

Vision-Language Model as a Judge for Robot Policy A/B Evaluation

Toan Nguyen¹

¹University of Southern California

Correspondence: tientoan@usc.edu

Abstract

Policy evaluation plays a critical role in robot learning. However, current robotic policy evaluation predominantly relies on unexpressive scalar metrics, while detailed, language-based assessments are largely performed by human evaluators, making the process slow, expensive, and difficult to scale. Meanwhile, vision-language models (VLMs) are increasingly leveraged as automated judges across a wide range of domains, owing to their broad world knowledge, commonsense reasoning, and multimodal understanding. Motivated by this, in this work, we formulate robot policy A/B evaluation as a multimodal learning problem and fine-tune VLMs on curated human policy evaluation data. Our experiments demonstrate that fine-tuning on high-quality human annotations dramatically improves the judgment capability of Qwen-VL models, yielding evaluations that more faithfully reflect human preferences. Furthermore, we show that our fine-tuned Qwen3-VL-8B judge can be effectively leveraged for data curation in robotic imitation learning: the policy trained on data selected by our model achieves an 80.5% success rate on the LIBERO-Object benchmark, significantly outperforming that with data curated by the base model.

“If you can’t measure it, you can’t improve it.”

— Lord Kelvin

1 Introduction

Policy evaluation plays a fundamental role in developing robot learning models (Barreiros et al., 2026). Common approaches to robot policy evaluation and benchmarking typically rely on scalar metrics like success rate or the dynamic time warping similarity (Senin, 2008) between expert and rollout robot trajectories. While these metrics are simple and efficient to implement, they fail to capture many important aspects of real-world robot

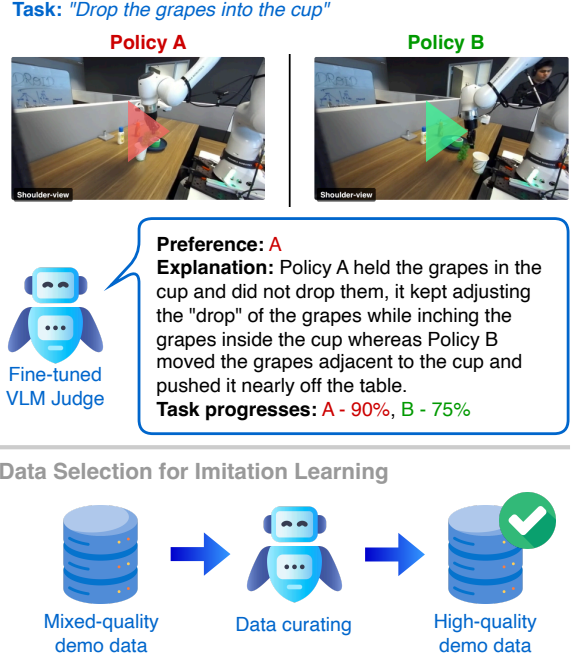


Figure 1: **VLM Judge for Policy Evaluation Overview.** *Top:* Our fine-tuned VLM judge can automate the robot policy A/B evaluation process (select the better policy, give task progress scores, and provide free-form text explanations); *Bottom:* The VLM judge can also be used for demonstration data selection for imitation learning.

behavior, including partial progress, efficiency, robustness, and adherence to language instructions, thus making it difficult to meaningfully distinguish between policies or identify specific failure modes in complex, long-horizon tasks.

As robot learning shifts toward generalist policies (Sapkota et al., 2025) capable of handling open-world tasks conditioned on language instructions, recent work has highlighted the importance of incorporating human preferences and free-form feedback to better capture nuanced performance differences (Atreya et al., 2025; Jangir et al., 2026). However, this evaluation paradigm remains fundamentally limited: collecting preference judgments, progress scores, and textual explanations from hu-

man evaluators is costly, time-consuming, and difficult to scale across the diverse task and configuration distributions that generalist robots are expected to handle. Recent advances in large language models (LLMs) and vision-language models (VLMs) offer a promising path forward, as these models have demonstrated strong multimodal reasoning abilities and broad world knowledge (Yin et al., 2024) that closely mirror the judgment capabilities required for holistic robot evaluation. This makes them compelling for automating the assessment of task completion, instruction grounding, and behavioral quality in a scalable and reproducible manner.

Motivated by this trend, in this work, we formulate robot policy A/B evaluation as a multimodal learning problem and investigate training VLMs as automated judges. Specifically, we fine-tune VLMs on high-quality human evaluation data (Atreya et al., 2025) such that, given a pair of rollout videos from two anonymized policies and a natural language task description, the model predicts (i) a pairwise preference indicating which policy performed better, (ii) a task progress score for each policy, and (iii) a free-form natural language explanation justifying the judgment. We demonstrate that fine-tuning on high-quality human annotations substantially improves VLMs’ ability to approximate human evaluator outputs, enabling efficient, consistent, and scalable policy assessment without requiring human involvement. We also show that our fine-tuned Qwen3-VL-8B judge generalizes to evaluations on an SO100 unseen robot arm (Cadenne et al., 2024). Beyond evaluation, we show that our VLM judge can also be effectively repurposed for filtering robot demonstration data, owing to its learned ability to assess task progress and behavioral quality from video observations. In particular, we show that data curated by the VLM judge can be used for training an imitation learning model that obtains an 80.5% success rate on the LIBERO-Object benchmark (Liu et al., 2023). The result suggests the broader utility of VLM judges across a range of robot learning applications. An overview of our work is illustrated in Figure 1.

In summary, our contributions are two-fold:

- We fine-tune Qwen-VL models for the robot policy evaluation task, demonstrating that fine-tuning on high-quality human evaluation data significantly boosts the judgment capability and enables generalizability to an unseen robot embodiment.

- We further utilize the trained VLM judge for the task of robotic demonstration data selection, showing that data curated by our judge is of high quality and effectively benefits the downstream imitation learning.

2 Related Work

2.1 Robot Policy Evaluation

Evaluating robot policies has traditionally relied on scalar metrics such as success rate or trajectory similarity (Zhu et al., 2020; Brohan et al., 2023; Senin, 2008; Ming et al., 2025). While simple and reproducible, these metrics reduce complex behaviors to a single number and often fail to capture important aspects such as motion smoothness, safety, and partial task completion. To address these limitations, RoboArena (Atreya et al., 2025) introduces a decentralized evaluation paradigm in which human evaluators compare policies via pairwise A/B judgments and provide natural language explanations, enabling more nuanced and human-aligned assessment across diverse tasks. However, this approach relies heavily on human annotators’ effort, making it costly and difficult to scale to the large number of evaluations required during policy development. In this work, we instead fine-tune VLMs as automated evaluators, effectively retaining the semantic richness of human judgment while removing the bottleneck of human involvement.

2.2 VLM-as-a-Judge

The use of language and vision-language models as automated judges has emerged as a promising paradigm for scalable, human-aligned evaluation. Prior work on LLM-as-a-Judge (Zheng et al., 2023) and its multimodal extension (Lee et al., 2024) demonstrate strong agreement with human preferences across diverse tasks. VLM-based evaluators have since been applied to domains such as image, video, and 3D generation (Lin et al., 2024; He et al., 2024; Wu et al., 2024). However, applying this paradigm to robot policy evaluation remains challenging, as it requires reasoning over temporally extended, contact-rich behaviors and task-relevant spatial interactions from video observations. Moreover, rollout videos usually present suboptimal, inefficient, or jittery robot actions, which are common in robotics but overlooked in the training data of VLMs (Tian et al., 2026). Recent works such as RoboMeter (Liang et al., 2026), RoboReward (Lee et al., 2026), and TOPReward (Chen et al., 2026)

explore VLM-based reward modeling by predicting scalar progress signals, but these approaches lack the interpretability and semantic richness of language-based evaluation. In contrast, our work fine-tunes VLMs to jointly produce pairwise preferences, progress scores, and natural language explanations, enabling more interpretable and human-aligned A/B evaluation of robot learning policies.

3 Methodology

3.1 Human Evaluation Data Collection

We leverage publicly available and high-quality human-annotated data from RoboArena (Atreya et al., 2025), a distributed robot policy evaluation platform. RoboArena hosts large-scale, real-world A/B evaluations in which human evaluators compare pairs of robot learning policies performing the same task under matched initial conditions. Each sample of our data corresponds to an A/B comparison and includes: (i) input including a textual task instruction (e.g., “put the banana in the orange bowl”) and a pair of 40-second rollout videos from two robot learning policies (anonymized as policy A and policy B) on a Franka Panda robot arm (Franka Robotics, 2026). Note that an evaluation session can contain multiple pairs of videos captured using different cameras; and (ii) output including a pairwise preference indicating the better policy (A, B, or TIE), a free-form textual justification for the preference, and two task progress scores for two policies. In total, we collect 4,465 high-quality human evaluation data samples, spanning across a wide range of robot evaluation scenarios and camera viewpoints, as exemplified in Figure 2. We use 90% of the data for training, 5% for validation, and reserve the remaining 5% for testing. Refer to Figure 3 for a sample in our dataset.

3.2 Training Details

Given the curated high-quality data, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) to perform parameter-efficient fine-tuning for widely-used VLMs in the Qwen-VL family, including Qwen2.5-VL (Bai et al., 2025b) and Qwen3-VL (Bai et al., 2025a). In particular, we fine-tune Qwen2.5-VL-3B-Instruct, Qwen3-VL-2B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct. Robot policy rollout videos are sampled at 1 FPS, resulting in 40 frames per video. During our training, we only update the LoRA adapter and the visual projector, while freezing

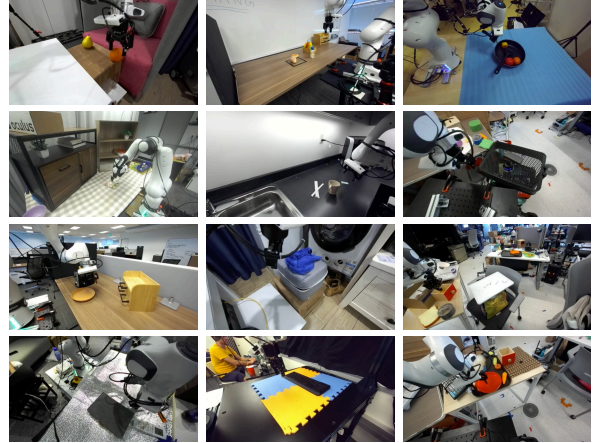


Figure 2: **Data Diversity.** RoboArena covers a wide range of tasks, scene settings, and camera viewpoints.

the LLM backbones and vision encoders. We set LoRA rank to 64 and the α scaling hyper-parameter to 128. All models are trained for 1 epoch with the learning rate of $1e-5$.

4 Experiments

4.1 Baselines and Setting

We benchmark the base and fine-tuned variants of Qwen2.5-VL-3B-Instruct, Qwen3-VL-2B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct. Temperature is set to 0 for all models to guarantee reproducibility. We evaluate models on 224 testing samples. Human preferences, progress scores, and textual explanations given in RoboArena data are treated as the ground truth. We prompt the VLMs to produce outputs in a predefined structured JSON format, enabling straightforward computation of evaluation metrics.

4.2 Metrics

We evaluate VLM judges using four metrics. First, **Preference Accuracy (PA)** measures the fraction of samples where the model’s pairwise preference (A, B, or TIE) matches the human annotator’s judgment, serving as a direct measure of judgment correctness. Second, we employ **Cohen’s kappa score (Kappa)** (McHugh, 2012), which is an accuracy-like metric adjusting for chance agreement given the class distributions. Third, we assess the quality of free-form explanations using **BERTScore** (rescaled with baseline) (Zhang et al., 2020), which computes token-level semantic similarity between predicted and reference explanations. Finally, we evaluate progress score estimation via **Mean Absolute Error (MAE)** on the $0 \rightarrow 100$ in-

```

{
  "id": "0498f9e6-a0ad-47aa-be33-d8d8c4958c6c_right",
  "video": [
    "0498f9e6-a0ad-47aa-be33-d8d8c4958c6c/A_paligemma_fast_droid/paligemma_fast_droid_2025-12-03T08:29:18.274515_video_right.mp4",
    "0498f9e6-a0ad-47aa-be33-d8d8c4958c6c/B_paligemma_fast_specialist_droid/paligemma_fast_specialist_droid_2025-12-03T08:34:03.351125_video_right.mp4"
  ],
  "conversations": [
    {
      "from": "human",
      "value": "<video>\n<video>\nYou are given two robot policy rollout videos (Video A and Video B) performing the same task: \"put the shoe on to floor\".\nEvaluate which policy performed better. Respond with a JSON object following this template exactly:\n{\n  \"preference\": \"<A | B | TIE>\",\n  \"explanation\": \"<free-form explanation of your preference>\",\n  \"progress_score_A\": <integer 0-100>,\n  \"progress_score_B\": <integer 0-100>\n}\nWhere:\n- \"preference\" is your preferred policy (\"A\", \"B\", or \"TIE\").\n- \"explanation\" is a free-form explanation of your preference.\n- \"progress_score_A\" and \"progress_score_B\" are integers from 0 to 100 proportional to the maximum progress each policy achieved (0 = no progress, 100 = task completed).\"
    },
    {
      "from": "gpt",
      "value": "{\n  \"preference\": \"TIE\",\n  \"explanation\": \"Policy A first moved the gripper toward the curtain, then moved back and forth for a while, grasped the curtain. Policy B first moved the gripper close to the controller, attempted to grasp it, then approached the curtain, attempted to grasp it. My instruction is to put the shoe on to floor, neither policy was able to complete the task. When comparing the performance of the two, both policies only moved the gripper, but did not find the shoe. Therefore, I choose tie.\",\n  \"progress_score_A\": 10,\n  \"progress_score_B\": 10\n}"
    }
  ]
}

```

Figure 3: **Data Example.** In total, we curate 4,465 high-quality human evaluation data samples from RoboArena. Each sample consists of two rollout videos of two policies, a task description, together with the human preference, a textual explanation, and two progress scores, one for each policy.

teger scores assigned to each rollout, where MAE is averaged across both rollouts and all samples. More concretely, given N evaluation samples, each with a predicted score pair $(\hat{s}_A^i, \hat{s}_B^i)$ and ground-truth pair (s_A^i, s_B^i) , the MAE is defined as:

$$\text{MAE} = \frac{1}{2N} \sum_{i=1}^N (|\hat{s}_A^i - s_A^i| + |\hat{s}_B^i - s_B^i|). \quad (1)$$

MAE measures how accurately VLM judges quantify task completion relative to human judgment. Together, these four evaluation metrics capture complementary aspects of evaluation quality: discrete preference correctness, explanation coherence, and progress estimation.

4.3 Results

Table 1 presents the performance of base and fine-tuned Qwen-VL models across all evaluation metrics. Among the base models, Qwen3-VL-8B-Instruct achieves the strongest performance, with a preference accuracy of 0.4598, a Kappa of 0.1274, a BERTScore of 0.2318, and an MAE of 35.73, suggesting that larger model capacity provides a modest advantage in judgment quality even without fine-tuning on human data. Fine-tuning with LoRA consistently and substantially improves per-

formance across all models and metrics, confirming that adapting VLMs to human judgment data is key for reliable policy evaluation. Notably, Qwen3-VL-8B-Instruct-LoRA achieves the best overall performance, attaining a PA of 0.6558, a Kappa of 0.3995, and a BERTScore of 0.3254, while reducing MAE to 22.04 – the lowest across all models. These results highlight that fine-tuning on high-quality human annotations is critical for driving evaluation quality. However, the moderate absolute scores suggest that despite their promise as automated judges, current VLMs still fall considerably short of human-level judgment, likely due to the inherent difficulty of interpreting nuanced robot behavior in diverse environments and reasoning across multiple video observations simultaneously. A qualitative comparison between a human response, one generated by the base Qwen3-VL-8B, and another produced by its version fine-tuned on the human data is presented in Figure 4. We can observe that while the base model’s response is largely incoherent and unreliable, the one generated by the fine-tuned model is much better and aligns more closely with the human evaluator’s response. For the remainder of this paper, unless otherwise stated, we refer to the fine-tuned Qwen3-VL-8B-Instruct-LoRA model as our VLM judge.

Model	PA $\uparrow$	Kappa $\uparrow$	BERTScore $\uparrow$	MAE $\downarrow$
Qwen2.5-VL-3B-Instruct	0.1818	-0.0102	0.2277	56.08
Qwen3-VL-2B-Instruct	0.4286	0.0023	0.2263	48.12
Qwen3-VL-4B-Instruct	0.3361	0.0170	0.2314	38.29
Qwen3-VL-8B-Instruct	0.4598	0.1274	0.2318	35.73
Qwen2.5-VL-3B-Instruct-LoRA	0.3020	0.0102	0.2499	34.22
Qwen3-VL-2B-Instruct-LoRA	0.4574	0.0452	0.2957	32.46
Qwen3-VL-4B-Instruct-LoRA	<u>0.5567</u>	<u>0.2490</u>	<u>0.3060</u>	<u>25.05</u>
Qwen3-VL-8B-Instruct-LoRA	0.6558	0.3995	0.3254	22.04

Table 1: **Results of Base and Fine-Tuned Qwen-VL Models.** For each metric, the best performance is **bolded**, and the second-best performance is underlined.

5 Ablation Study

5.1 Multi-View Rollout Videos

Robot learning setups usually come with multiple cameras to provide comprehensive visual observations for the robot policy. For example, a static side-view camera captures scene-level visual cues, while a dynamic wrist-view camera (attached to the robot’s wrist) provides detailed, fine-grained images of subtle robot-object interactions. Similarly, in RoboArena, rollout videos from the robot’s wrist camera are included alongside those from the side-view camera, which we use in our main experiments. In this section, we investigate whether incorporating wrist-view videos can improve the evaluation performance of the VLM judge. The trade-off is that while wrist-view videos supply additional visual context for more informed judgments, they also make the problem considerably more challenging, requiring the VLM to process four videos (two per policy) simultaneously in a single evaluation session. To study this, we fine-tune a Qwen3-VL-8B-Instruct model on multi-view video observations and also include wrist-view videos at test time. The results in Table 2 show that adding wrist-view rollout videos does not improve evaluation performance and, in fact, slightly worsens all scores. We hypothesize that this is because the manipulation tasks evaluated in RoboArena are not particularly contact-rich, such that side-view observations alone provide sufficiently informative visual cues, while reasoning across four videos imposes an additional burden on the VLM model.

5.2 Cross-Embodiment Generalization

Although our VLM judge is trained on RoboArena’s evaluation data collected exclusively with the Franka Panda robot, we expect its learned understanding of robot behaviors and

Setting	PA $\uparrow$	Kappa $\uparrow$	BERTScore $\uparrow$	MAE $\downarrow$
Multi-view	0.6309	0.3912	0.2904	25.71
Ours	0.6558	0.3995	0.3254	22.04

Table 2: **Comparison with Multi-View Setting.** In the multi-view setting, the videos captured from the wrist camera are also used for evaluation.

physical interactions to transfer to other robot embodiments. To verify this, we test the judgment ability of our VLM judge on two manipulation tasks conducted with the SO100 robot arm (Cadene et al., 2024), using the same prompting setup as our main experiments. The first task is to place a cup on a plate, and the second is to fold a towel. We use two policies trained on demonstration data of these tasks. The results are shown in Figure 5. In both cases, the VLM judge correctly identifies the better policy – policy B in the first task and policy A in the second. We observe that the textual explanations and task progress scores provided for the folding task are reasonable and consistent with our own judgment. However, for policy B in the first task, the VLM judge incorrectly states that the robot successfully placed the cup on the plate, when in fact the robot dropped the cup at the end of the episode. Overall, these results suggest that our VLM judge can produce reasonable evaluations for robot embodiments unseen during fine-tuning, indicating that the learned knowledge generalizes well and that the judge can potentially be applied across a wide range of robot platforms.

6 Data Selection for Imitation Learning

In this section, we evaluate the usefulness of our VLM judge for an application beyond policy evaluation, which is demonstration data curation for training robotic imitation learning. The expectation is that with the knowledge learned in the evalua-

Task: "Pick up the knife"

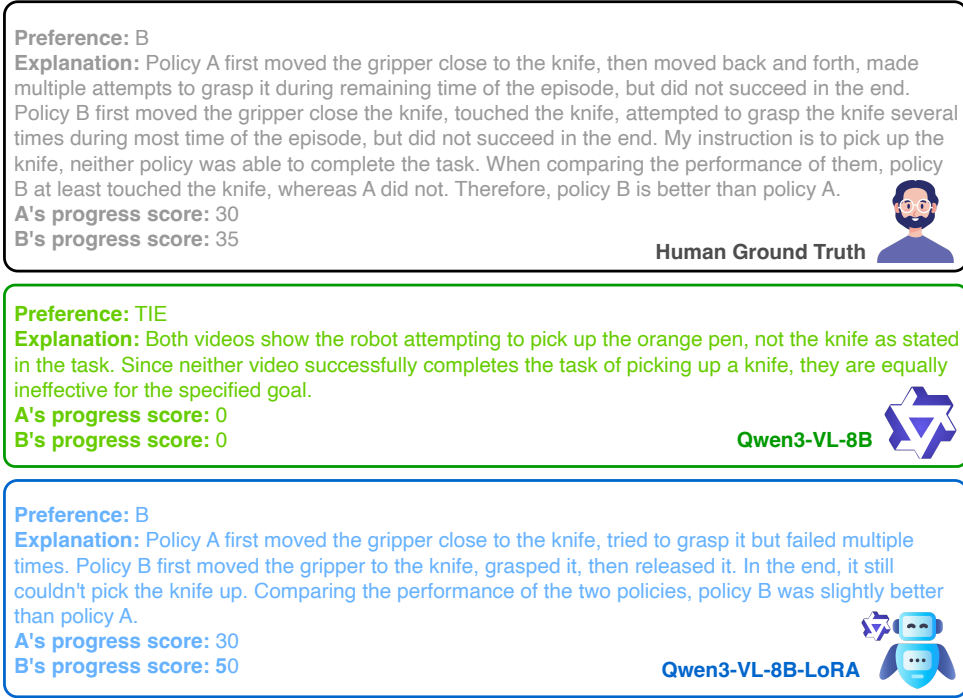


Figure 4: **Qualitative Result.** Sample responses of a human evaluator (ground-truth), the base Qwen3-VL-8B model, and our fine-tuned VLM judge. Compared to the base model, our judge produces the response that aligns substantially more closely with the human judgment.

tion fine-tuning process, the VLM judge should possess a strong capability in assessing the quality of robotic demonstrations by watching the corresponding video observations.

6.1 Setup

Mixed-quality data construction. We investigate the utility of our VLM judge for selecting high-quality data from a mixed-quality demonstration dataset, and subsequently use the curated data to train a robot learning policy on the LIBERO-Object robotic manipulation benchmark (Liu et al., 2023). The benchmark consists of 10 manipulation tasks, each with 50 expert demonstrations provided. To construct the mixed-quality dataset, we combine the 50 expert demonstrations with 50 suboptimal and 50 failure demonstrations generated by an OpenVLA model (Kim et al., 2025) trained on the LIBERO-Object expert data. Note that this model achieves a success rate of 88.4% on the benchmark. We define suboptimal demonstrations as those in which the robot completes the task but takes at least 30 more time steps than the per-task average of the expert demonstrations. In total, the mixed dataset comprises 1,500 demonstrations, 150 per task.

Data selection. From the mixed-quality dataset,

we employ the base Qwen3-VL-8B model and our fine-tuned VLM judge to select 50 demonstrations per task for training a new OpenVLA policy. Specifically, for each demonstration, we prompt the VLM to assign a task progress score from 0 to 100 given the rollout video and the corresponding text description of the task. For every task, the 50 highest-scoring demonstrations are then selected for policy training. All training and evaluation settings follow those already established in LIBERO (Liu et al., 2023).

6.2 Results

Can the VLM judge effectively select high-quality demonstration data? To answer this question, we analyze the composition of data selected by the base model and our VLM judge in terms of the ratio of expert, suboptimal, and failure demonstrations. Figure 6 reports the average counts out of 50 selected demonstrations for each demonstration type, averaged across the 10 LIBERO-Object tasks. The results show that our VLM judge selects a higher proportion of expert demonstrations and a lower proportion of failure demonstrations compared to the base model. Specifically, more than half of the demonstrations curated by our judge are

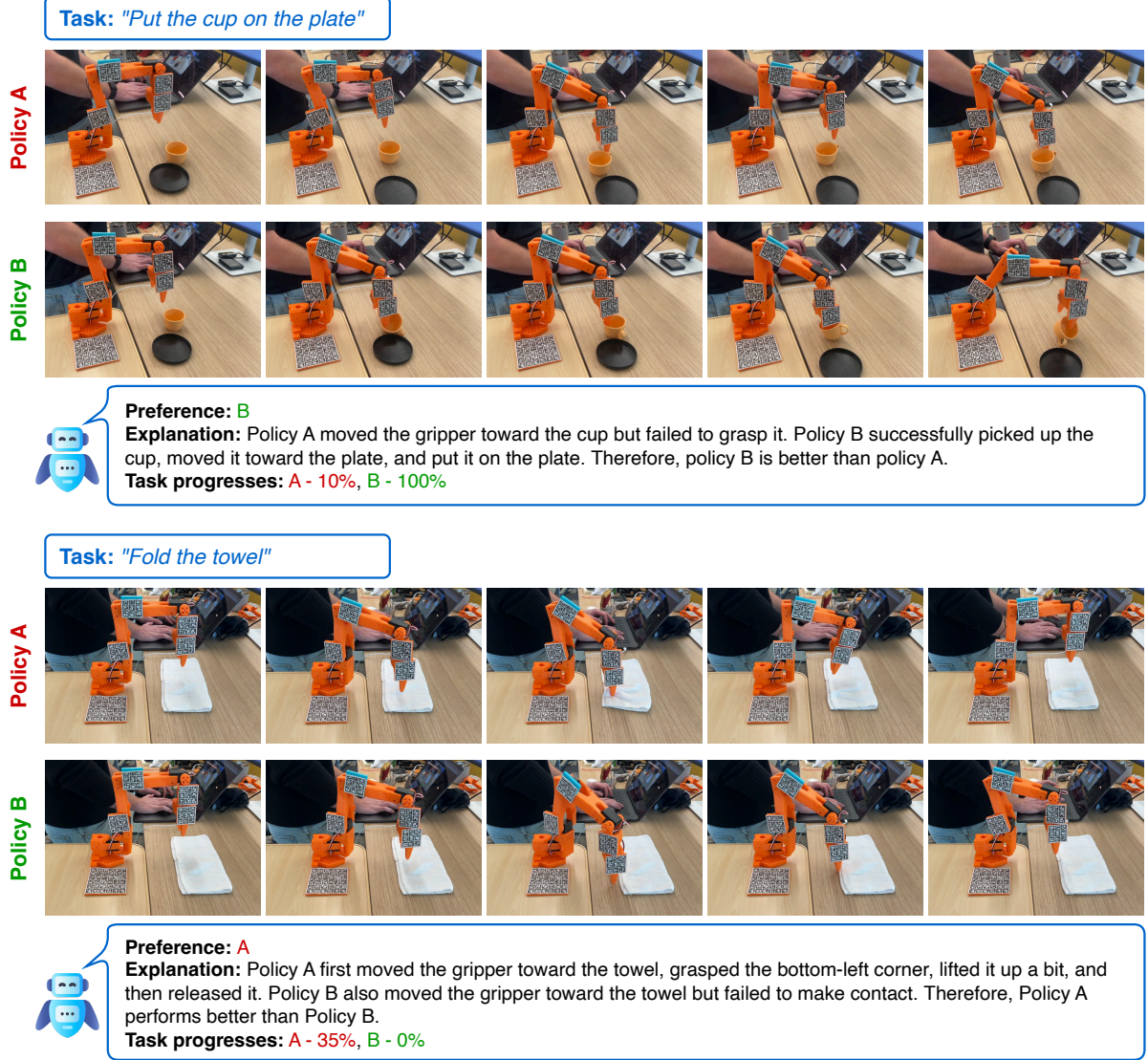


Figure 5: **Cross-Embodiment Evaluation Result.** Rollout videos are captured by running two robot policies for two manipulation tasks on an SO100 robot arm.

Ours	28.74	14.05	7.21
Base model	21.30	14.94	13.76

■ Expert demos
 ■ Suboptimal demos
 ■ Failure demos

Figure 6: **Data Selection Result.** We report ratios averaged across 10 tasks of the LIBERO-Object benchmark.

high-quality expert demonstrations (28.74 out of 50 on average), while only 7.21 out of 50 are failure demonstrations. This suggests that fine-tuning on the A/B policy evaluation task also benefits demonstration data selection, which is intuitive given that both tasks require the model to comprehend physical robot behaviors and interaction outcomes. Qual-

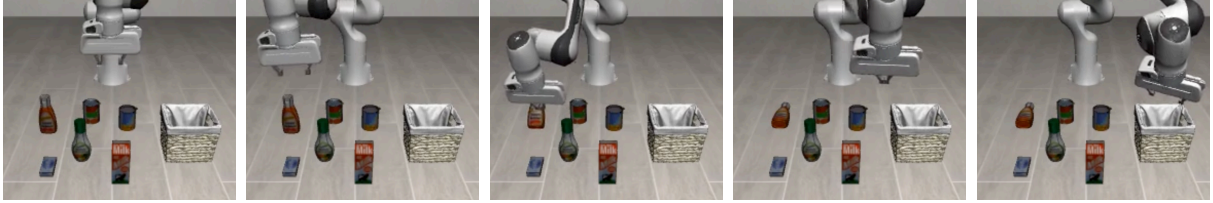
itative examples of demonstrations and their corresponding scores assigned by our VLM judge are shown in Figure 7.

How good is the policy trained on data curated by our VLM judge? Given that our VLM judge selects higher-quality training data, we expect the OpenVLA policy trained on such data to achieve stronger performance. Table 3 validates this expectation. The results show that the policy trained on data curated by our VLM judge substantially outperforms the one trained on data selected by the base model. While it does not reach the performance of the policy trained exclusively on expert demonstrations, its success rate is considerably higher than that of the base model, which remains very low, demonstrating the practical value of our VLM judge for downstream imitation learning.

Task: "Pick up the *orange juice* and put it in the basket"



Task: "Pick up the *ketchup* and put it in the basket"



Task: "Pick up the *tomato sauce* and put it in the basket"



Figure 7: **Data Selection Qualitative Result.** In the first video, the robot grasps the wrong object, which is the cream cheese. In the second, the robot moves toward the ketchup but knocks it off before moving to the basket. The robot successfully completes the task in the third video.

Data from	Success Rate $\uparrow$
Base model	$61.9 \pm 0.8\%$
Our VLM judge	$80.5 \pm 0.9\%$
Expert demos	$88.4 \pm 0.8\%$

Table 3: **Imitation Learning Result.** The final success rate is averaged across 10 tasks of LIBERO-Object.

7 Discussion

Despite encouraging results, there remains substantial room for improvement in this research direction. First, we only evaluate our VLM judge on single-arm robots, specifically the Franka Panda and the SO100 arm. Extending the evaluation to bimanual robotic systems or humanoid robots would provide a more comprehensive picture of the judge’s generalizability across diverse robot embodiments. Second, throughout our experiments, we observe that the VLM judge primarily focuses on semantic and task progress properties of rollout videos, while exhibiting limited sensitivity to other behaviorally important attributes such as motion smoothness, reactivity, and failure recovery – qualities that are crucial for robust robot learning but remain

difficult for current VLMs to reliably assess. Finally, we expect that augmenting the fine-tuning data with more general robotic visual question answering data (Zhao et al., 2025) could strengthen the model’s grasp of spatial and physical concepts, thereby further improving its assessment capability. We leave these directions for future work.

8 Conclusion

We investigate training vision-language models for the task of robot policy A/B evaluation and show that fine-tuning on high-quality human evaluation data significantly improves the models’ evaluation ability. Additionally, we show that the trained VLM judge can generalize to a different robot embodiment that is not present in its fine-tuning data. Finally, we demonstrate that the VLM judge can also be repurposed to select training demonstration data for robotic imitation learning, yielding a policy that substantially outperforms one trained on data selected by the base VLM. Together, these findings suggest that modern vision-language models can serve as reliable automated judges in the development of generalist robot learning models.

Acknowledgment

We sincerely thank the RoboArena team for developing the evaluation framework and releasing the A/B evaluation data that made this research possible. We also thank Cameron Smith for providing the SO100 robot rollout videos used in our cross-embodiment generalization experiment.

References

- Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Cyrus Neary, Edward S Hu, Kanav Arora, Kirsty Ellis, and 1 others. 2025. Roboarena: Distributed real-world evaluation of generalist robot policies. In *Conference on Robot Learning*, pages 336–364. PMLR.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-v1 technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-v1 technical report. *arXiv preprint arXiv:2502.13923*.
- Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, and 1 others. 2026. A careful examination of large behavior models for multitask dexterous manipulation. *Science Robotics*, 11(113):eaea6201.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, and 1 others. 2023. Rt-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems XIX*.
- Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Jess Moss, and Thomas Wolf. 2024. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. <https://github.com/huggingface/lerobot>.
- Shirui Chen, Cole Harrison, Ying-Chun Lee, Angela Jin Yang, Zhongzheng Ren, Lillian J Ratliff, Jiafei Duan, Dieter Fox, and Ranjay Krishna. 2026. Topreward: Token probabilities as hidden zero-shot rewards for robotics. *arXiv preprint arXiv:2602.19313*.
- Franka Robotics. 2026. Franka robotics. <https://franka.de/>.
- Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyang Jiang, Aaran Arulraj, and 1 others. 2024. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Y Jangir, Y Zhang, K Yamazaki, and 1 others. 2026. Robotarena ∞ : Scalable robot benchmarking via real-to-sim translation. In *International Conference on Learning Representations*.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, and 1 others. 2025. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR.
- Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11286–11315.
- Tony Lee, Andrew Wagenmaker, Karl Pertsch, Percy Liang, Sergey Levine, and Chelsea Finn. 2026. Roboreward: General-purpose vision-language reward models for robotics. *arXiv preprint arXiv:2601.00675*.
- Anthony Liang, Yigit Korkmaz, Jiahui Zhang, Minyoung Hwang, Abrar Anwar, Sidhant Kaushik, Aditya Shah, Alex S Huang, Luke Zettlemoyer, Dieter Fox, and 1 others. 2026. Robometer: Scaling general-purpose robotic reward models via trajectory comparisons. *Robotics: Science and Systems*.
- Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.
- Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791.
- Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- Chenlin Ming, Zitong Wang, Boxuan Zhang, Zhanxiang Cao, Xiaoming Duan, and Jianping He. 2025. Stochastic trajectory optimization for robotic skill acquisition from a suboptimal demonstration. *IEEE Robotics and Automation Letters*.

- Ranjan Sapkota, Yang Cao, Konstantinos I Roulis, and Manoj Karkee. 2025. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*.
- Pavel Senin. 2008. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40.
- Ran Tian, Yilin Wu, and A Bacjsy. 2026. Position: Good embodied reward models need bad behavior data.
- Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. 2024. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22227–22238.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Enyu Zhao, Vedant Raval, Hejia Zhang, Jiageng Mao, Zeyu Shangguan, Stefanos Nikolaidis, Yue Wang, and Daniel Seita. 2025. Manipbench: Benchmarking vision-language models for low-level robot manipulation. In *Conference on Robot Learning*, pages 3413–3462. PMLR.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. 2020. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*.