MotIF: Motion Instruction Fine-tuning

Minyoung Hwang[1], Joey Hejna[2], Dorsa Sadigh[2], Yonatan Bisk[3]
[1]MIT, [2]Stanford, [3]CMU

Abstract

While success in many robotics tasks can be determined by only observing the final state and how it differs from the initial state – e.g., if an apple is picked up – many tasks require observing the full motion of the robot to correctly determine success. For example, brushing hair requires repeated strokes that correspond to the contours and type of hair. Prior works often use off-the-shelf vision-language models (VLMs) as success detectors; however, when success depends on the full trajectory, VLMs struggle to make correct judgments for two reasons. First, modern VLMs often use single frames, and thus cannot capture changes over a full trajectory. Second, even if we provide state-of-the-art VLMs with an input of multiple frames, they still fail to correctly detect success due to a lack of robot data. Our key idea is to fine-tune VLMs using abstract representations that are able to capture trajectory-level information such as the path the robot takes by overlaying keypoint trajectories on the final image. We propose motion instruction fine-tuning (MotIF), a method that fine-tunes VLMs using the aforementioned abstract representations to semantically ground the robot's behavior in the environment. To benchmark and fine-tune VLMs for robotic motion understanding, we introduce the MotIF-1K dataset containing 653 human and 369 robot demonstrations across 13 task categories with motion descriptions. MotIF assesses the success of robot motion given task and motion instructions. Our model significantly outperforms state-of-the-art API-based single-frame VLMs and video LMs by at least twice in F1 score with high precision and recall, generalizing across unseen motions, tasks, and environments. Finally, we demonstrate practical applications of MotIF in ranking trajectories on how they align with task and motion descriptions. Dataset, code, and checkpoints are in https://motif-1k.github.io/

1. Introduction

Measuring success in robotics has focused primarily on *what* robots should do, not *how* they should do it. Concretely, *what* is determined by the *final state* of an object, robot, or end-effector [2, 5]. However, not all trajectories that achieve the same final state are *equally successful*. When transporting a fragile object, a path through safer terrain could be considered *more successful* than a shorter yet riskier route (Fig. 1a). Similarly, in the presence of humans a robot's actions when navigating, holding objects, or brushing human hair (Fig. 1 b-d) can cause surprise, discomfort, or pain, making such motions *less successful*.

Success detectors play an important role in robot learning since they evaluate whether or not a robot has completed a task. However, most overlook the importance of "*how*" the task is accomplished, focusing on the initial and final states of the trajectory [2, 3]. This simplification fails to account for tasks that fundamentally require evaluating the entire trajectory to assess success. As we incorporate robots into everyday scenarios, the manner in which they complete tasks will become increasingly important given the context of a scene and its semantic grounding (*e.g.*, avoid collision). Therefore, a more holistic approach to success detection is needed that considers both the task and how the agent should move to complete it.

While modern vision-language models (VLMs) have recently been used as promising tools for success detection [2, 3], they are unable to capture complex notions of how a task is completed for two reasons. First, the majority of VLMs are designed to reason over single images, while success detection in robotics is inherently sequential. Second, even models trained on multiple frames, like video LMs, struggle to recognize fine-grained motion due to a lack of training data. To bridge this gap, we explore how the choice of abstract motion representations, such as visualizing trajectories, affects the performance of both VLMs and video LMs. We propose a trajectory based visual motion representation which overlays a robot's past trajectory on the current or final frame, capturing both the path shape and its semantic connections to the environment. This approach leverages the world knowledge encoded in VLMs and refines it to assess robotic behaviors more effectively.

We propose **motion instruction fine-tuning**, a method that fine-tunes pre-trained VLMs to equip the capability to distinguish nuanced robotic motions with different shapes and semantic groundings. Using the aforementioned trajectory representation, we query our model to output a binary

(a) Outdoor Navigation (b) Indoor Navigation (c) Object-Interactive Manipulation (d) User-Interactive Manipulation

Figure 1. Different robotic motions for various tasks. For each task, we visualize two different motions (path 1 and 2) from real robot demonstrations, where the trajectories share the same initial and final states. Most existing success detectors ignore intermediate states, thereby cannot distinguish them.

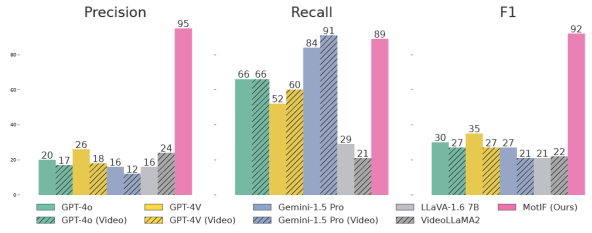


Figure 2. Comparison with off-the-shelf VLMs and Video LMs.

Figure 3. Co-training on Human and Robot Data.

value indicating whether the motion is *correct* (1) or *incorrect* (0). To do so, we collect the **MotIF-1K** dataset, due to limited availability of robot data with diverse semantically grounded motions. We find that co-training mostly on human data with limited robot data enables transfer to robotic motion understanding effectively. **MotIF-1K** contains a variety of motions with 653 human and 369 robot demonstrations across 13 task categories, offering extensive coverage of both the *what* and the nuanced *how* of motion, complete with detailed annotations. It identifies common types of motions featuring varying degrees of semantic grounding, such as the robot's relationship with objects or humans in the environment. The dataset also captures diverse path shapes, in terms of directionality, concavity, and oscillation. For instance, paths in Fig. 1a differ in terms of semantic grounding, where it might be undesirable for a robot to pass over the grass. Fig. 1d describes how straight and curly hairs require different brushing techniques. Notably, MotIF-1K includes subtle motions that are often indistinguishable solely by their start and end states (see project page).

MotIF, a motion discriminator developed by fine-tuning on MotIF-1K, shows further improved success detection on nuanced robot motions. We evaluate MotIF on the test split of MotIF-1K and demonstrate generalization to unseen motions, tasks, and environments (Fig. 2). We significantly outperform state-of-the-art (SoTA) VLMs (*e.g.* GPT-4o, GPT-4V, Gemini-1.5 Pro, LLaVA-1.6 [4], VideoLLaMA2 [1]) with both single and multi-frame (video) input, with at least twice higher in both precision and F1, while maintaining high recall. Co-training human data (653 demos) with minimal robot data (20 demos) significantly improved recall by

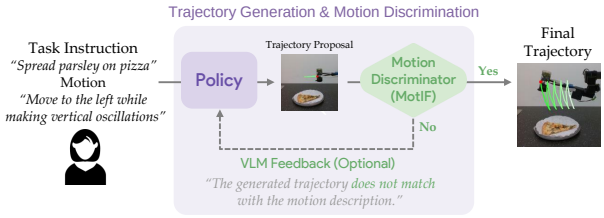
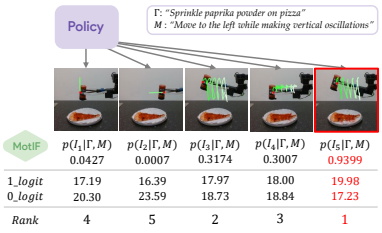


Figure 4. Refining and Terminating Robot Planning. MotIF can close the loop of any existing open-loop controlled system by determining success and giving this as a feedback to the system.



(a) Ranking Example

Model	Win Rate
GPT-4V	0.57
GPT-4o	0.65
Gemini	0.68
MotIF	**0.82**

(b) Win Rate

Figure 5. Ranking Trajectories. We can use MotIF to rank trajectories. (a) $p(I_k|\Gamma, M)$ denotes how likely the motion in the k^{th} image corresponds to the given task instruction Γ and motion description M. (b) Win rate evaluates each model by measuring the prediction accuracy of pairwise rankings.

151.5% over robot-only baselines, demonstrating positive transfer from human to robot data (Fig. 3). As shown in Fig. 4, we can use MotIF to refine and terminate robot trajectories. Also, we demonstrate using MotIF on ranking real robot trajectories from a planner, outperforming SoTA VLMs by at least 20.6% higher win rate.

References

[1] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv:2406.07476*, 2024.

[2] Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-language models as success detectors. *CVPR*, 2023.

[3] Lin Guan, Yifan Zhou, Denis Liu, Yantian Zha, Heni Ben Amor, and Subbarao Kambhampati. "task success" is not enough: Investigating the use of video-language models as behavior critics for catching undesirable agent behaviors. *COLM*, 2024.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[5] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, et al. Homerobot: Open-vocabulary mobile manipulation. *NeurIPS Competition*, 2023.