FREE-VIEW MANIPULATION VIA DUAL-FLOW GENERATION POLICY: CANONICAL IMAGES AND ACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visuomotor policies have become the mainstream action generation head in robot manipulation research. However, in open-world environments, the camera viewpoint varies arbitrarily and breaks the perception–action spatial alignment, leading to poor generalization and unstable execution. To address this issue, we propose a novel Dual-Flow Generation Policy (DFGP) that implicitly learns viewpoint-invariant visual representations to adapt to free-view observations and guide accurate manipulation action generation. Our framework consists of two collaborative pathways: an image mapping pathway that infers canonical-view images from free-view inputs via flow matching, and an action generation pathway that predicts robot actions conditioned on the shared canonical visual features. We further introduce a Bi-directional Cross-Attention (BiXT) module to adaptively fuse multi-source features for stronger cross-view action generalization. Extensive experiments conducted on two simulation benchmarks (Free-View and LIBERO) —as well as on physical world robotic platforms demonstrate that DFGP achieves statistically significant improvements over prior visuomotor policies.

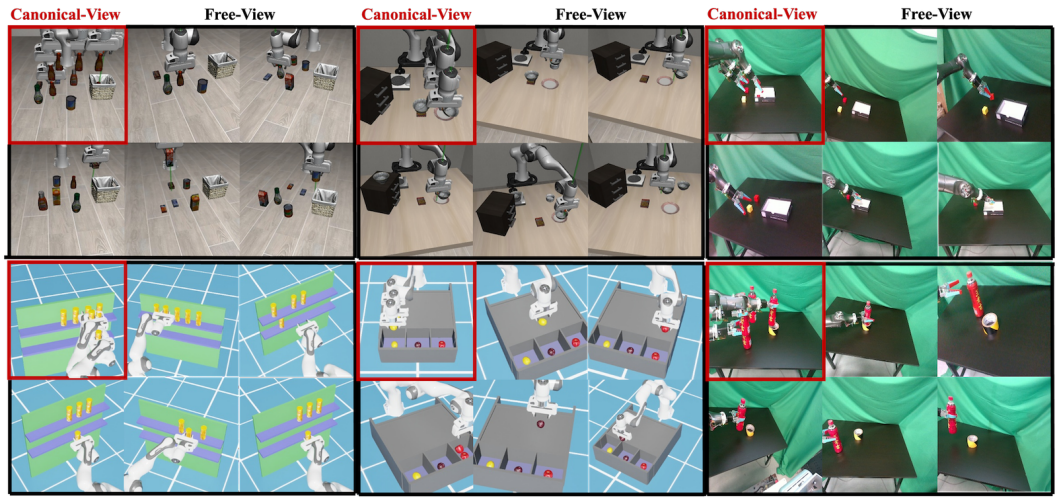


Figure 1: Free View Manipulation. To enable robots to perform manipulation tasks from arbitrary observation viewpoints (marked in black), we propose a novel Dual-Flow Generation Policy that simultaneously generates canonical-view images and subsequent manipulation actions (marked in red). The inferred canonical-view images further assist in generating more accurate actions.

1 INTRODUCTION

In recent years, robotic hardware systems have undergone rapid and diverse advancements, leading to the widespread integration of various robot types into human society Heravi et al. (2023); Duan et al. (2025). Concurrently, both academic and industrial communities have sought to deploy visuomotor policies across different robotic platforms through data-driven approaches, enabling autonomous execution of complex tasks Xu & She (2025). However, due to inherent differences in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

robotic hardware architectures, data collected from a specific platform often fail to generalize effectively to others Jangir et al. (2022); Li et al. (2023). The main cause of this limitation is camera mounting variations, which break the perception-action spatial alignment and lead to poor motor behavior alignment with tasks Meng et al. (2025). This results in trajectory deviations and lower operational precision. Their hardware dependency raises cross-platform deployment costs and hinders the broad application of visuomotor policies in multi-robot systems.

Recent studies have recognized this challenge and proposed various approaches to address it. iDP3 Ze et al. (2025) introduces a method that leverages point clouds from the camera's viewpoint to eliminate dependence on camera calibration entirely. However, this approach requires continuous simultaneous observation of both the robot's end effector and the target object; otherwise, the policy cannot determine their relative spatial configuration. Calibration Diffusion Policy Jin et al. (2025) proposes the concept of free-view manipulation by explicitly incorporating camera calibration parameters into action generation across different camera viewpoints. Nevertheless, obtaining accurate calibration parameters remains challenging in real-world environments. This paper focuses on the free-view manipulation task, eliminating the reliance on calibration parameters and enhancing the robot's adaptability to different viewpoints.

In free-view tasks, issues such as image-action misalignment and occlusion of visual information caused by view deviation are prone to result in a significant decline in the stability of policy outputs. To address this problem, we have designed a view mapping mechanism innovatively, which enables the policy to infer the image under the canonical view based on the observation information from the free view, thereby generating manipulation actions more stably and accurately. In detail, we propose a Dual-Flow Generation Policy (DFGP) that integrates Flow Policy Fang et al. (2026); Zhang et al. (2025) for action generation and Flow Matching Lipman et al. (2023) for image generation. For free-view tasks, each episode corresponds to an independent free observation view. On one hand, we construct an image mapping pathway for the image stream, which extracts visual features from the current observation view and generates the canonical view image at the current timestamp via the flow matching method. On the other hand, we design an action generation pathway for the action stream, which leverages the shared visual representation from the image mapping pathway to guide the generation of subsequent manipulation actions. In addition, we employ a dedicated bi-directional cross-attention transformers mechanism (BiXT) Hiller et al. (2024) to fuse and concatenate historical state information and free-viewpoint information, thereby providing fine-grained visual guidance information. In this way, even though the policy does not receive direct observations from the canonical view, it can still infer the canonical view scene representation through the proposed image mapping pathway.

We conduct experiments in both simulation and real-world scenarios. For the simulation experiments, we regenerated task-specific simulation data across various scenarios based on the LIBERO dataset (built on MuJoCo) Liu et al. (2023) and the Free-View Dataset (built on Isaac Sim) Jin et al. (2025). We further constructed image pairs of free-view and canonical-view for each episode. For the real-world experiments, we collected action data via teaching-based learning using a self-developed robot, and simultaneously captured image pairs of free-view and canonical-view with multiple RGBD cameras. Although canonical-view images are required as supervision during training, such additional information is not required during the testing process. Experimental results demonstrate that the proposed image mapping pathway can effectively enhance the policy's adaptability to free-view tasks and improve the overall task success rate.

The main contributions of our work are summarized as follows:

(1) We present a novel Dual-Flow Generation Policy (DFGP) for free-view manipulation tasks, which consists of an image mapping pathway and an action generation pathway, without relying on extra calibration parameters.

(2) Our method leverages the image mapping pathway to generate canonical-view images from free-view observations, and uses a bi-directional cross-attention transformer module (BiXT) to guide action generation, thereby enhancing the adaptability of the visuomotor policy to different viewpoints.

(3) We recollect partial data and carry out extensive experiments in both simulation and real-world environments. Experimental results demonstrate that our method achieves state-of-the-art average success rates on free-view tasks.

2 Related Works

2.1 Visuomotor Policy

With the development of humanoid robots and embodied intelligence, an increasing number of researchers are focusing on mapping raw visual observations into motor actions directly, namely the visuomotor policy. Early visuomotor policies adopted simple deep learning methods to generate robotic actions directly from images Florence et al. (2019); Chen et al. (2025). Despite their low operational success rates or poor generalization performance, these approaches have garnered researchers' attention by virtue of the convenience offered by their end-to-end action generation mechanism. Action Chunking with Transformers (ACT) Zhao et al. (2023) employs a Conditional Variational Autoencoder and a Transformer backbone to predict sequences of future actions in 'chunks,' effectively mitigating compounding errors. Diffusion Policy Chi et al. (2023); Wang et al. (2025) formulates robot action generation as a conditional denoising process, enabling the model to learn complex, high-dimensional action distributions with superior stability and precision compared to traditional regression-based methods. Flow Policy Fang et al. (2026); Zhang et al. (2025) learns a time-dependent velocity field via flow matching to replace the denoising process of Diffusion Policy, which is a more efficient action generation policy.

In addition, there are also improved works based on these fundamental visuomotor policies to enhance their performance. Hierarchical ACT Park et al. (2024) extend the ACT method for refined manipulation tasks. DP3 Ze et al. (2024), built on Diffusion Policy, is a policy for generating actions by replacing images with point clouds. Adaptive DP Chen et al. (2025) uses temporal-aware diffusion models and low-rank adaptation to learn skills from few demonstrations and fine-tune quickly online. Any2Policy Zhu et al. (2024) generates manipulation actions by aligning various modalities such as images, videos, target states, and point clouds, thereby enhancing the diversity of policy inputs. π_0 Black et al. (2024) and DreamZero Ye et al. (2026) respectively utilized the Vision-Language-Action model and the world action model combined with flow matching to generate generalized robot actions. Building upon prior work in visuomotor policy learning, this paper investigates viewpoint variation in robotic manipulation tasks and proposes a framework that enables policies to generalize across diverse operational viewing perspectives.

2.2 Viewpoint in Manipulation

Early visuomotor policies Florence et al. (2019); Chi et al. (2023); Jin et al. (2024) typically adopt a single viewpoint for observation, and the viewpoint of training and test data are identical. All these methods require a fixed camera position and thus cannot be directly applied to free-viewpoint tasks. Some research Liu et al. (2024); Dong et al. (2025) implement multi-view visuomotor policies with multiple cameras to address issues such as a narrow field of view and occlusions inherent in a single viewpoint. Robouniview Liu et al. (2024) introduces an approach to extract a unified view representation from multiple viewpoints to eliminate camera perspective interference. Luo et al. (2024) leverage multiple fixed camera viewpoints and trains the policy to collaboratively accomplish a single manipulation task. However, both single-view and multi-view manipulation require consistent camera distributions between the training and testing phases. Even a slight variation in viewpoint can render the trained policy ineffective.

Recently, several works have begun to progressively investigate the impact of viewpoint variations on manipulation. Yuan et al. (2025) explicitly align visual features between the canonical viewpoint and other viewpoints to mitigate the interference caused by viewpoint changes. Pang et al. (2025) leverage large-scale multi-view data to enhance the policy's robust performance under viewpoint disturbance. Jin et al. (2025) propose Free-View tasks and corresponding dataset, and improves the adaptability of Diffusion Policy to different viewpoints via additional calibration parameters. We extend the Free-View task by generating canonical viewpoint observations conditioned on arbitrary free viewpoint, thereby enabling implicit alignment of viewpoint-dependent features without requiring explicit camera calibration parameters.

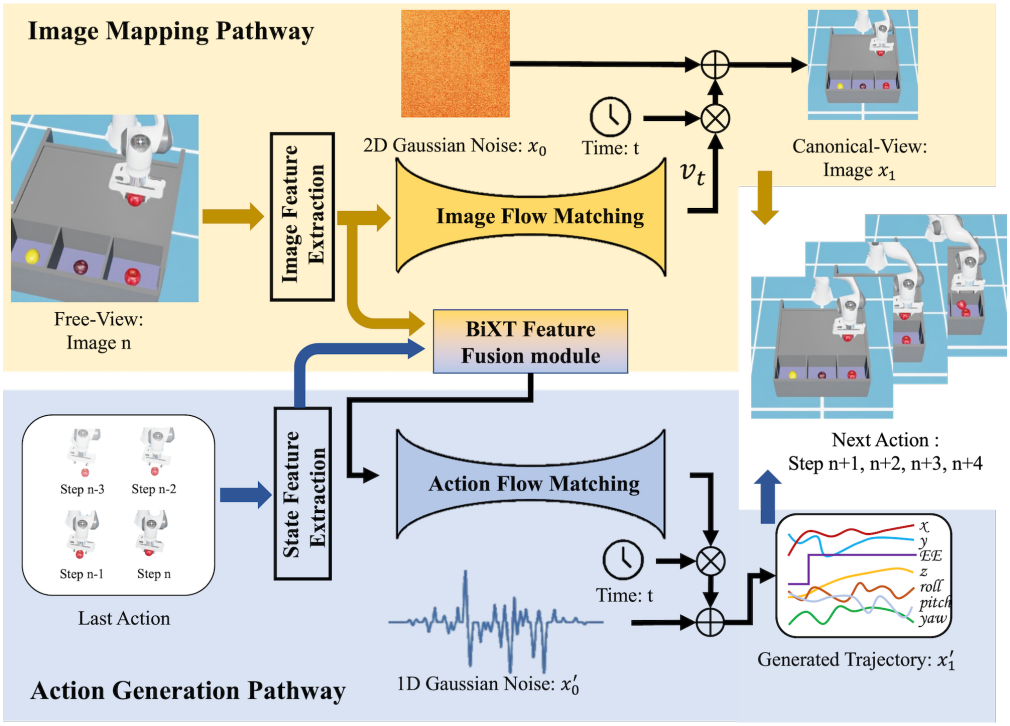


Figure 2: Overall Framework of the Dual-Flow Generation Policy. It consists of an Image Mapping pathway, an Action Generation Pathway, and a dedicated BiXT fusion module. The Image Mapping pathway takes free-view images as conditional input and generates imagined canonical-view images from random 2D Gaussian noise. The BiXT fusion module fuses features to aggregate distinct image-view features and robot state features, producing joint attention-aware fused representations. The Action Generation Pathway takes the fused features as conditions and generates subsequent actions from random 1D Gaussian noise.

3 METHOD

3.1 OVERVIEW

In this paper, we propose a framework for the Dual-Flow Generation Policy, which consists of two components: an image generation pathway and an action generation pathway. The overall framework is illustrated in Figure 2. For the free-viewpoint manipulation task, the observation viewpoint varies in each episode. The Image Mapping Pathway maps images from different viewpoints to the canonical view. Specifically, in Figure 2, the image observed at the current timestep from the current viewpoint is fed into an image feature extraction network to extract image features. These image features are then fed into the Image Flow Matching module as conditional features. The Image Flow Matching module addresses viewpoint inconsistency by establishing a smooth continuous flow transformation between current viewpoint features and canonical view feature, and finally generates the corresponding canonical view image.

The Action Generation Pathway, on the one hand, records the robot's historical actions and extracts action features through a state feature network; on the other hand, employs a BiXT-based feature fusion module to incorporate the intermediate features from the Image Mapping Pathway as visual conditional features. We replace the traditional concat feature fusion strategy with a BiXT based feature fusion module, which adaptively achieves fine-grained alignment between current-view visual features and historical robot-state features. Since the implicit image features are inherently used to generate the canonical view image, they contain features of images under the canonical view to a certain extent. The Action Flow Matching module can utilize historical action features and image features under the canonical view to generate subsequent action steps. The overall loss function con-

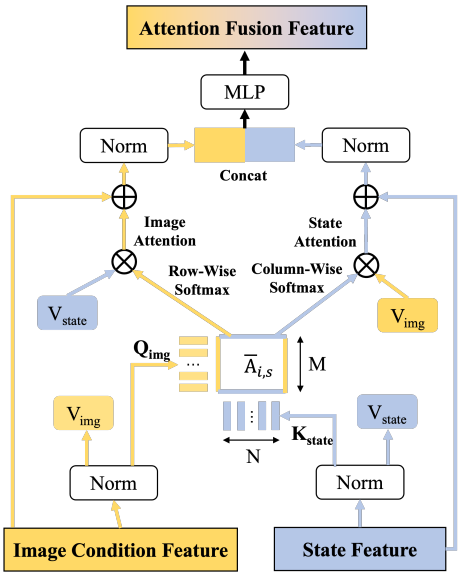


Figure 3: Bi-directional Cross-Attention Transformer Module.

Algorithm 1
Dual-Flow Generation Policy Training

Dataset:
  Image: free-view I_{free}, canonical-view I_c
  Trajectory: history T_h, future T_f
Hyperparameters:
  Total epochs N, Loss Weights λ_1, λ_2
Training:
  Initialize policy $\mathcal{F}$ parameters θ
for *epoch* $= 1$ to N **do**
  $t \sim \mathcal{U}(0, 1)$
  $T_t, I_t \leftarrow Euler(I_c, T_f, t)$
  Predict image field $\hat{\boldsymbol{v}}_t^{img} = \mathcal{F}(I_t, I_{free})$
  Predict action field $\hat{\boldsymbol{v}}_t^{act} = \mathcal{F}(T_t, I_{free}, T_h)$
  Ground-truth image field $\boldsymbol{u}_t^{img}$ from I_c, I_t, t
  Ground-truth action field $\boldsymbol{u}_a^{act}$ from T_f, T_t, t
  $\mathcal{L}_{image} \leftarrow \text{MSE}(\hat{\boldsymbol{v}}_t^{img}, \boldsymbol{u}_t^{img})$
  $\mathcal{L}_{action} \leftarrow \text{MSE}(\hat{\boldsymbol{v}}_t^{act}, \boldsymbol{u}_t^{act})$
  $\mathcal{L}_{total} \leftarrow \lambda_1 \mathcal{L}_{image} + \lambda_2 \mathcal{L}_{action}$
  Update θ via gradient descent $\nabla_\theta \mathcal{L}_{total}$
end for

sists of two components: the loss for canonical-view image generation, and the loss for subsequent action generation.

3.2 Dual Flow Matching

Flow Matching is a powerful framework for generative modeling that simplifies the training of continuous normalizing flows. Instead of dealing with the complex simulation required by Diffusion Models, Flow Matching allows to regress directly onto a fixed vector field. For generative tasks, our goal is to transform a simple probability distribution p_0 (usually random Gaussian noise) into a complex data distribution p_1 (like images or action). This transformation is defined through a Probability Path p_t and a Vector Field $v_t(x)$ that generates this path. For the sample x_t at time t, $x_t \sim p_t(x)$, we can define a conditional vector field $v_t(x_t|x_1)$ that generates a path between x_0 and a specific data point x_1. The loss function is a simple Mean Squared Error (MSE) between our neural network $u_\theta(x_t, t)$ and the target vector field:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,q(x_1),p_0(x_0)} \left[\|u_\theta(x_t, t) - v_t(x_t|x_1)\|^2 \right]. \tag{1}$$

In our experiments, we employ the optimal transport path, which transports points along a straight line:

$$\begin{aligned} \text{Path: } x_t &= (1-t)x_0 + tx_1, \\ \text{Target Vector Field: } v_t(x_t|x_1) &= x_1 - x_0. \end{aligned} \tag{2}$$

Our Dual-Flow Generation Policy consists of two components: the generation of canonical view images and the generation of subsequent manipulation actions. For Image Mapping Pathway, the free-view image features are used as conditions c. x_0 is a 64 * 64 random 2D Gaussian noise, and x_1 is a 64×64 resolution canonical view image. Image Flow Matching employs a 2D UNet to construct an image vector field $u_\theta(x_t, t, c)$, such that the network learns to predict the evolution velocity of samples at time t, thereby transforming the noise x_0 along an optimal transport path toward the canonical-view image x_1 conditioned on free-view image features c. For Action Generation Pathway, we share the image features from the Image Mapping Pathway as the condition c'. Both the initial action noise x_0' and the final generation action x_1' have a dimension of 16×7. It includes 8 past actions and 8 future actions, where each action consists of a 6D end-effector pose and the gripper state, 7 parameters per action. Action Flow Matching employs another 1D UNet to construct an action vector field $u_\theta'(x_t', t, c')$.

3.3 Bi-Directional Cross Attention Feature Fusion

In the free-view task, the camera viewpoint varies across different episodes. Policies need to adapt to changes in viewpoints as much as possible. Traditional feature fusion method merely simply concatenates the features of different modalities and inputs them into a shared neural network, without making a fine distinction between the image features from different viewpoints. To address this issue, we adopt a bi-directional cross attention fusion architecture Vaswani et al. (2017); Hiller et al. (2024) rather than simple feature concatenation Jin et al. (2025). Crucially, we establish fine-grained, bidirectional interaction between different viewpoint observations and robot state representations.

As illustrated in Figure 3, the bidirectional cross-attention transformer (BiXT) module Hiller et al. (2024) is designed for bidirectional fine-grained fusion between image-condition visual features and robot state features. Given the input image-condition feature and state feature, they are individually normalized to yield $\boldsymbol{V}_{img}$ and $\boldsymbol{V}_{state}$. We compute the shared cross-attention matrix $\bar{\boldsymbol{A}}_{i,s}$ via $\boldsymbol{Q}_{img}$ projected from visual feature tokens and $\boldsymbol{K}_{state}$ projected from state feature tokens. Row-wise softmax is applied over $\bar{\boldsymbol{A}}_{i,s}$ to produce image-oriented attention, which updates the visual feature tokens using state-side value $\boldsymbol{V}_{state}$. Meanwhile, column-wise softmax is performed on the same shared matrix to obtain state-oriented attention, refining the state-feature tokens conditioned on $\boldsymbol{V}_{img}$. The updated image-branch and state-branch features are normalized, concatenated, and further fed into an MLP to generate the final attention-fusion feature. Different from conventional sequential bidirectional cross-attention that computes two independent attention matrices, BiXT reuses one common affinity matrix for dual-direction feature update, reducing computational overhead while enabling mutual refinement of visual and state representations.

3.4 Training

Following the training settings of Calibration Diffusion Policy Jin et al. (2025), we also conduct our training on an RTX 4090 GPU. The loss function consists of two components: image loss L_{image}, action loss L_{action}, as shown in Algorithm 1. Both the L_{image} and L_{action} are calculated using the mean squared error (MSE). In our experiments, λ_1 denotes the weight of the image loss and λ_2 corresponds to the weight of the action loss. We set $\lambda_1 = 10$ and $\lambda_2 = 100$ to prioritize accurate action prediction. We employ the flow matching framework for both image and action generation. During training, time points are uniformly sampled from $[0, 1]$ to compute the flow fields. At inference time, we perform generation with 10 discrete steps. Compared with diffusion policies, this design yields substantial speed gains for both training and inference. The input image resolution is set to 256×256, while the resolution of the generated canonical-view images is 64×64, and the length of historical robot states is configured to 8. We use the AdamW optimizer for parameter optimization, and the batch size and learning rate are set to 64 and 5e-4, respectively.

4 Experiments and Discussion

4.1 Experiment Setting

To evaluate the effectiveness of the Dual-Flow Generation Policy in free-view manipulation tasks, we conducted extensive experiments across two simulation benchmarks and a real-world experimental setup. (1) First, within the Free-View Dataset—a benchmark explicitly designed for viewpoint-agnostic manipulation—we systematically regenerate the training data using deterministic, reproducible automated scripts. Meanwhile, we select four tasks: Pull Drawer (**P. D.**), Classify Fruits (**C. F.**), Arrange Bottle (**A. B.**), and Push Stick (**P. S.**). The observation range under our free-view setup is consistent with that of Calibration Diffusion Policy Jin et al. (2025). We collect canonical images from the canonical viewpoint to facilitate model training, with 300 samples gathered for each task. (2) Second, we extended the LIBERO dataset Liu et al. (2023)—a benchmark for generalizable embodied AI—by re-collecting free-view data across three primary task suites: LIBERO_object, LIBERO_goal, and LIBERO_spatial. The free viewpoint is randomly perturbed by ±0.1 m along the x-axis, ±0.5 m along the y-axis, and ±0.15 m along the z-axis. The camera orientation is adjusted to keep the manipulated object within the camera's field of view. Similarly, we collect only 300 samples for each task. Given that the robot actions in LIBERO are conditioned on natural language instructions within the same scene, we integrate an auxiliary language-based task feature extraction network into the Action Generation Pathway to encode and fuse semantic instruction embeddings.

(3) Finally, we establish a physical experimental platform, as illustrated in Figure 4, to collect robot manipulation data via human teleoperation, and design two real-world tasks: Pick Bottle (**P. B.**) and Select Cube (**S. C.**). A primary camera positioned directly in front of the workspace is used to record canonical-view images, while multiple RealSense cameras deployed at various external positions captured diverse free-view observations. To perform the evaluation, we run the trained policy in both simulated and real-world environments. We execute 50 task episodes per evaluation trial to obtain the success rate. The evaluation is repeated five times by using different random seeds, and we report the corresponding median and extremum deviation.

Table 1: Main Results in Free-View tasks

Method	Free-View				Free-LIBERO			Physical World	
	P. D.	**C. F.**	**A. B.**	**P. S.**	Object	Spatial	Goal	**P. B.**	**S. C.**
BC-T	0%	18%	6%	26%	/	/	/	/	/
ACT	4%	8%	0%	10%	/	/	/	/	/
DP	54%	32%	52%	62%	72%	40%	42%	14%	30%
FP	44%	40%	54%	54%	70%	32%	50%	26%	24%
Cali. DP	58%	52%	74%	80%	/	/	/	/	/
Ours	72±4%	60±6%	78±4%	82±6%	78±2%	58±4%	54±2%	36±10%	56±8%

4.2 Main Results

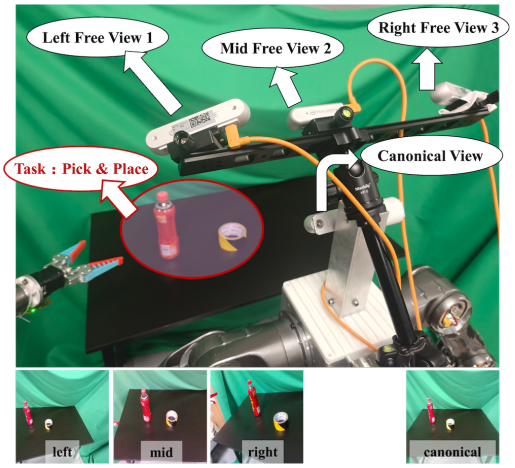


Figure 4: Physical World Experimental Setup. We use a fixed primary camera to capture canonical-view images. The other three cameras are mounted on the same stand, which is randomly positioned before each data collection episode. This setup ensures that all cameras can simultaneously observe the robot manipulation.

The main experimental results are shown in Table 1. Since this work only introduces improvements to the Action Head module for action generation, we primarily compare it with prior Action Head methods. Behavior Cloning Transformer (BC-T) Florence et al. (2022) is a traditional imitation learning approach that uses a Transformer for action generation, yet its predicted actions exhibit substantial uncertainty. Action Chunking with Transformers (ACT) Zhao et al. (2023) alleviates action uncertainty via action chunking. However, the raw actions directly generated by the transformer are not well-suited for free-viewpoint manipulation tasks. Visuomotor policies based on probability distributions are more capable of adapting to changes in viewpoints. From the experimental results, both Diffusion Policy Chi et al. (2023) and Flow Policy Fang et al. (2026) achieve higher manipulation success rates. Calibration Diffusion Policy Jin et al. (2025) leverages additional calibration information to further improve the success rate of free-viewpoint manipulation, but it is quite sensitive to calibration errors. Our method has achieved further improvements and does not require accurate calibration information; instead, by introducing supervision from the canonical viewpoint, more accurate operational actions can be generated.

By comparing experimental results across different datasets, we can observe that the performance in simulation is evidently better than that in the real physical world. This is because the data collected in simulation contains much less inherent noise. In this work, only a small amount of real-world data is collected for physical validation, making the real-world experiments more susceptible to interference from illumination changes, visual sensor noise, and other environmental perturbations. We believe that introducing large multimodal models or world models can further bridge the performance gap between simulation and physical scenarios, and effectively improve the generalization and robustness of the model in physical world environments.

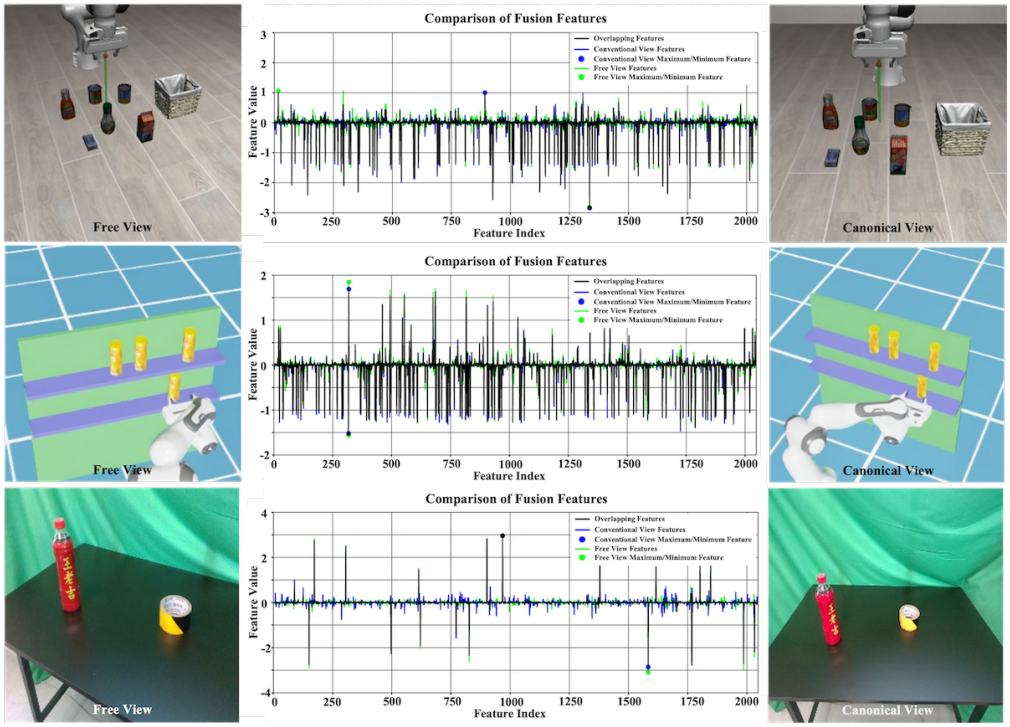


Figure 5: Compare image features from different viewpoints. Left: free-view images; Right: canonical viewpoint images. Middle: the one-dimensional features obtained by flattening the extracted image features, where the black parts represent the overlapping regions of feature values.

4.3 Quantitative Comparisons on Fusion Features

In this paper, we employ the image mapping pathway to learn visual features under the canonical viewpoint from images captured at different viewpoints, and fuse them with robot state information via the BiXT module, thereby obtaining fused features. To examine whether this mapping effectively bridges free-view and canonical-view representations, we visualize the fused features in Figure 5. Specifically, for the trained DFGP model, we input the free-view image and the canonical-view image at the same timestep respectively, extract the intermediate fused features, and flatten them into one-dimensional vectors for comparison.

Table 2: Ablation Study on Fusion Methods

Method	Free-LIBERO		
	Object	Spatial	Goal
Concat	64%	40%	24%
MoE	66%	48%	40%
I2S Attention	32%	0%	0%
S2I Attention	24%	8%	0%
Dual Attention	70%	56%	46%
BiXT	78%	58%	54%

The middle feature plots reveal that for the identical task, the fused features from the free viewpoint achieve high similarity and overlap with those from the canonical viewpoint, whereas distinct differences exist for different tasks. The peak and trough feature values further verify the effectiveness of our mapping module. Based on these observations, we believe that the image mapping pathway can adapt to different viewpoints to ensure that the policies are not disturbed by viewpoint shifts.

4.4 Ablation Study on Different Fusion Methods

In this paper, we propose a novel BiXT feature fusion module that jointly encodes each different view visual features and robot state representations through bidirectional cross-attention, enabling adaptive aggregation of free-view features tailored to diverse manipulation tasks. We conduct ablation studies on the LIBERO simulation dataset to compare the traditional concatenation fusion and our BiXT fusion module. We explore simple Concatenation, Mixture-of-Experts (**MoE**), Im-

age to State unidirectional attention (**I2S Attention**), State to Image unidirectional attention (**S2I Attention**), conventional bidirectional cross-attention (**Dual Attention**), and our BiXT with shared similarity matrices. Detailed architectures of these fusion methods are provided in Appendix A.1. As shown in Table 2, the task success rate of MoE-based fusion is only slightly higher than that of traditional concatenation fusion. We further observe that the MoE model tends to activate only a single expert during training. We conjecture that this issue arises from limited training data, making it difficult for MoE to learn to select distinct experts for different viewpoints. In addition, we also evaluate various attention fusion schemes. Unidirectional attention fusion may incur feature loss and cannot boost the task success rate. While independent bidirectional cross-attention yields beneficial fusion performance, it demands two separate cross-attention computations and subsequent concatenation. Inspired by this observation, we develop BiXT for our Dual-Flow Generation Policy, which utilizes a shared similarity matrix and attains a higher manipulation success rate.

4.5 Quality of Generated Canonical Images

Our image generation pathway synthesizes canonical-view images from free-view observations. Here, we evaluate the quality of the generated canonical-view images by adopting evaluation metrics from image generation. PSNR (Peak Signal-to-Noise Ratio) is a classic pixel-level metric that quantifies reconstruction fidelity based on mean squared error; and SSIM (Structural Similarity Index) Wang et al. (2004) is a perception-oriented metric that evaluates image similarity in terms of luminance, contrast, and structure. As shown in Table 3, we select one task from each of the three datasets for evaluation. In general, the generated canonical-view images are consistently close to the ground truth. This is because the target canonical-view images mapped from different viewpoints share high similarity, making the image generation task prone to overfitting. The generated canonical-view images on the simulation datasets LIBERO and Free-View are significantly superior to those generated in the physical world. This is because images obtained in simulation are much cleaner, whereas images captured by cameras in real-world scenarios inherently contain a certain amount of noise. We visualize the generated canonical-view images in Figure 6 (more results are shown in Appendix A.4). It can be observed that in physical world experiments, the policy sometimes fails to generate accurate canonical-view images. Common causes of errors are object offset and missing.

Table 3: Canonical-View Image Generation.

Datasets	Metric	
	PSNR↑	SSIM↑
LIBERO (Goal)	96.0	0.999
Free-View (Classify Fruits)	95.4	0.997
Physical (Pick Bottle)	82.7	0.814

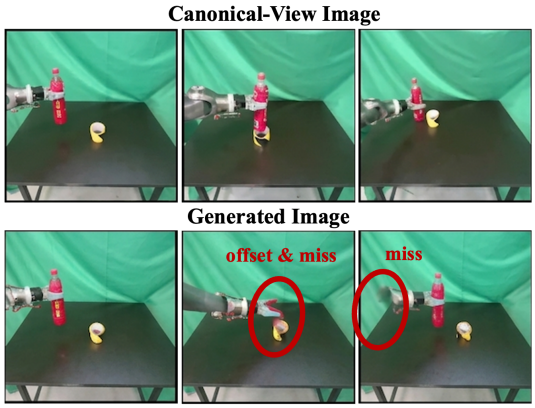


Figure 6: Visualization results of generated canonical-view images. In the generated images, the first one is accurate, while the second and third ones both have problems of object offset and loss.

5 Conclusion

In this work, we address the challenging problem of free-view robot manipulation by introducing a novel Dual-Flow Generation Policy (DFGP). The core idea is to implicitly learn viewpoint-invariant visual representations to adapt to arbitrary free-view observations. Our method jointly performs canonical-view image generation and robot action generation through two collaborative flow-matching pathways, and uses a shared similarity matrix BiXT fusion module to adaptively fuse visual features and robot states. By inferring canonical-view representations from arbitrary free-view observations, DFGP effectively alleviates view misalignment and improves the stability and generalization of visuomotor policies. Extensive experiments on simulation benchmarks and real-world platforms show that our approach outperforms previous methods. In the future, we will extend DFGP to long-horizon and multi-task manipulation scenarios to further enhance its versatility.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript polishing, sentence rephrasing, and grammar correction of the English text. We also leveraged generative AI tools to assist with literature lookup and retrieval of related papers and methods. We have not used generative AI tools for research idea formulation, algorithm design, data collection, experimental implementation, or result analysis. All AI-assisted outputs, including the revised text and retrieved literature summaries, have been manually reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work conducts robot manipulation experiments in simulated environments and a physical laboratory setup. No human subjects, private personal data, or sensitive information were involved in data collection. The proposed visuomotor policy is designed for benign robotic manipulation tasks and carries no foreseeable harmful applications. We confirm there are no conflicts of interest, discrimination, privacy or legal compliance concerns associated with this research.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, all simulation datasets are recollected based on a prior public framework. Upon publication, we will release the source code along with complete experimental configurations, including model hyperparameters, environment settings, and evaluation protocols. The implementation details of our DFGP and BiXT modules are described in the main paper and appendix to support the reproduction of our free-view robot manipulation experiments.

REFERENCES

Kevin Black, Zhidao Bu, Roberto Calandra, Nick Cass, Siyuan Chen, Yunzhu Chu, Coline Costello, Yilun Cui, Yilun Du, Alexei A Efros, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Dechao Chen, Zhengwen Chen, Xiangyan Zheng, Weiling Xu, Chencong Ma, and Chentao Mao. Adp: Adaptive diffusion policy energizes robots thinking in both learning and practice. *IEEE Transactions on Automation Science and Engineering*, 22:21585–21594, 2025. doi: 10.1109/TASE.2025.3612396.

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.

Ju Dong, Lei Zhang, Liding Zhang, Yao Ling, Yu Fu, Kaixin Bai, Zoltán-Csaba Márton, Zhenshan Bing, Zhaopeng Chen, Alois Christian Knoll, et al. M4diffuser: Multi-view diffusion policy with manipulability-aware control for robust mobile manipulation. *arXiv preprint arXiv:2509.14980*, 2025.

Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models. In *Conference on Robot Learning (CoRL)*, pp. 5326–5350, 2025.

Yu Fang, Xuehe Zhang, Haoshu Cheng, Xizhe Zang, Rui Song, and Jie Zhao. Flow policy: Generalizable visuomotor policy learning via flow matching. *IEEE/ASME Transactions on Mechatronics*, 31(1):140–150, 2026.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning (CoRL)*, pp. 158–168, 2022.

Peter Florence, Lucas Manuelli, and Russ Tedrake. Self-supervised correspondence in visuomotor policy learning. *IEEE Robotics and Automation Letters*, 5(2):492–499, 2019.

Negin Heravi, Ayzaan Wahid, Corey Lynch, Pete Florence, Travis Armstrong, Jonathan Tompson, Pierre Sermanet, Jeannette Bohg, and Debidatta Dwibedi. Visuomotor control in multi-object scenes using object-aware representations. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9515–9522, 2023.

Markus Hiller, Krista A Ehinger, and Tom Drummond. Perceiving longer sequences with bi-directional cross-attention transformers. *Advances in Neural Information Processing Systems*, 37:94097–94129, 2024.

Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):3046–3053, 2022.

T. Jin, W. Song, T. Zhou, et al. Free-view robot manipulation: Visuomotor policy by calibration diffusion. *Available at SSRN*, 6379322, 2025.

Tianlei Jin, Hongwei Zhu, Jiakai Zhu, Shiqiang Zhu, Zaixing He, Shuyou Zhang, Wei Song, and Jason Gu. Whole-body inverse kinematics and operation-oriented motion planning for robot mobile manipulation. *IEEE Transactions on Industrial Informatics*, 20(12):14239–14248, 2024.

Guotao Li, Xu Liang, Yifan Gao, Tingting Su, Zhijie Liu, and Zeng-Guang Hou. A linkage-driven underactuated robotic hand for adaptive grasping and in-hand manipulation. *IEEE Transactions on Automation Science and Engineering*, 21(3):3039–3051, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems (NIPS)*, 36:44776–44791, 2023.

Fanfan Liu, Feng Yan, Liming Zheng, Chengjian Feng, Yiyang Huang, and Lin Ma. Robouniview: Visual-language model with unified view representation for robotic manipulation. *arXiv preprint arXiv:2406.18977*, 2024.

Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. Multistage cable routing through hierarchical imitation learning. *IEEE Transactions on Robotics*, 40:1476–1491, 2024.

Qiwei Meng, Jason Gu, and Yun-Hui Liu. Gpd: Learning geometric primitive deformation for unseen object pose estimation. *IEEE Transactions on Automation Science and Engineering*, 22: 9903–9922, 2025.

Jing-Cheng Pang, Nan Tang, Kaiyuan Li, Yuting Tang, Xin-Qiang Cai, Zhen-Yu Zhang, Gang Niu, Masashi Sugiyama, and Yang Yu. Learning view-invariant world models for visual robotic manipulation. In *International Conference on Learning Representations (ICLR)*, 2025.

J Hyeon Park, Wonhyuk Choi, Sunpyo Hong, Hoseong Seo, Joonmo Ahn, Changsu Ha, Heungwoo Han, and Junghyun Kwon. Hierarchical action chunking transformer: Learning temporal multimodality from demonstrations with fast imitation behavior. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12648–12654. IEEE, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Zhendong Wang, Max Li, Ajay Mandlekar, Zhenjia Xu, Jiaojiao Fan, Yashraj Narang, Linxi Fan, Yuke Zhu, Yogesh Balaji, Mingyuan Zhou, et al. One-step diffusion policy: Fast visuomotor policies via diffusion distillation. In *International Conference on Machine Learning (ICML)*, pp. 63399–63416, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Zhengtong Xu and Yu She. Leto: Learning constrained visuomotor policy with differentiable trajectory optimization. *IEEE Transactions on Automation Science and Engineering*, 22:8567–8578, 2025. doi: 10.1109/TASE.2024.3486542.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Zhecheng Yuan, Tianming Wei, Shuiqi Cheng, Gu Zhang, Yuanpei Chen, and Huazhe Xu. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. In *Conference on Robot Learning (CoRL)*, pp. 1815–1833, 2025.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Robotics: Science and Systems (RSS)*, 2024.

Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with 3d diffusion policies. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2873–2880. IEEE, 2025.

Qinglun Zhang, Zhen Liu, Haoqiang Fan, Guanghui Liu, Bing Zeng, and Shuaicheng Liu. Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 14754–14762, 2025.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

Yichen Zhu, Zhicai Ou, Feifei Feng, and Jian Tang. Any2policy: Learning visuomotor policy with any-modality. *Advances in Neural Information Processing Systems (NIPS)*, 37:133518–133540, 2024.

A Appendix

A.1 Different Feature Fusion Methods

In Section 4.4, we evaluate different fusion strategies. Figure 7 illustrates the detailed fusion architectures. For the MoE architecture, when we set four experts with only one activated per forward pass, training tends to route all inputs to a single expert. When two experts are selected, training routes all inputs through both experts. After adding the load-balancing loss, the success rate during inference drops substantially. We hypothesize that this arises because our policy is trained on merely 300 trajectories, and the limited data volume is insufficient to support the training of multiple experts.

Unidirectional attention architectures, including state-to-image attention and image-to-state attention, use one modality as queries (Q) and the other as keys (K) and values (V). The results suggest that a single unidirectional attention branch tends to discard partial features. Concatenating the features from both attention branches yields higher task success rates compared with naive concatenation. Our BiXT is also a bidirectional cross-attention module. The difference lies in that it constructs a single similarity matrix and computes attention along both row and column directions, achieving higher computational efficiency.

A.2 Visualization Shared Cross-Attention Matrix

We visualize the shared cross-attention similarity matrix learned by the BiXT module in Figure 8. We set the number of BiXT attention heads to 16. Using the Pull Drawer task in Figure 8 as an example, we record the average values of different attention heads throughout inference steps. Combined with viewing directions and the trends of different heads over inference steps, attention heads under nearby viewpoints (the first and second rows) show more similar patterns. The third row captures observations from the opposite side, yielding distinct attention-head trends. This partly indicates that the similarity matrix in BiXT is capable of learning viewpoint-associated features.

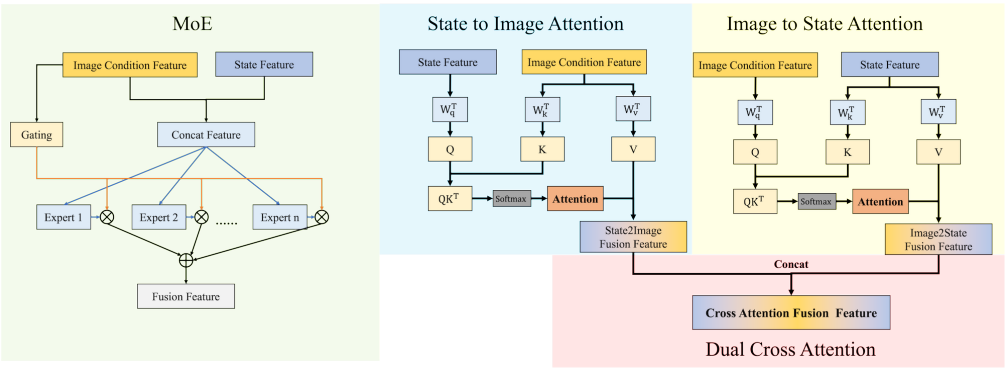


Figure 7: Network architectures of different feature fusion methods. Left: Mixture-of-Experts (MoE) fusion module, where image and state features are concatenated and routed to different experts via a gating network. Right: Independent Cross Attention, which contains two decoupled cross-attention branches: state-to-image attention and image-to-state attention. The two branches compute attention independently, and their outputs are concatenated to produce the final fused feature.

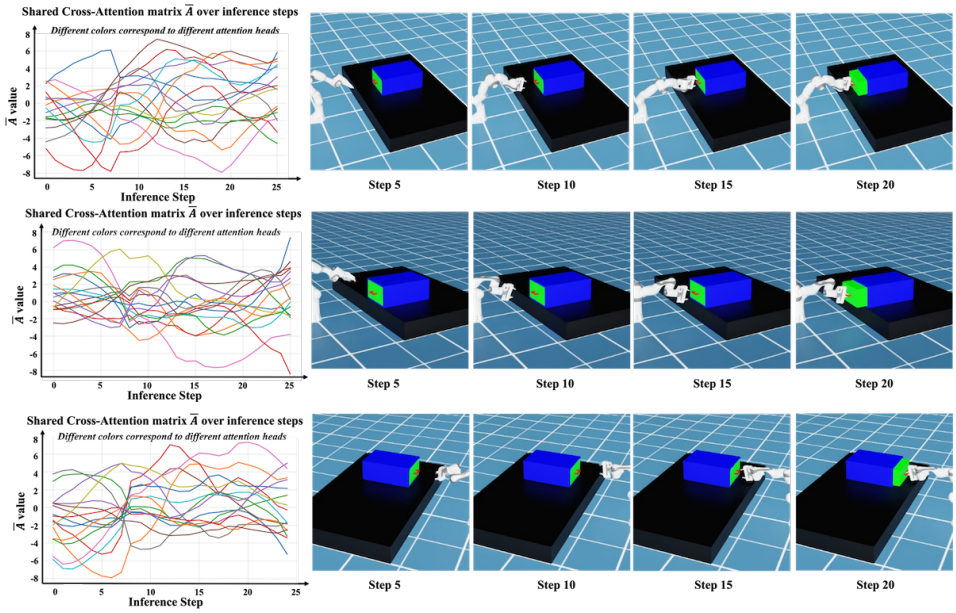


Figure 8: Dynamics of the shared cross-attention matrix $\overline{A}$ over inference steps. Each colored curve visualizes the evolution of $\overline{A}$ for a separate attention head. Corresponding robot simulation snapshots at inference steps 5, 10, 15 and 20 are shown on the right.

A.3 Ablation Study on Policy

In this work, we mainly adopt flow policy as our action generation strategy. In principle, this policy can be replaced by other generative learning policies, such as diffusion policy. In Table 4, we present results using diffusion policy for action generation within our dual generation policy, and further investigate how the number of denoising steps influences manipulation performance. We observe that when the number of denoising steps is small, diffusion policy yields unstable actions due to insufficient denoising iterations. As the number of denoising steps increases, diffusion policy achieves increasingly satisfactory performance. In contrast, flow matching learns a continuous vector field and inherently supports one-step generation. Only a few Euler steps are sufficient to produce stable manipulation actions, and increasing the number of steps brings only marginal improvements.

Notably, flow matching relies on deterministic Euler sampling, which provides significant advantages in generation efficiency compared with denoising-based probabilistic models such as diffusion policy.

Table 4: Ablation on Policy.

Backbone Policy	Sampling Steps	Task			FPS
		Pull Drawer	Classify Fruits	Arrange Bottle	
Diffusion Policy	10 step	10%	20%	8%	32.3
	100 step	62%	42%	48%	4.2
Flow Policy	10 step	72%	60%	78%	31.6
	100 step	74%	60%	72%	4.6

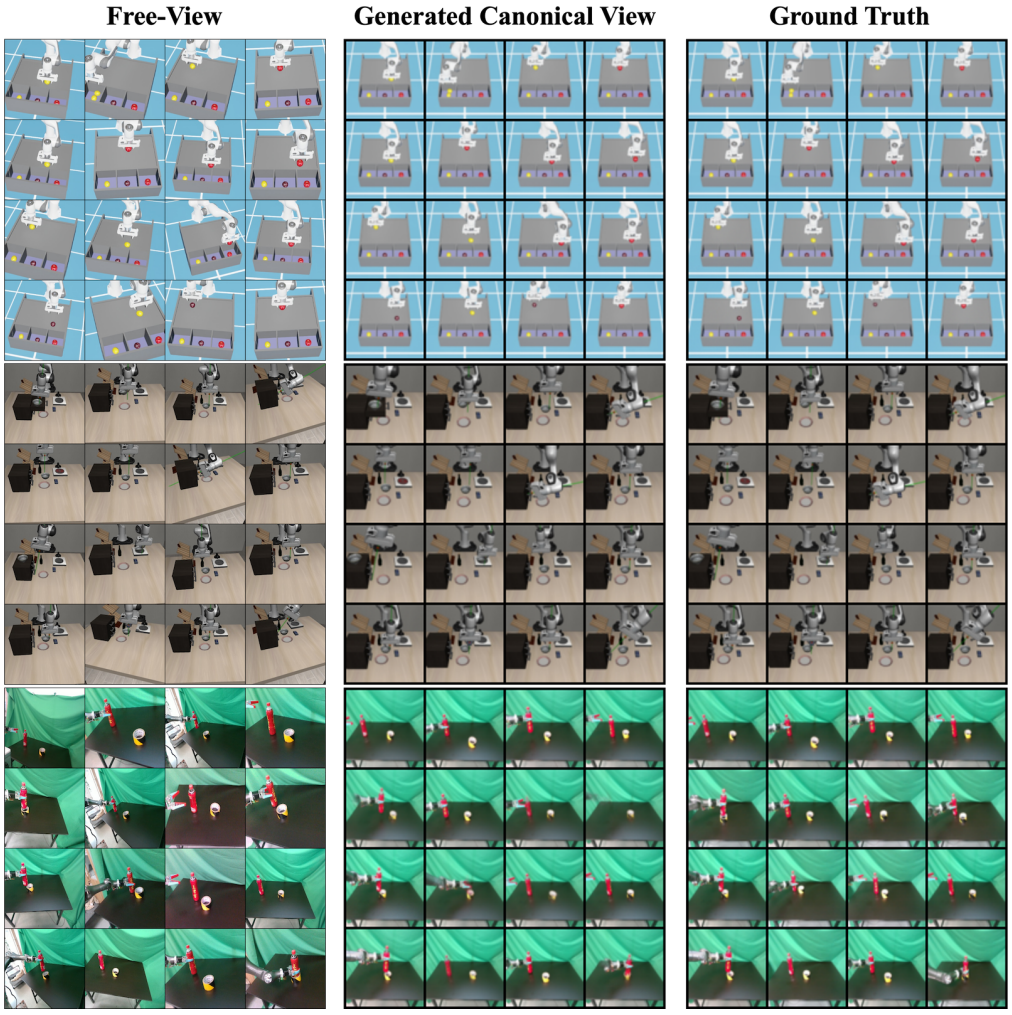


Figure 9: Visual comparison of canonical-view image generation from free-view inputs.. From left to right: free-view input observations with a resolution of 256×256, canonical-view images generated by our DFGP at 64×64 resolution (which appear slightly blurry), and ground-truth canonical-view images of 64×64 resolution used for supervision during training.

A.4 Qualitative Results of Canonical-View Generation

As shown in Figure 9, given input images captured from arbitrary free viewpoints, our model can generate plausible canonical-view images that closely match the ground truth. The generated images preserve the spatial layout of objects and the state of the robot end-effector, even when the input observation is captured from a side or oblique viewpoint. This capability verifies that the image mapping pathway successfully transfers visual information from varying free viewpoints into the unified canonical viewpoint space. Such consistent canonical representations provide reliable visual features for the subsequent action generation branch of our policy. From the generation results, the generated canonical-view images in the simulated environment closely match the ground truth. In contrast, canonical-view images generated from free-view observations captured in the real world are more prone to misalignment or missing visual details. Consistent with this observation, the success rate of manipulation tasks in the real world is also relatively lower.

A.5 Limitation and Future Work

Our method requires using the canonical-view images as the supervisory signal during the training process. Although this information is not needed during testing, collecting such paired data in the real world requires additional multi-view camera setups and manual synchronization, which increases the engineering burden. The current experiments are mainly focused on static and rigid desktop-level operation tasks. The generalization to more complex scenarios, such as deformable object manipulation, multi-object interaction, or long-duration sequential tasks, still needs to be verified. In particular, in dynamic environments, the temporal consistency of the generated canonical-view images may decrease, which may introduce errors in the action generation path.

In the future, we will leverage the scene perception capabilities of multimodal large models to further enhance the success rate and generalization of tasks in the physical world. At present, numerous vision-language-action models and world models have been widely studied for robotic manipulation, yet few of them can adapt well to free-view tasks. Most existing methods rely on fixed camera viewpoints and stable spatial alignment between visual inputs and robot actions. We will also integrate vision-language foundation models to enhance semantic cross-view alignment, and explore self-supervised learning policy to eliminate reliance on supervision from paired canonical views. In addition, we also expect future work to incorporate temporal modeling to improve the stability of canonical-view image generation in dynamic scenes.