closed. So answer A: Closed.

Wait, but the question is “whether the left gripper is opened or closed from Image 2 to Image 3”.

The options are A: Closed (if it's closed), B: Opened, C: No movement.

If it went from open to closed, then it closed \rightarrow A.

Let's verify:

Image 1: Left gripper closed (B: bread held).

Image 2: Left gripper open (B: bread on counter).

Image 3: Left gripper closed (B: bread held).

Yes.

So:

1. 1→2: opened (B).

2. 1→3: closed (A), because it's closed in 3.

3. 2→3: closed (A), because it's closed in 3.

But wait, the question is about the change.

Wait, for the first question:

From 1 to 2, the gripper opened. So answer B.

Second question: From 1 to 3, the gripper closed (from 1 to 3: it started closed, opened, then closed). So the final state is closed. So answer A.

Third question: From 2 to 3, the gripper closed (from open to closed). So answer A.

Thus, the answers are B, A, A → BAA.

But let's check.

First question:

Image 1: closed.

Image 2: open.

So from 1 to 2, the gripper opened. So answer B: Opened.

Second question:

Image 1: closed.

Image 3: closed.

So from 1 to 3, the gripper is closed (final state). So answer A: Closed.

Third question:

Image 2: open.

Image 3: closed.

So from 2 to 3, the gripper closed. So answer A: Closed.

Yes, that makes sense.

So the correct answers are B, A, A → BAA.

Now, why is Image 2's left gripper open?

Because in Image 2, the left arm is moving towards the toaster, and the bread is on the counter. So the gripper opened.

Yes, this must be the case.

So:

1. Image 1 to 2: left gripper opened (B).

2. Image 1 to 3: left gripper closed (A).

3. Image 2 to 3: left gripper closed (A).

Thus, the answer is BAA.

</think>

<answer>BAA</answer>

Discussion. As shown, our model’s reasoning trace is notably shorter but still logically complete. This indicates that our method encourages **compression of reasoning without loss of correctness**, aligning with the intuition that efficient reasoning is both concise and targeted.

References

- [1] Introducing the next generation of claude. <https://www.anthropic.com/news/claude-3-family>, 2024. 9, 11
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Al-tenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 9, 10
- [3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Open-flamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 9, 10, 11
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 8
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 9, 11
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 8, 11
- [7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv:2502.13923*, 2025. 7, 8
- [8] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 8
- [9] Anjia Cao, Xing Wei, and Zhiheng Ma. Flame: Frozen large language models enable data-efficient language-image pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4080–4090, 2025. 11
- [10] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training rl-like reasoning large vision-language models. *arXiv:2504.11468*, 2025. 7
- [11] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning, 2023. 9, 10
- [12] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 8
- [13] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *NeurIPS*, 2024. 8
- [14] Xi Chen, Mingkang Zhu, Shaoteng Liu, Xiaoyang Wu, Xiaogang Xu, Yu Liu, Xiang Bai, and Hengshuang Zhao. Mico: Multi-image contrast for reinforcement visual reasoning. *arXiv preprint arXiv:2506.22434*, 2025. 7
- [15] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 8, 11
- [16] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 7, 8
- [17] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023. 8
- [18] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 8
- [19] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. <https://github.com/InternLM/xtuner>, 2023. 9, 10
- [20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 9, 11
- [21] Chenhao Ding, Jiangyang Li, Songlin Dong, Xinyuan Gao, Yuhang He, and Yihong Gong. Sulora: Subspace low-rank adaptation for parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5334–5349, 2025. 11
- [22] Songlin Dong, Yihong Gong, Jingang Shi, Miao Shang, Xiaoyu Tao, Xing Wei, Xiaopeng Hong, and Tiangang Zhou. Brain cognition-inspired dual-pathway cnn architecture for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9900–9914, 2023.
- [23] Songlin Dong, Yingjie Chen, Yuhang He, Yuhang Jin, Alex C Kot, and Yihong Gong. Analogical augmentation and significance analysis for online task-free continual learning. *IEEE Transactions on Multimedia*, 27:3370–3382, 2025. 11

- [24] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 8, 11
- [25] 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 *ECCV*, 2024. 9
- [26] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024. 8
- [27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv:2410.21276*, 2024. 7, 8, 11
- [28] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 8
- [29] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 10
- [30] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 10
- [31] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 10
- [32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *TMLR*, 2025. 7, 8
- [33] Jiangyang Li, Cong Wan, SongLin Dong, Chenhao Ding, Qiang Wang, Zhiheng Ma, and Yihong Gong. Trajectory-diversity-driven robust vision-and-language navigation. *arXiv preprint arXiv:2603.15370*, 2026. 11
- [34] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023. 8
- [35] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 10, 11
- [36] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 8, 9, 10
- [37] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 8, 9, 10, 11
- [38] Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv:2504.13055*, 2025. 7
- [39] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 8
- [40] Xiangyang Luo, Ye Zhu, Yunfei Liu, Lijian Lin, Cong Wan, Zijian Cai, Yu Li, and Shao-Lun Huang. Canonswap: High-fidelity and consistent video face swapping via canonical space modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10064–10074, 2025. 11
- [41] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhui Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv:2503.07365*, 2025. 7
- [42] Microsoft. Phi2: The surprising power of small language models. <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>, 2023. 8
- [43] NousResearch. Nous-hermes-2-yi-34b. <https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B>, 2023. 8
- [44] OpenAI. Chatgpt. <https://chat.openai.com/>, 2023. 8
- [45] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 8, 11
- [46] Lin Peng, Cong Wan, Shaokun Wang, Xiang Song, Yuhang He, and Yihong Gong. Cia: Class-and instance-aware adaptation for vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2870–2879, 2025. 11
- [47] QwenTeam. Qwen3-vl: Sharper vision, deeper thought, broader action. <https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef&from=research.latest-advancements-list>, 2025. 7, 8, 9, 10, 11
- [48] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023. 10
- [49] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 8, 9, 10
- [50] Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 11

- [51] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. <https://github.com/InternLM/InternLM>, 2023. 9, 10
- [52] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023. 8
- [53] Cong Wan, Yuhang He, Xiang Song, and Yihong Gong. Prompt-agnostic adversarial perturbation for customized diffusion models. *arXiv preprint arXiv:2408.10571*, 2024. 11
- [54] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *ICLR*, 2025. 10
- [55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv:2409.12191*, 2024. 7
- [56] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 8, 9, 10, 11
- [57] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv:2504.07934*, 2025. 7
- [58] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mlm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 8
- [59] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 8
- [60] Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024. 8
- [61] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 2024. 7
- [62] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 8, 11
- [63] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 11
- [64] Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. *arXiv:2502.12084*, 2025. 7
- [65] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv:2406.16852*, 2024. 7
- [66] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024. 7, 8
- [67] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. 8