

Video-Only ToM: Enhancing Theory of Mind in Multimodal Large Language Models

Supplementary Material

6. MLLMs with EgoToM baseline

We provide a complete introduction here to all the multimodal large language models (MLLMs) that we compared or used in the EgoToM benchmark. Specifically, the EgoToM benchmark includes human baselines, LLMs baselines, and MLLMs baselines. These models represent the state-of-the-art in text and vision-language processing. In our work, we focus on the performance of MLLMs, selecting representative open-source and closed-source models from the EgoToM benchmark, including GPT-4-Turbo [37], Video-Llama2-72B [9], and CogVLM2 [15]. Additionally, we also added two widely-used closed-source models: GPT-4o [36] and Gemini-2.5-Flash [11], as well as open-source models LLaVA-Next-Video-7B [56], Qwen2.5-VL-7B [3], and the reasoning model GLM-4.1V-9B-Thinking [16]. Furthermore, we selected LLaVA-Next-Video-7B and Qwen2.5-VL-7B models as the base models for VisionToM, and through our method, both models achieved better performance.

We conducted quantification based on the metric charts in the EgoToM [30] benchmark and list the complete data baseline here in Table 4, along with additional experiments we added.

Table 4. Full Baseline and results

Method	Setting	Context	Nframe	Accuracy		
				Goal	Belief	Actions
Humans	Video	last 30sec	-	0.88	0.72	0.78
		last 5sec		0.89	0.71	0.77
Llama3.1-405b-instruct	Text	full context	-	0.82	0.44	0.48
		last 30sec		0.80	0.46	0.46
		last 5sec		0.62	0.45	0.43
		last action		0.58	0.40	0.38
		no context		0.20	0.30	0.15
Llama3.1-70b-instruct	Text	full context	-	0.80	0.34	0.47
		last 30sec		0.80	0.42	0.45
		last 5sec		0.65	0.41	0.42
		last action		0.60	0.36	0.38
		no context		0.28	0.25	0.18
Llama3.1-8b-instruct	Text	full context	-	0.80	0.40	0.36
		last 30sec		0.78	0.42	0.38
		last 5sec		0.65	0.41	0.40
		last action		0.67	0.39	0.34
		no context		0.35	0.36	0.22
GPT-4-Turbo	Video	full context	20	0.83	0.45	0.42
		last 30sec		0.87	0.53	0.44
		last 5sec		0.85	0.51	0.47
		last action		0.78	0.50	0.41
		no context		0.15	0.18	0.06
GPT-4-Turbo	Text	full context	-	0.85	0.47	0.44
		last 30sec		0.82	0.48	0.45

Continued on next page

Method	Setting	Context	Nframe	Accuracy		
				Goal	Belief	Actions
GPT-4-Turbo	Text	last 5sec	-	0.68	0.44	0.36
		last action		0.60	0.34	0.32
		no context		0.15	0.18	0.06
GPT-4	Text	full context	-	0.86	0.46	0.47
		last 30sec		0.82	0.48	0.43
		last 5sec		0.70	0.42	0.41
		last action		0.61	0.40	0.38
		no context		0.20	0.28	0.18
GPT-3.5-Turbo	Text	full context	-	0.70	0.29	0.23
		last 30sec		0.70	0.32	0.21
		last 5sec		0.65	0.34	0.22
		last action		0.58	0.30	0.21
		no context		0.15	0.23	0.15
VideoLLaMA2-72B	Video	full context	8	0.85	0.46	0.40
		last 30sec		0.86	0.48	0.42
		last 5sec		0.85	0.50	0.45
		last action		0.83	0.54	0.47
		no context		0.21	0.30	0.14
VideoLLaMA2-7B-16F	Video	full context	16	0.67	0.33	0.30
		last 30sec		0.71	0.34	0.32
		last 5sec		0.73	0.41	0.34
		last action		0.66	0.39	0.36
		no context		0.32	0.25	0.19
VideoLLaMA2-7B	Video	full context	8	0.79	0.41	0.31
		last 30sec		0.75	0.42	0.32
		last 5sec		0.75	0.40	0.40
		last action		0.52	0.33	0.35
		no context		0.32	0.28	0.21
CogVLM2	Video	full context	24	0.73	0.39	0.36
		last 30sec		0.75	0.40	0.38
		last 5sec		0.77	0.42	0.41
		last action		0.53	0.34	0.32
		no context		0.21	0.29	0.30
GPT-4o				0.69	0.20	0.23
Gemini-2.5-Flash				0.86	0.47	0.40
GLM-4.1V-9B-Thinking	Video	full context	24	0.80	0.31	0.26
LLaVA-Next-Video-7B				0.62	0.39	0.24
Qwen2.5-VL-7B				0.87	0.36	0.31

7. Additional Generalization Results

We report two additional experiments promised in the rebuttal: scaling VisionToM to a stronger backbone and transferring the learned directions to a second video-only ToM benchmark.

7.1. Large-Backbone Results

Table 5. Additional experiments with a larger MLLM backbone. VisionToM continues to improve ToM reasoning when scaled to Qwen2.5-VL-72B.

Method	Setting	Nframe	Accuracy (%) \uparrow		
			Goal	Belief	Actions
Gemini-3-Flash-Preview	Baseline	24	91.5	51.3	55.6
Qwen3-VL-235B-A22B-Instruct			90.2	51.3	44.4
Qwen2.5-VL-72B-Instruct	Baseline	24	87.0	43.6	37.1
	w/o $\delta_{T,l}^b$		89.8	45.3	40.3
	w/o $\delta_{V,l}^b$		89.0	50.9	49.2
	Rnd- Δ		86.2	43.6	39.1
	$-\alpha\Delta$		79.7	35.5	25.8
	$+\alpha\Delta$	91.5	59.8	57.3	

Table 5 shows that VisionToM remains effective on Qwen2.5-VL-72B. The intervention improves the 72B backbone on all three EgoToM tasks and surpasses strong large-model baselines on Belief and Actions, indicating that the method remains beneficial even when the base MLLM is already strong.

7.2. Cross-Dataset Transfer on MMTToM-QA

Table 6. Experiments on the MMTToM-QA benchmark under the video-only setting. “Transfer” directly applies the intervention vector learned on EgoToM without retraining on MMTToM-QA.

Method	Setting	Nframe	Belief	Goal	All
BIP-ALM-LLaMA2	Video Only	-	64.0	58.3	61.2
Qwen2.5-VL-7B-Instruct		47.0	29.3	38.2	
VisionToM w/ Qwen2.5-VL-7B-Instruct		8	70.7	62.0	66.3
Qwen2.5-VL-7B-Instruct Transfer		64.3	56.7	60.5	

On MMTToM-QA, we evaluate both in-domain generalization and zero-shot transfer under the video-only setting. Following the same protocol as EgoToM, we compute intervention vectors from the MMTToM-QA training split and evaluate on its benchmark. VisionToM achieves the best overall performance, while directly transferring the intervention vector learned on EgoToM yields results close to the strongest video-only baseline. These findings suggest that the learned directions capture transferable ToM reasoning patterns rather than dataset-specific shortcuts.

8. Additional Probing Results

The probing results on the LLaVA-Next-Video and Qwen2.5-VL models are shown in Figures 4 and 5, covering two stages: visual attention probing and ToM reasoning probing. Each stage includes independent probing of Goal, Belief, and Actions tasks, with the y-axis representing attention layers and the x-axis representing attention heads.

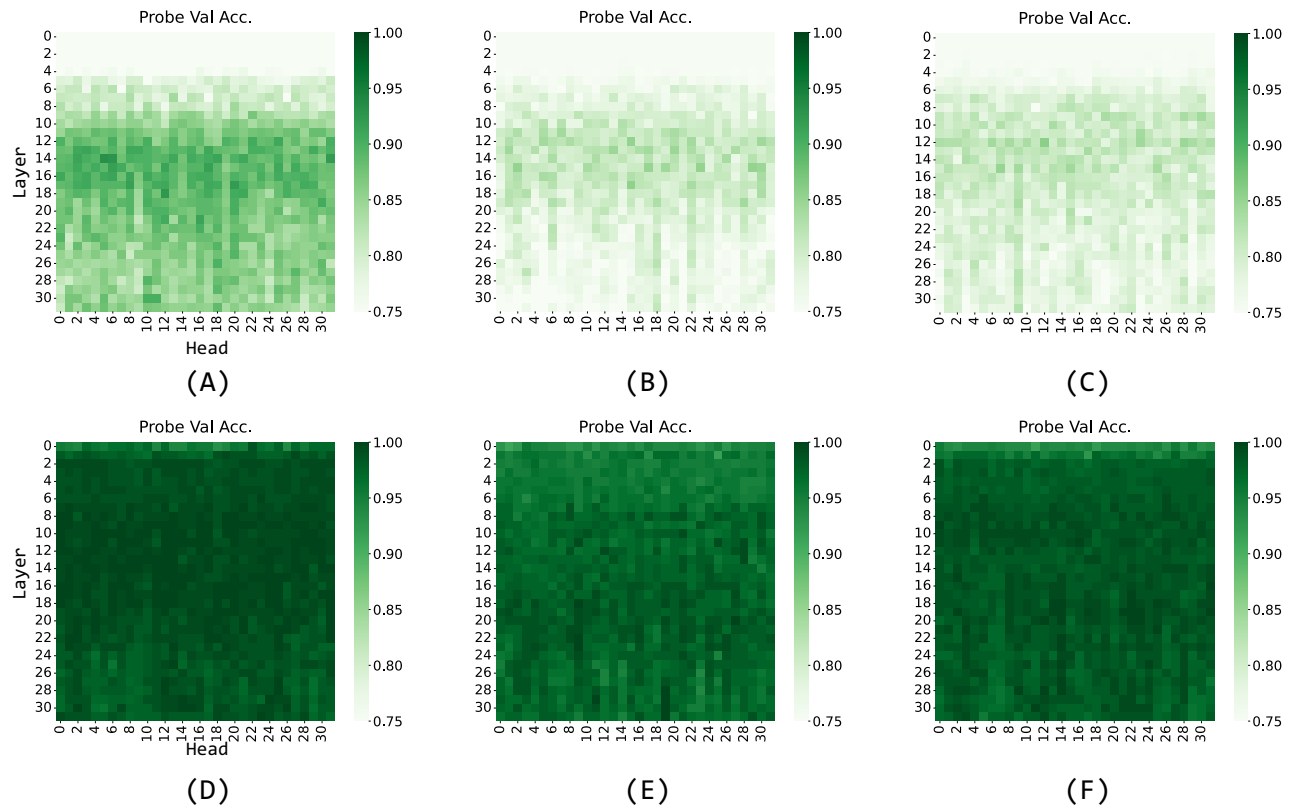


Figure 4. Probe validation accuracies for the three EgoToM tasks, based on activations from each attention head across all layers of LLaVA-Next-Video-7B. Subfigures (A)–(C) correspond to the ToM reasoning stage, showing accuracies for the (A) goal prediction, (B) belief inference, and (C) actions inference tasks, respectively. Subfigures (D)–(F) correspond to the visual attention stage, showing the same tasks in the order: (D) goal prediction, (E) belief inference, and (F) actions inference. Darker shades indicate higher probing accuracy, suggesting stronger task-relevant signals in specific heads and layers.

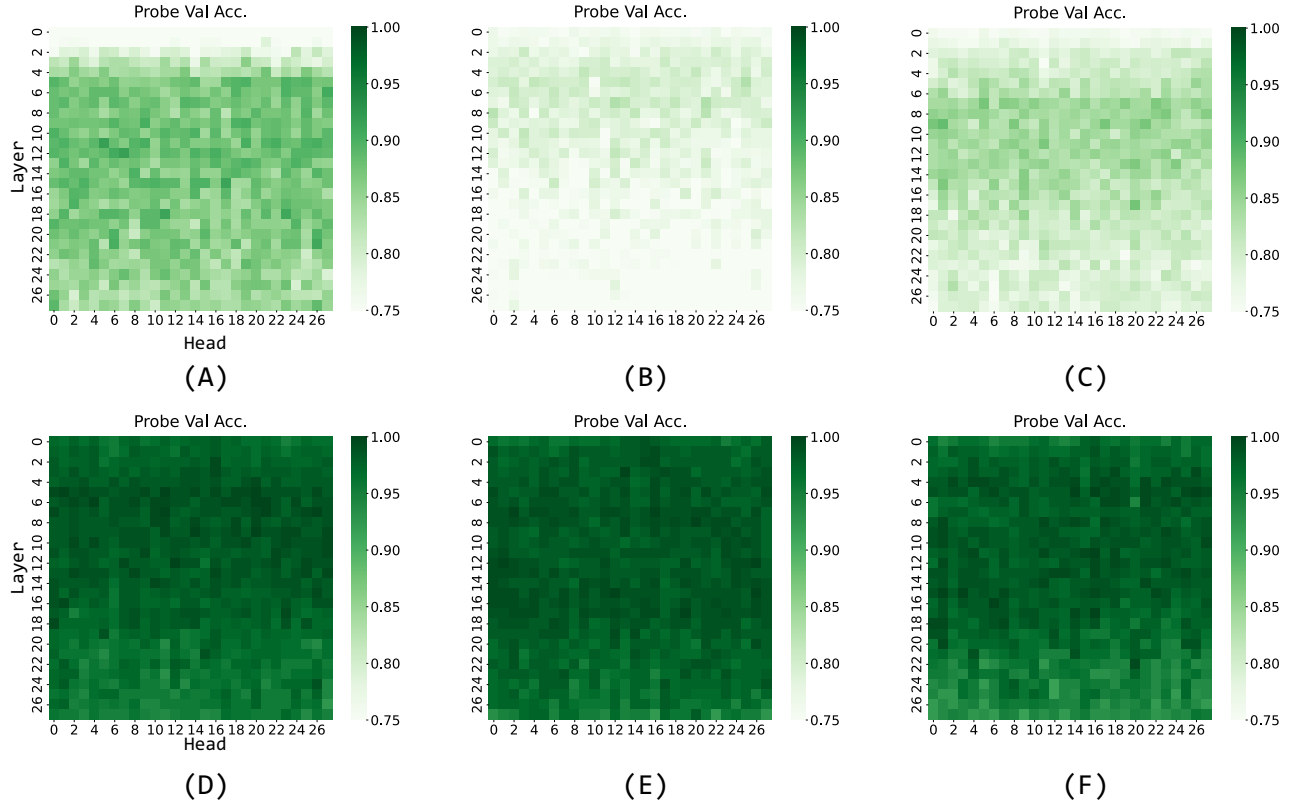


Figure 5. Probe validation accuracies for the three EgoToM tasks, based on activations from each attention head across all layers of Qwen2.5-VL-7B. Subfigures (A)–(C) correspond to the ToM reasoning stage, showing accuracies for the (A) goal prediction, (B) belief inference, and (C) actions inference tasks, respectively. Subfigures (D)–(F) correspond to the visual attention stage, showing the same tasks in the order: (D) goal prediction, (E) belief inference, and (F) actions inference. Darker shades indicate higher probing accuracy, suggesting stronger task-relevant signals in specific heads and layers.

9. Hyperparameters’ Analysis

Figure 6 and Figure 7 respectively show the intervention effects of the number of editing heads K and intervention strength α on the LLaVA-Next-Video model and Qwen2.5-VL model across three tasks in the EgoToM benchmark. The three subplots correspond to: (A) Goal Task, (B) Belief Task, and (C) Actions Task.

For the editing heads K of the LLaVA-Next-Video model, we choose to use 16, 32, 64 based on its attention head count of 32. For the Qwen2.5-VL model with an attention head count of 28, we choose to use 14, 28, 56 as the editing heads K . We did not search for the optimal editing heads K to achieve the best results.

Theoretically, whether for visual attention enhancement or ToM reasoning guidance, the obtained δ represents a correction from negative samples pointing to positive samples, so $+\alpha$ intervention will bring positive gains, while $-\alpha$ will weaken model capabilities. The results in Figure 6 and 7 strongly support our hypothesis, with intervention effects showing monotonic behavior around the baseline ($\alpha = 0$) and presenting uniform and coherent characteristics within the effective range. Additionally, the improvement effects brought by VisionToM intervention are not unlimited, but are only effective within a certain intervention strength range (for the LLaVA-Next-Video model, this range is $\alpha \in [-5, 5]$). Beyond this range, all responses become invalid. Specifically, as shown in Figure 6(A), when $\alpha=4$, the accuracy rates for $K=16, 32$, and 64 all show a declining trend. Unlike the decline when $\alpha=-1$, the performance degradation here is mainly due to excessive interference intensity, causing some responses to become invalid (such as outputting garbled text or infinitely repeating words). During the statistical process, we retained all samples and treated invalid responses as errors to ensure consistency in comparison. The same phenomenon was observed in other experiments, indicating that our method has controllability and remains effective within a certain range of interference.

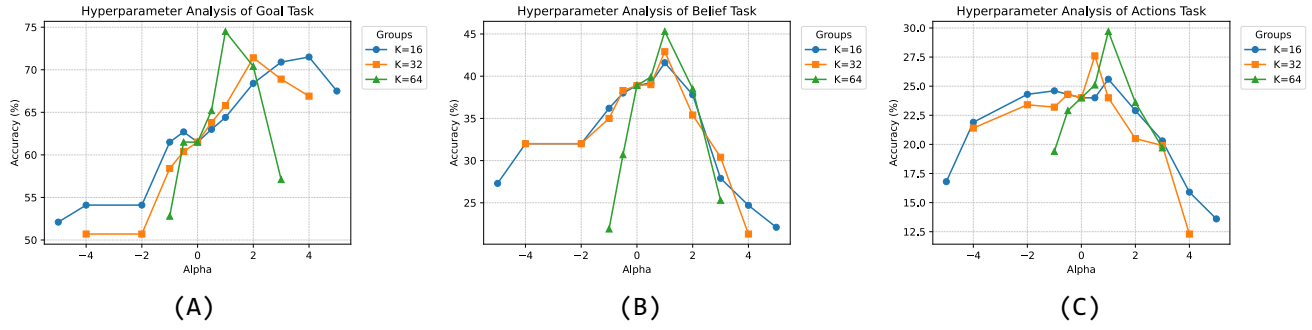


Figure 6. Analysis of the Impact of Hyperparameter of LLaVA-Next-Video on Three Tasks

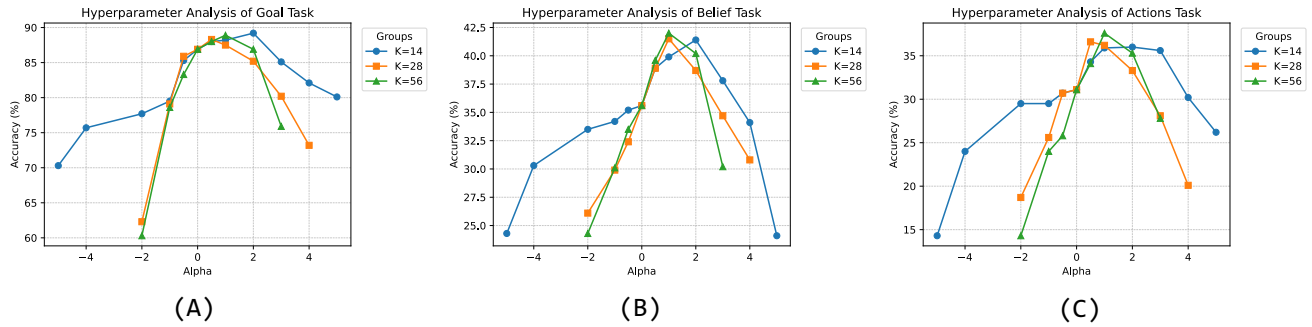
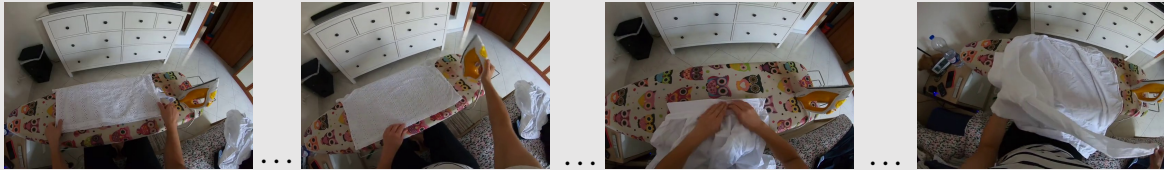


Figure 7. Analysis of the Impact of Hyperparameter of Qwen2.5-VL on Three Tasks

10. Open-ended Generation

We have listed examples of open-ended generation here, including the responses from the base model as well as the improved effects after applying our method.

INPUTS: For the photographer named C of this video, at the end of these actions, what will C most likely do next?

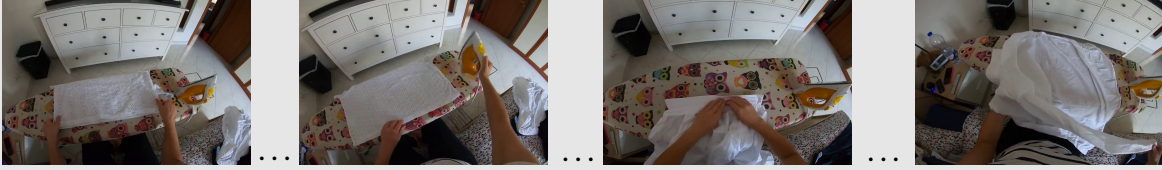


Answer: C adjusts the sleeve, opens the button of the sleeve, and irons the sleeve.

LLaVA-Next-Video: Based on the information in the video, I cannot make an inference.

LLaVA-Next-Video+VisionToM: Based on the video, it appears that C is in the process of ironing a piece of clothing, possibly a shirt or a garment. At the end of this action, C is likely to adjust the sleeve and then ironing it.

INPUTS: For the photographer named C of this video, at the end of these actions, what will C most likely do next?

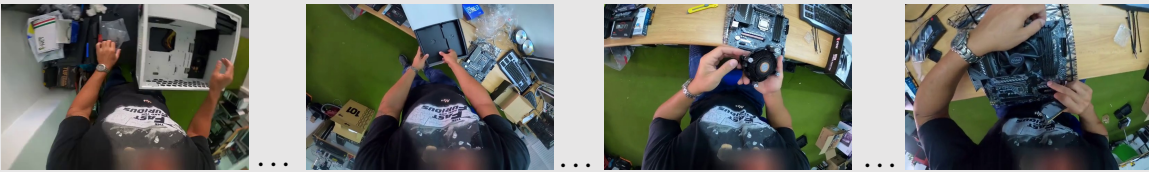


Answer: C adjusts the sleeve, opens the button of the sleeve, and irons the sleeve.

Qwen2.5-VL: Based on the information in the video, I cannot make an inference.

Qwen2.5-VL+VisionToM: The video shows C ironing what appears to be a shirt, and once that's done, C will realign the sleeve and iron it.

INPUTS: For the photographer named C of this video, at the end of these actions, what does C most likely believe?

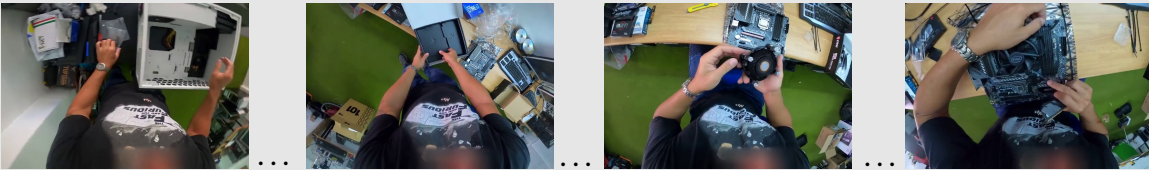


Answer: C believes that the CPU assembly requires additional components to function optimally.

LLaVA-Next-Video: Based on the information in the video, I cannot make an inference.

LLaVA-Next-Video+VisionToM: In the video, C is installing computer equipment, and by the end of the video the equipment is still not ready, appearing to require further installation.

INPUTS: For the photographer named C of this video, at the end of these actions, what does C most likely believe?



Answer: C believes that the CPU assembly requires additional components to function optimally.

Qwen2.5-VL: Based on the video, it is likely that C believes that the computer hardware they are working on is now fully assembled and functional. This conclusion can be drawn from the fact that they are shown carefully placing the CPU into the motherboard, which is a critical step in the assembly process, and the subsequent actions of connecting the power supply and other components. The final shot shows the completed setup, suggesting that the hardware is now ready for use.

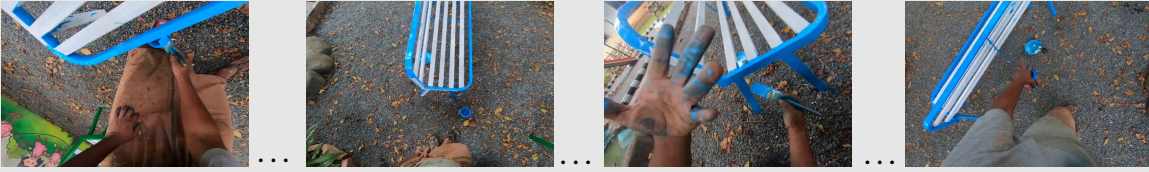
Qwen2.5-VL+VisionToM: From the video, it can be seen that C installs the CPU into the motherboard socket. The final shot shows scattered hardware on the motherboard and does not present the complete powered-on state, indicating that the hardware assembly is not yet complete and still requires adding and properly connecting the remaining components before it can actually be put into use.

11. Experiments Settings

11.1. Data preprocessing

We follow the experimental setup of the EgoToM dataset, extracting corresponding video segments from the Ego4D dataset based on the timeline it provides. EgoToM includes three ToM tasks: Goal, Belief, and Actions, with sample sizes of 351, 335, and 354 respectively. According to the experimental protocol, we sample video frames at equal intervals from each video

INPUTS: For the photographer named C of this video, what is C's future goal?

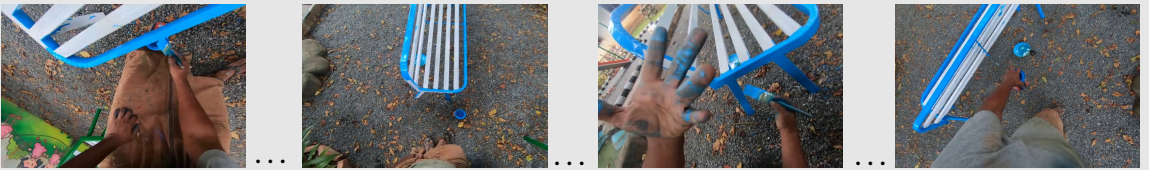


Answer: To complete painting the playground equipment.

LLaVA-Next-Video: Based on the video, C's future goal appears to be to become a professional photographer. The video shows C painting a chair with a brush, which suggests that they are actively engaged in creating or improving their work, which is a common pursuit for artists and photographers. Additionally, C is seen holding a camera and taking pictures, which indicates their interest in photography. The fact that they are painting a chair could be seen as a metaphor for their desire to improve their craft and create visually appealing work, which is a common goal for photographers. Therefore, it is likely that C's future goal is to become a professional photographer.

LLaVA-Next-Video+VisionToM: Based on the video content, C's future goal should be to complete the painting work on the playground equipment. In the video, C uses a brush to paint the equipment, and the camera finally focuses on parts of the equipment that have not yet been painted, suggesting that the entire project has not been finished.

INPUTS: For the photographer named C of this video, what is C's future goal?



Answer: To complete painting the playground equipment.

Qwen2.5-VL: The video does not provide any information about the photographer's future goals. The video only shows a person painting a bench.

Qwen2.5-VL+VisionToM: According to the video information, C is in a park and is painting a bench.

segment and input these frames along with the corresponding questions into the model as the sole source of information for reasoning. For each task, we use a 30% calibration split to train the probe and encoder and compute intervention vectors, and a disjoint 70% evaluation split for final testing. No labels or answers from the evaluation split are used when learning the intervention directions.

11.2. Computing infrastructure

To ensure reproducibility, all experiments were conducted under the following computing environment: Ubuntu 22.04; 14 vCPUs on an Intel® Xeon® Gold 6348 @ 2.60 GHz; 8×NVIDIA A800 GPUs; and 100 GB system memory. The software stack consists of Python 3.12, PyTorch 2.5.1, and CUDA 12.4. We fixed the global random seed to 42 and enabled deterministic settings to eliminate randomness from data loading and operator-level execution. Both training and inference were performed in FP16 precision.

11.3. Calibration Cost

VisionToM keeps the MLLM backbone frozen. On the hardware reported in Section 11.2, the one-time calibration stage takes approximately 0.2 hours for probe training and 1 hour for encoder training. All downstream experiments, including multiple-choice QA, open-ended generation, large-model evaluation, and MMTToM-QA transfer, directly apply the resulting precomputed intervention vectors without further training.

11.4. Open-ended Evaluation Details

For each open-ended answer, two DeepSeek-R1 judges are prompted independently, and we accept a label only when both judges agree. The prompt explicitly defines the “true” and “info” criteria and standardizes edge cases. In particular, an answer

is marked “false” if any factual statement is incorrect, hallucinated, logically contradictory, or inconsistent with the reference facts; answers that mix correct and incorrect claims are also marked “false”. An answer is marked “info” only if it contains substantive task-relevant content rather than vague restatements. We additionally performed manual verification with three volunteers and observed human–LLM agreement rates of 96.2% for the “true” label and 93.5% for the “info” label.