OLAF Vision: Interactive Robot Learning with Visual States and Verbal Feedback

Pei Chen, Lee
B10902032

Yu Chen, Lin
B10902033

Bing Yi, Fan
B10902117

Abstract: This paper introduces OLAF Vision, a novel approach to interactive robot learning that combines Vision-Language Models (VLMs) and fine-tuned diffusion models with human verbal feedback. Unlike traditional methods that rely on pre-programmed behaviors or extensive expert demonstrations, OLAF Vision enables robots to learn from real-time user corrections while operating in unstructured environments. The system leverages visual state representations and verbal feedback to dynamically update neural policies, aiming to reduce dependency on expert-crafted prompts and enhance generalization across tasks. However, our experimental results on MetaWorld tasks reveal that OLAF Vision currently underperforms compared to its predecessor OLAF, highlighting challenges in VLM-based spatial reasoning and the critical importance of feedback quality. These findings provide valuable insights for future development of more robust and adaptable robotic learning systems that can effectively integrate human feedback.

1 Introduction

The integration of robots into daily life has the potential to revolutionize how we approach routine and complex tasks, ranging from household chores to industrial processes. However, achieving this vision requires robots that can adapt to unstructured environments and learn from user feedback in real-time. Current robotic systems often depend heavily on pre-programmed behaviors or reinforcement learning, which limits their adaptability and scalability in dynamic settings.

Imitation learning, a subset of machine learning, has emerged as a promising approach to train robots by mimicking expert demonstrations. While effective, this paradigm faces significant challenges, including the reliance on large datasets of high-quality expert demonstrations, which are costly and time-consuming to collect. Furthermore, the static nature of these datasets restricts the robot's ability to adapt to unforeseen scenarios in real-world applications.

In this work, we propose **OLAF Vision**, a novel system that combines Vision-Language Models (VLMs), fine-tuned diffusion models, and human verbal feedback to address these challenges. OLAF Vision allows users to interactively teach robots by providing corrective verbal feedback when the robot exhibits suboptimal behavior. For example, a user can say, "You should move up to reach the button," enabling the robot to refine its policies and improve future performance.

Unlike traditional approaches that rely on reinforcement learning or static datasets, OLAF Vision leverages visual state representations and verbal corrections to dynamically update the robot's neural policies. This interactive learning paradigm enhances the robot's ability to generalize across tasks and environments, reducing the dependency on expert-crafted prompts or demonstrations. Additionally, the incorporation of VLMs enables the system to interpret and act upon multimodal data, bridging the gap between visual and linguistic inputs.

Our contributions are as follows:

1. We introduce a framework that generates corrective demonstration data using VLMs and diffusion models, enabling robots to learn interactively from human feedback.

Robot Perception and Learning (RPL 2024 Fall), National Taiwan University.

2. We explore the capability of VLMs to understand and process robotic state images, identifying key limitations and potential areas for improvement.
3. We demonstrate the generalizability of our approach across different tasks without relying on extensive prompt engineering or domain-specific expertise.
4. We provide an open-source implementation to facilitate further research and development in interactive robot learning.

By addressing the limitations of current imitation learning systems and incorporating human feedback, OLAF Vision aims to pave the way for more adaptable and user-friendly robotic systems. This paper evaluates the system's performance, identifies key challenges, and outlines future directions for research in this emerging field.

2 Related Work

2.1 Imitation Learning

Imitation Learning is a prominent paradigm in machine learning, particularly within the realm of reinforcement learning. It enables agents to learn tasks by mimicking expert behavior rather than relying solely on trial-and-error methods typical of traditional reinforcement learning. This approach, often referred to as Learning from Demonstration (LfD) [1], involves collecting data from expert demonstrations, which the agent uses to learn a policy that maps observations to actions. The primary advantage of imitation learning is its ability to bypass the challenges associated with defining reward functions, making it especially useful in complex environments where such definitions are difficult or impractical. Behavior Cloning (BC) is one of the main methods to approach an imitation learning problem.

2.2 Human Feedback

Human feedback plays a crucial role in enhancing the performance and adaptability of machine learning systems and also in reinforcement learning, also known as Reinforcement learning from Human Feedback (RLHF)[2]. While RLHF is not the method we use, we still gather human feedback as one of our inputs of our system. Shi et al.[3] investigates how real-time language corrections from users can be utilized to improve robotic performance. Torne et al.[4] focuses on goal-conditioned exploration strategies that leverage human feedback to shape the exploration process, guiding AI agents toward desired outcomes.

OLAF (**O**peration relabeled learning with **LA**nguage **F**eedback)[5] not only uses human feedback to improve policy but also represents a significant development within imitation learning frameworks. Specifically, it enables agents to learn from a single demonstration (which includes previous states and human feedback) rather than requiring extensive datasets. This method emphasizes efficiency and adaptability, allowing agents to quickly assimilate new tasks with minimal input. OLAF employs advanced **neural architectures** that facilitate rapid learning and generalization, making it particularly suitable for dynamic environments where quick adaptation is crucial. In this paper, we build upon the OLAF architecture, introducing a crucial modification—data synthesis. By incorporating a visual-based approach, we offer a novel perspective to address the problem effectively.

2.3 Vision-Language Models / VLM for Labeling

Vision-Language Models (VLMs) have emerged as powerful tools in the realm of machine learning, particularly for tasks involving multimodal data such as images and text. GPT-4o [6], LLaVA [7] are two notable advancements in the field of VLMs, each designed to enhance multimodal interactions by integrating various forms of input and output.

Recent work shows that VLMs effectively detect and correct noisy labels through their understanding of visual and textual relationships [8]. The distillation of large-scale VLMs into task-specific

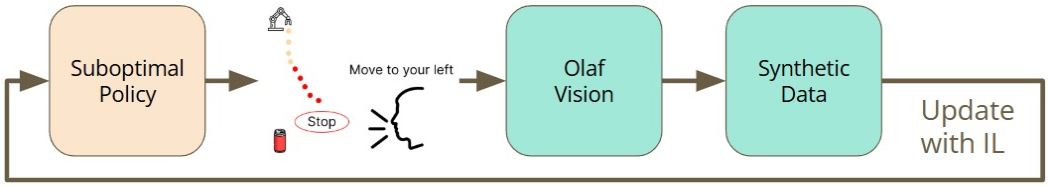

Figure 1: Main problem formulation method

relabeling functions facilitates high-quality labeling and knowledge transfer [9]. In robotics, VLMs enable skill acquisition by combining visual and textual instructions, allowing robots to adapt to complex tasks and environments [10]. In our work, we utilize VLMs to relabel a new action based on given trajectory image and human feedback, expecting that VLMs would determine the correct label for imitation learning.

2.4 Text-to-image Diffusion Models

Recent advancements in text-to-image diffusion models have significantly enhanced their ability to generate high-quality images from textual prompts. In addition to improving control over text-to-image generation through orthogonal finetuning techniques [11], integrates policy optimization with KL regularization in text-to-image diffusion models [12], the **InstructPix2Pix** model [13] introduces a method for editing images based on natural language instructions. This model is also used in our system. Unlike traditional models that require specific labels or example images, InstructPix2Pix interprets user-provided text to perform edits directly on the input image. The model leverages a combination of a language model (GPT-3) [14] and a text-to-image model (Stable Diffusion) [15] to generate a diverse dataset of image editing examples, which it uses for training. The result is a conditional diffusion model capable of executing complex edits in seconds without needing fine-tuning for each example. This model best fits our system as it responds fast and provides outstanding image outputs.

3 Problem Formulation

This section addresses the challenge of using human feedback to train robots to recover from suboptimal behavior, as shown in Figure 1. When a robot trained in a controlled factory environment is deployed to a user's environment, it often encounters unfamiliar conditions, resulting in poor performance. To address this, we propose leveraging user feedback during operation to refine the suboptimal robot's behavior.

The user observes the robot's trajectory, defined as a sequence of observation-action pairs, and intervenes upon detecting errors by halting the robot and providing corrective feedback. While the feedback does not explicitly identify which steps are incorrect, it indicates a general region of the trajectory where errors are likely to have occurred. To address these errors, we assume that incorrect actions occurred within a predefined window before the intervention, based on findings such as those in [5]. Synthetic corrective actions are then generated for these steps to augment the dataset. Using this augmented dataset, we apply behavior cloning to update the robot's policy.

The primary challenge lies in effectively leveraging the human feedback and trajectory data to infer and generate the most accurate corrective actions. This step is crucial, as the precision of these corrective actions directly influences the robot's ability to adapt and improve its performance in the user's environment. By tackling this challenge, we aim to establish a framework that enhances robotic behavior in dynamic, real-world conditions.

3.1 Suboptimal Agents

The suboptimal robot model used in this study is trained on the dataset provided by [16], utilizing the behavior cloning (BC) methodology outlined by [1]. The training configuration is provided in the Appendix A.

3.2 Human Feedback Generation

To generate human feedback, we created a web interface, as described in the Appendix B. Using this interface, we recorded the robot's full trajectory, including observation states, actions, 30 frames-per-second rendered videos, and textual feedback from a human participant. The participant was tasked with observing the robot's behavior, identifying when it made errors, and providing suggestions on how the robot should behave instead. To ensure meaningful feedback, the participant had prior knowledge of the robot's task and was instructed to give corrections relative to the robot's position. Based on the approach outlined in [5], we defined the 30 states leading up to the human intervention as erroneous and focused on correcting these states.

3.3 Behavior Cloning Updates

After generating the synthetic data and augmenting the dataset, we update the base suboptimal models using the same behavior cloning (BC) method. This process incorporates a new training configuration, as detailed in Appendix C.

3.4 Previous Solution - Olaf

[5] proposes Olaf, an approach that leverages human feedback and observational data to generate corrective actions through large language models (LLMs) using expert-crafted prompts. While the method presents an interesting direction, several limitations warrant consideration:

1. **State-Space Representation Challenges**: LLMs, primarily trained on natural language, may not optimally process or reason about numerical state representations that are fundamental to robotics.
2. **Limited Model Transparency**: Despite generating explanatory text, the decision-making process within the LLM remains opaque due to the black-box nature of these models.
3. **System-Specific Dependencies**: The approach's reliance on extensive prompt engineering for performance optimization suggests limited transferability across different robotic systems. This system-specific customization requirement may impede scalable deployment and broader applicability in diverse robotics applications.

4 Method

To address the challenges outlined above, we propose a novel system, Olaf-Vision, which leverages visual state image representations to generate corrective action data from human feedback and trajectory data.

4.1 Our system - Olaf Vision

The Olaf-Vision system begins by identifying the initial observation state image associated with an erroneous action. It then predicts the next observation state images for each of eight predefined actions: "Move right," "Move left," "Move forward," "Move backward," "Move up," "Move down," "Close gripper," and "Open gripper." These predicted images, referred to as candidate next states, represent potential outcomes of each action.

Next, the system uses the original observation state image, the candidate next state images, and the user-provided feedback to prompt a visual-language model (VLM). In our implementation, we use

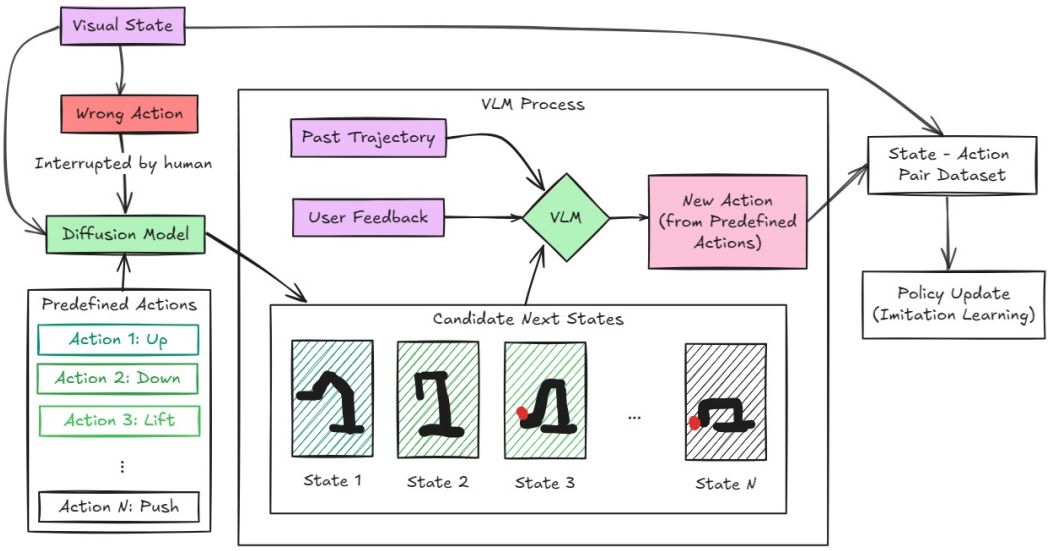

Figure 2: System Design of Olaf-Vision

GPT-4o-mini ([6]) as the VLM to generate the corrective action. The complete system design is illustrated in Figure 2.

4.2 Diffusion Model Fine-tuning Details

To fine-tune our diffusion model, we collected 5,000 data samples, each comprising an observation image, an action, and the resulting next observation image. The data was gathered from 50 suboptimal agents, each trained on one of the 50 tasks provided in the MetaWorld environment ([17]).

Each agent had a 60% probability of performing its predicted action and a 40% probability of collecting a random data sample. When collecting a sample, the agent first saved its current observation image. It then performed a randomly selected action from the eight predefined actions for several steps and saved the resulting next observation image. This process was repeated until each agent collected 100 data samples.

We trained our diffusion model using the collected dataset, starting with the pretrained stable-diffusion-v1-5 model from [18]. Training followed the methodology of InstructPix2Pix, as described by [13]. The specific training configuration is detailed in Appendix D.

4.3 Benefits

We believe that this approach will bring the following benefits.

1. **Enhanced Representations with Visual States**: By incorporating visual states, the system captures richer 3D spatial information, which is essential for robotics applications.
2. **Improved Generalization and Simplicity in Prompting**: The VLM requires significantly less expert-driven prompt engineering due to its ability to understand information directly from images. Compared to Olaf's original method, Olaf-Vision reduces the number of words used in prompts by two-thirds (300 words vs. 100 words).
3. **Clearer State Interpretation for Language Models**: Providing actual next-state images makes the system's reasoning process more transparent to the language model. Additionally, when past trajectories (a sequence of previous states) are included, the language model can more accurately determine the correct state. Unlike methods that rely solely on action descriptions, Olaf-Vision enables humans to verify the VLM's choices by presenting predicted next-state images, improving trust and interpretability.

5 Experiments

To validate our hypotheses, we designed two experiments addressing the following research questions:

1. Does Olaf-Vision outperform Olaf?
2. Does increasing the quantity of human feedback improve performance?

5.1 Experiment Setup

We selected two tasks from the MetaWorld environment: "button-press" and "button-press-topdown." For each task, we collected over 100 instances of human feedback and trajectory data using the method outlined in Section 3.2. Subsequently, subsets of 30, 60, and 100 feedback instances were randomly sampled for training. Olaf (described in Section 3.4) and Olaf-Vision (described in Section 4.1) were employed to generate synthetic, relabeled corrective data for each dataset. The suboptimal models were then retrained using the procedure detailed in Section 3.3.

For evaluation, each agent was assessed every 20 epochs, with performance measured by the success rate on 20 randomly selected testing environments. As a baseline, the suboptimal model achieved success rates of 59% on the "button-press" task and 15% on the "button-press-topdown" task.

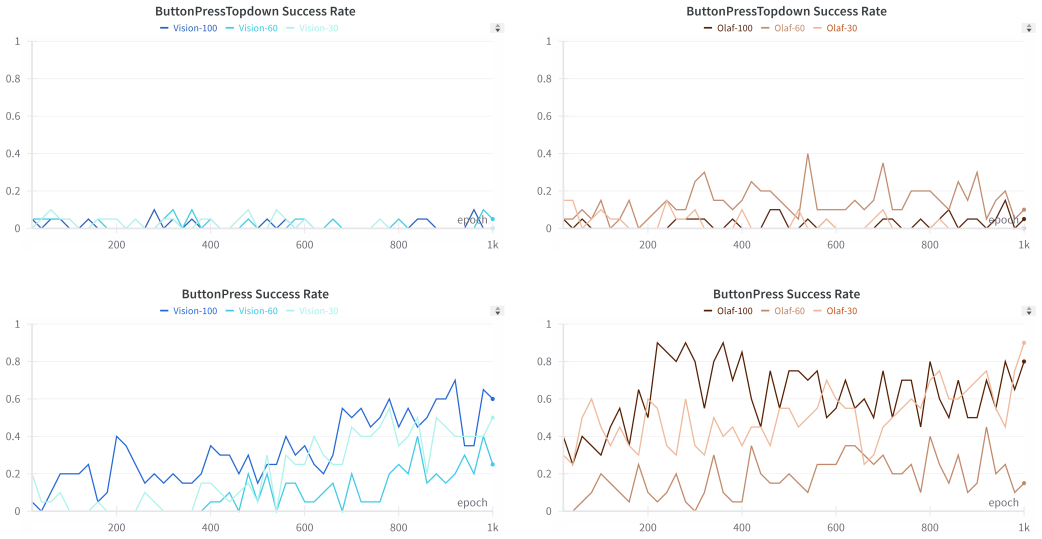

Figure 3: Comparison of success rates across tasks and methods. The baseline model achieved success rates of 59% for the "button-press" task and 15% for the "button-press-topdown" task. Additional detailed results are available in Appendix E.

5.2 Comparison between Olaf-Vision and Olaf

Figure 3 illustrates that Olaf-Vision underperforms compared to Olaf across both tasks. This difference is particularly evident when considering the success rates of the baseline models.

For the "button-press" task, Olaf-Vision achieves success rates comparable to the baseline (59%) only when trained on 100 human feedback samples. In contrast, Olaf consistently improves the baseline model by approximately 20%, even with only 30 or 100 feedback samples.

Similarly, for the "button-press-topdown" task, Olaf-Vision's performance remains similar to the baseline (15%), whereas Olaf enhances the baseline model by approximately 20%, particularly when trained on 60 feedback samples.

5.3 Comparison of Feedback Quantities

As depicted in Figure 3, increasing the quantity of human feedback does not always correlate with improved performance.

For the "button-press" task, models trained on 30 and 100 feedback samples exhibit similar performance, while the model trained on 60 samples performs worse.

Conversely, for the "button-press-topdown" task, the model trained with 60 feedback samples demonstrates the best performance when using Olaf, whereas models trained on 30 and 100 samples yield comparably poorer results.

6 Discussions

6.1 Why wasn't the performance as expected?

As shown in Section5, we observe that the results for OLAF Vision (employ VLM for decision making) do not perform as well as those for OLAF. One possible reason for this is that the VLM struggled to clearly interpret the provided images such as trajectory images and next-state images, which limited its ability to make accurate decisions. In addition, VLM fails to determine the correct action according to the current-state image and next-state images, despite being distinguishable by humans. This can be evident by the sample response shown in Appendix F. On the other hand, LLM could understand the difference in terms of Cartesian coordinate and therefore output a reliable decision. Another possible reason is that the experimental environment may not have been broad enough to examine the model and display more aspects of the capability of OLAF Vision in statistics. In the end, we believe that the use of a better model will address some of limitations on the current model and improved performance. Our current model, GPT-4o mini, struggles with interpreting images effectively. Its limitations in understanding image content result in an inability to accurately extract its meaningful information. Instead of GPT-4o mini, we will try models which experts in visual interpretation tasks.

6.2 Why didn't more feedback generate better results?

As shown in section 5.3, we observed that the success rate was not proportional to the number of feedback instances. The primary reason is that **the quality of human feedback may not have been sufficient**. Although we selected relatively fair feedback, it was challenging to ensure consistency, as two similar scenes could result in different yet reasonable feedback. To cope with this situation, we should aim to eliminate such inconsistencies and focus on generating diverse, high-quality human feedback. Another contributing factor is the **instability of imitation learning (IL) agents during training**. Training neural network agents often results in overfitting due to various reasons, for instance excessive training epochs, the absence of test data during continual training stage, or biased training data. We expect that more feedback will lead to improved results, while more structural modifications and careful hyperparameter tuning work should be taken into consideration in our future work to later validate our expectations.

6.3 Future Works

Future research should address the challenges and limitations identified in this study by exploring the following directions:

1. **Fine-tuning Vision-Language Models (VLMs) for 3D Spatial Understanding**: Enhancing the capacity of VLMs to comprehend 3D spaces could significantly improve decision-making performance. This may involve fine-tuning VLMs on larger and more diverse datasets or incorporating 3D data representations, such as point clouds or meshes, to enable more accurate spatial reasoning.

2. **Incorporating and Evaluating Diverse Human Feedback**: Human feedback is inherently variable, and its quality can significantly influence the model's performance. Future work should focus on developing methods to automatically assess and prioritize high-quality human feedback. This could include mechanisms to detect inconsistencies, filter noisy data, and integrate diverse feedback types to enhance the training process.
3. **Generalizing Olaf to a Broader Range of Tasks**: Improving the adaptability of Olaf across diverse tasks is critical for broader applicability. A potential direction is to develop techniques for generating task-specific, expert-level prompts automatically. Such methods could enhance the generalizability of the model while minimizing manual intervention in task-specific adaptations.

7 Conclusions

In this work, we presented OLAF-Vision, a novel system designed to improve robot learning through the integration of visual states and human feedback. Our approach leverages Vision-Language Models (VLMs) and fine-tuned diffusion models to address the limitations of traditional imitation learning frameworks and improve the adaptability of robots in dynamic environments.

Our findings reveal several key insights:

1. While OLAF-Vision introduces a new paradigm for robot learning, its current implementation underperforms compared to OLAF in terms of success rates. This can be attributed to limitations in the visual interpretation capabilities of the employed VLM.
2. The quality and consistency of human feedback are critical to model performance. We observed that an increased volume of feedback does not necessarily lead to better outcomes, highlighting the need for mechanisms to ensure high-quality feedback and reduce variability.
3. The experimental results underscore the potential of incorporating visual state representations for improved robot policy learning, though further refinement of the underlying models is required.

Looking forward, this research opens avenues for future exploration, including enhancing VLMs for better 3D spatial understanding, automating the assessment of human feedback quality, and generalizing the OLAF-Vision framework across diverse robotic tasks.

By providing an open-source implementation, we aim to encourage further research and development in this domain, fostering advancements in robot learning systems that are capable of leveraging human feedback to adapt and improve in real-world scenarios.

References

[1] S. Schaal. Learning from demonstration. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/68d13cf26c4b4f4f932e3eff990093ba-Paper.pdf.

[2] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

[3] L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, and C. Finn. Yell at your robot: Improving on-the-fly from language corrections, 2024. URL https://arxiv.org/abs/2403.12910.

[4] M. Torne, M. Balsells, Z. Wang, S. Desai, T. Chen, P. Agrawal, and A. Gupta. Breadcrumbs to the goal: Goal-conditioned exploration from human-in-the-loop feedback, 2023. URL https://arxiv.org/abs/2307.11049.

[5] H. Liu, A. Chen, Y. Zhu, A. Swaminathan, A. Kolobov, and C.-A. Cheng. Interactive robot learning from verbal correction, 2023. URL https://arxiv.org/abs/2310.17555.

[6] OpenAI. "hello gpt-4o.", 2024. URL https://openai.com/index/hello-gpt-4o/.

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

[8] C. Feng and I. Patras. Cleaning label noise with vision-language models, 2024. URL https://openreview.net/forum?id=1rgMkDWfYV.

[9] T. Sumers, K. Marino, A. Ahuja, R. Fergus, and I. Dasgupta. Distilling internet-scale vision-language models into embodied agents, 2023. URL https://arxiv.org/abs/2301.12507.

[10] T. Xiao, H. Chan, P. Sermanet, A. Wahid, A. Brohan, K. Hausman, S. Levine, and J. Tompson. Robotic skill acquisition via instruction augmentation with vision-language models, 2023. URL https://arxiv.org/abs/2211.11736.

[11] Z. Qiu, W. Liu, H. Feng, Y. Xue, Y. Feng, Z. Liu, D. Zhang, A. Weller, and B. Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning, 2024. URL https://arxiv.org/abs/2306.07280.

[12] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2305.16381.

[13] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[16] T. Schmied, M. Hofmarcher, F. Paischer, R. Pascanu, and S. Hochreiter. Learning to modulate pre-trained models in rl. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. *CoRR*, abs/1910.10897, 2019. URL http://arxiv.org/abs/1910.10897.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

A Suboptimal Agent Training Configuration

```
model = nn.Sequential(
    nn.Linear(input_dim, 1024),
    nn.ReLU(),
    nn.Linear(1024, 1024),
    nn.ReLU(),
    nn.Linear(1024, output_dim)
)

config = TrainingConfig(
    batch_size=64,
    training_size_per_epoch=0.05,
    epochs=200,
    lr=0.001,
    lr_step_size=50,
    lr_gamma=0.3,
    env_eval_freq=10,
    log_freq=10
)
```

Listing 1: Suboptimal Agent Training Configuration

B Human Feedback Generator Image

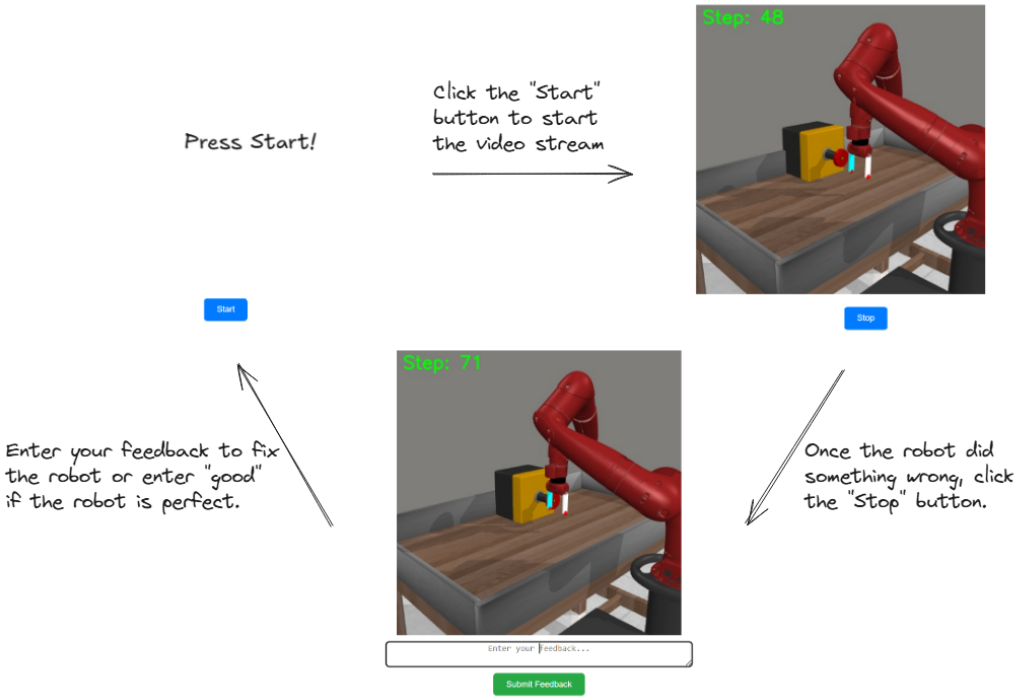

Figure 4: Human Feedback Generator

C Update Agent Training Configuration

```
config = UpdatingConfig(
    batch_size=64,
    training_size_per_epoch=1,
    epochs=1000,
    lr=0.0001,
    eval_count=10,
    env_eval_freq=20,
    eval_episodes=20,
    log_freq=10,
)
```

Listing 2: Update Agent Training Configuration

D Diffusion Training Configuration

```
--resolution=480
--train_batch_size=8
--gradient_checkpointing
--gradient_accumulation_steps=1
--max_train_steps=20000
--checkpointing_steps=2000
--learning_rate=5e-05
--max_grad_norm=1
--lr_warmup_steps=0
--conditioning_dropout_prob=0.05
--seed=42
```

Listing 3: Diffusion Training Configuration

E Wandb Training Results

See Figure 5 for training loss and Figure 6 for evaluation loss.

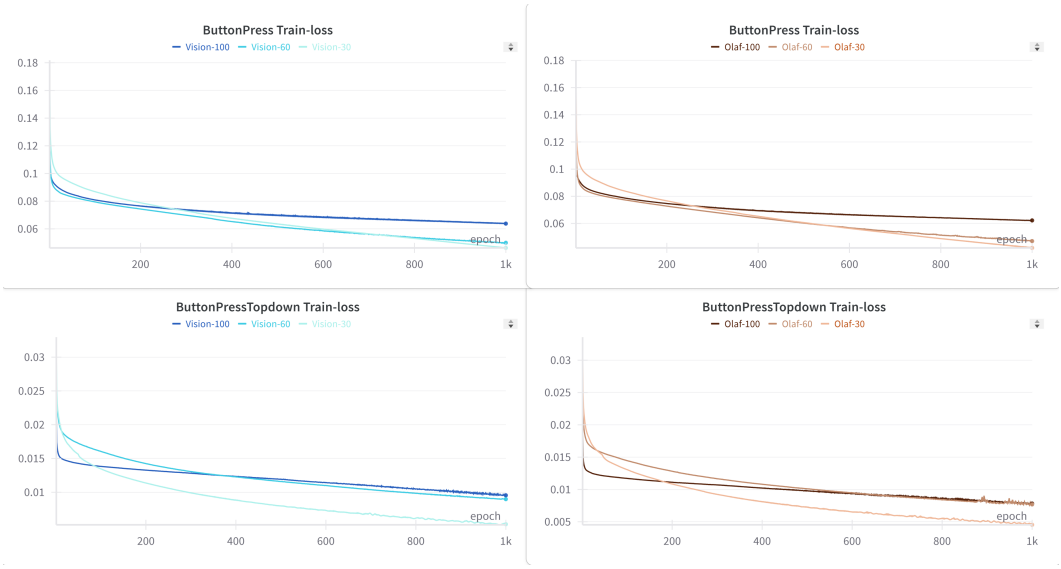

Figure 5: Train Loss

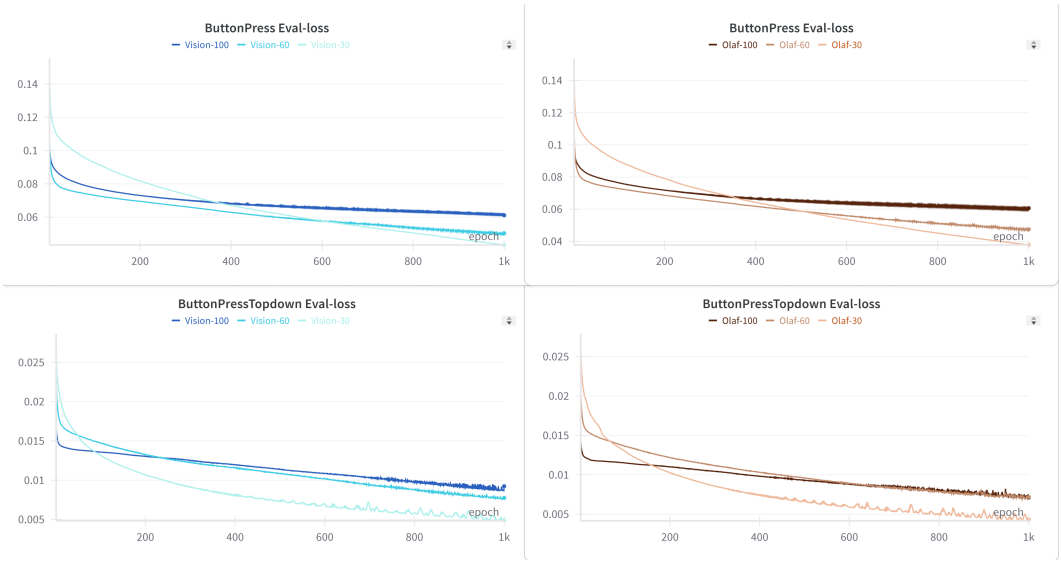

Figure 6: Eval Loss

F Bad Olaf-Vision Response

See Figure7 for an example.

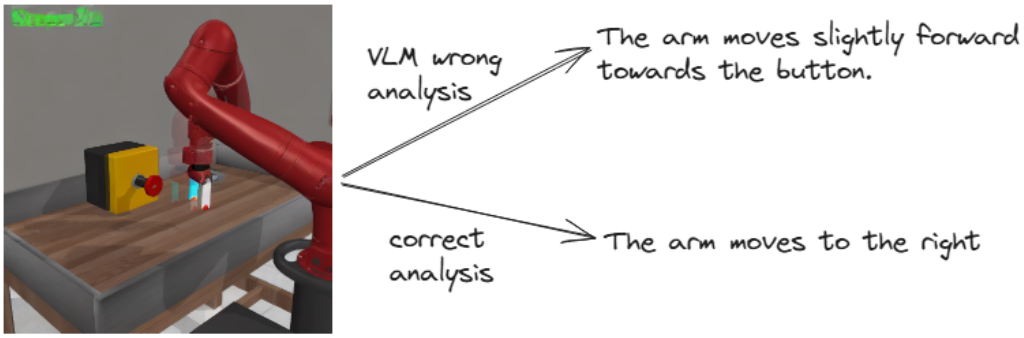

Figure 7: Bad Olaf-Vision Response Example(view our github for more example)