CONDITIONING ROBOTIC ACTION PREDICTION USING PRETRAINED VIDEO GENERATION MODELS

Adam Imdieke
Video Embodiment
imdie022@umn.edu

ABSTRACT

We propose a novel, zero-shot approach to robotic manipulation that leverages pretrained video generation models to predict intermediate and goal states, for conditioning a low-level policy. We demonstrate the effectiveness of our method on a variety of manipulation tasks, showing performance on tasks never seen during training.

1 INTRODUCTION

Robotics has witnessed significant advancements in recent years due to the integration of generative AI techniques. This is evident in the large parameter foundation models that have revolutionized fields like computer vision and natural language processing. These models utilize internet scale data and immense computational resources to achieve unprecedented levels of performance and generalization.

Despite these advancements, the robotics community still faces challenges in effectively using these models for real-world robotic control and decision-making tasks, especially in dynamic and unstructured environments. The specific knowledge required to understand 3D space and the physical interactions of the robot with its environment is referred to as embodied knowledge. Embodied knowledge varies between robots and environments, making it difficult use generalized models as they lack this crucial context. Another downside of these large models is that they often require significant computational resources and computational time, which can be prohibitive for real-time robotic applications.

We propose a novel approach that leverages pretrained video generation models to condition robotic action prediction. By utilizing the large scale datasets and powerful prediction capabilities of these models, we can predict intermediate and goal images that guide the robot's actions, enabling our lower level policy to use a better aligned goal representation than text.

2 RELATED WORKS

Recent works have explored the use of video prediction for robotic control, directly extracting poses or trajectory information from predicted future frames (Li et al., 2025; Patel et al., 2025), by predicting depth frames from each predicted frame, providing a dense 3D representation of the scene.

Black et al. (2023) propose a method that uses a pretrained image-editing diffusion model to perform zero-shot robotic manipulation. They create a goal image from a text description and a current image using the diffusion model, and then train a goal conditioned behavior cloning (GCBC) model to imitate the predicted actions. We aim to build upon this by integrating video prediction to generate more informative goal and sub-goal representations.

Patel et al. (2025) proposes a method that leverages generated videos and predicted depth images to predict the motion of an object in the scene, enabling the robot to imitate the predicted motion without requiring physical demonstrations. This has similar limitations to Li et al. (2025) where the object must start grasped, and the object should be large to ensure a quality prediction of the pose.

Li et al. (2025) proposed a zero-shot manipulation method that predicts a flow of keypoints on the object to be manipulated, enabling the robot to infer the necessary actions to achieve the desired

state. This approach leverages the generalization of large-scale video prediction models to understand object dynamics, and uses the motion to create a trajectory for the robot to follow. This approach requires the object to be grasped at the start of the interaction, and does not address the full capabilities of the robot's embodiment. The predicted flow is also generated once and cannot allow for a lower level policy to react to new observations or changes in the environment.

Diffusion Policy Chi et al. (2023) is a recent manipulation method that formulates policy learning as a conditional diffusion process, enabling complex, long-horizon robotic tasks to be modeled effectively. It provides state-of-the-art performance on a variety of challenging manipulation tasks, but is limited in generalization by the lack of widespread labeled robotic manipulation data. We aim to address this limitation by defining tasks implicitly using visual goal states, allowing for unlabeled data to be used for training.

We will build upon this line of work by using the same video prediction model Wan et al. (2025) as Li et al. (2025) to predict a sequence of future frames to be used as goal conditioning for a lower level Diffusion Policy Chi et al. (2023). Diffusion policy is usually condition on the current observation and a text embedding representing the task, but we propose to replace the text embedding with the predicted future frames, providing a better alignment between the embeddings for the observations and goals. This approach also reduces the need for the high frequency diffusion policy to explicitly model the entire task, focusing instead on short-term goal achievement.

3 METHODOLOGY

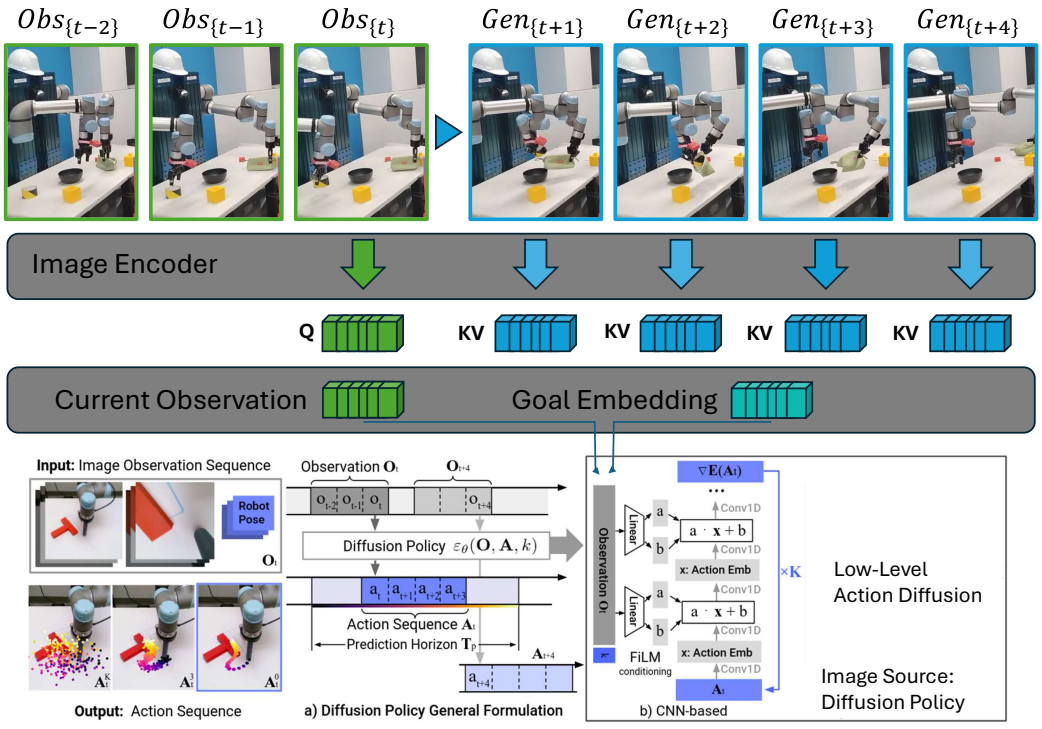

Figure 1: Overview of our approach: We use a pretrained video generation model to predict a sequence of future frames conditioned on the current observation and a high-level task description. These predicted frames serve as a goal for a low-level policy that executes actions to achieve the desired outcome.

We propose a hierarchical approach that integrates video prediction with diffusion-based policy learning. We take advantage of the complementary strengths of both paradigms to address the

challenges of long-horizon tasks, with the reactive lower-level policy handling the complexities of real-time control.

3.1 High-Level Video Prediction

We utilize the video prediction model from Wan et al. (2025) to generate our sequence of future frames, which will act as the high-level plan for the robot to follow. The video prediction model F_{video} takes as input the current visual observation v_t and a text command c_t, outputting a sequence of predicted future frames $\hat{V}_{t+1:t+H} = \{\hat{v}_{t+1}, \hat{v}_{t+2}, ..., \hat{v}_{t+H}\}$, where H is the prediction horizon. An example of the predicted frames is shown in Figure 1 for the text prompt: "Pick up the red block".

3.2 Low-Level Policy with Diffusion

Using the predicted frames at time step t, we define our sequence of goal states $\hat{G}_{t+1:t+H} = \{\hat{g}_{t+1}, \hat{g}_{t+2}, ..., \hat{g}_{t+H}\}$ where $\hat{G} \subset \hat{V}$ that we will use to condition the lower-level policy. The low-level policy $F_{action}(\tau)$ is then responsible for generating actions that move the robot towards these goals, where τ represents the time step of the current state or observation. The low level actions occur at a much higher frequency than the video prediction where $\Delta\tau << \Delta t$.

We will test several goal conditioning methods $\dot{G}$, including using the final predicted frame $\dot{G} = \hat{G}_{t+H}$ as a goal, using a single intermediate frame $\dot{G} = \hat{G}_{t+k}$ where $1 < k < H$, or using attention between the current observation o_t over the entire sequence $\dot{G} = Attention(o_t, \hat{G}_{t+1:t+H})$. For all approaches we use a common ResNet that we process our observations and goals through, maintaining a consistent embedding space.

The low-level policy $F_{action}(\tau)$ will generate a sequence of actions $\hat{A}_\tau = \{\hat{a}_\tau, \hat{a}_{\tau+1}, ..., \hat{a}_{\tau+K}\}$ conditioned on the current observation and the goal embedding, where K is the action horizon. Refer to Chi et al. (2023) for all the details on the diffusion policy architecture. The key idea is that $\hat{A}_\tau$ contains an array of action vectors $\hat{a}$ of the same dimension as the robot's action space, representing a plan of low-level controls to achieve the high-level goal.

The Diffusion Policy starts with $\hat{A}_\tau$ as random noise and iteratively denoises given the conditioning $\dot{G}$ and the current observation embedding, gradually refining the action sequence to produce a coherent plan that makes progress towards the specified goal.

3.3 Training

We use the pretrained video model without fine-tuning, as it is extremely computationally intensive to train. We produce one video per command, with an average generation time over 20 minutes on a single NVIDIA A5500 GPU. Because of the high computational cost, we only generate videos during rollout, instead training the diffusion policy on offline data.

To collect data for the Diffusion Policy, we teleoperate the robot to perform exploratory actions. There is no need to use labeled data, as the diffusion model can learn from the raw state-action trajectories in an autoregressive manner. For this reason, we do not require expensive annotation or task-specific training, preserving the generality of our approach.

We aim to collect a dataset of 2 hours of robot teleoperation data in the tabletop setting with diverse objects and configurations, capturing a wide range of manipulation behaviors.

Our dataset $\mathcal{D} = \{(\tau_i, a_i)\}_{i=1}^N$ consists of action-observation pairs, where the model learns to predict the next K actions given the current observation and goal embedding.

References

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.

Hongyu Li, Lingfeng Sun, Yafei Hu, Duy Ta, Jennifer Barry, George Dimitri Konidaris, and Jiahui Fu. Novaflow: Zero-shot manipulation via actionable flow from generated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12345–12354, 2025. URL https://api.semanticscholar.org/CorpusID:281950672.

Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations, 2025. URL https://arxiv.org/abs/2507.00990.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.