2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Factorizing Diffusion Policies for Observation Modality Prioritization

Omkar Patil[1], Prabin Kumar Rath[1], Kartikay Pangaonkar[1], Eric Rosen, Nakul Gopalan[1]

***Abstract*— Diffusion models have been extensively leveraged for learning robot skills from demonstrations. These policies are conditioned on several observational modalities such as proprioception, vision and tactile. However, observational modalities have varying levels of influence for different tasks that diffusion polices fail to capture. In this work, we propose 'Factorized Diffusion Policies' abbreviated as FDP, a novel policy formulation that enables observational modalities to have differing influence on the action diffusion process by design. This results in learning policies where certain observations modalities can be prioritized over the others such as `vision>tactile` or `proprioception>vision`. FDP achieves modality prioritization by factorizing the observational conditioning for diffusion process, resulting in more performant and robust policies. Our factored approach shows strong performance improvements in low-data regimes with 15% absolute improvement in success rate on several simulated benchmarks when compared to a standard diffusion policy that jointly conditions on all input modalities. Moreover, our benchmark and real-world experiments show that factored policies are naturally more robust with 40% higher absolute success rate across several visuomotor tasks under distribution shifts such as visual distractors or camera occlusions, where existing diffusion policies fail catastrophically. FDP thus offers a safer and more robust alternative to standard diffusion policies for real-world deployment. Code and videos are available at `https://fdp-policy.github.io/fdp-policy/`.**

I. Introduction

Humans prioritize different sensory modalities according to the specific requirements of the task [1]. For instance, Wahn and König [1] note that participants engaged in visually demanding tasks are comparatively less receptive to auditory stimuli. They argue that this flexible allocation of human attentional capacity maximizes the capability to process relevant information. Further, humans have been shown to prioritize the more reliable modality between vision and haptics in different situations [2]. This naturally raises the question of whether policy learning could benefit from such prioritization of observational modalities influencing robot actions. **Prioritization of the more influential or reliable modality could enable robot policies to learn skills more efficiently and avoid developing spurious correlations with noisy modalities**. Moreover, the number of possible skills is vast and skills may depend more strongly on certain observational modalities over others. For instance, repetitive motions like sweeping are more likely to depend on the robot's proprioception, while locating an object for manipulation is conditioned strongly on its vision. This necessitates a need for an efficient skill learning method that considers the varying levels of influence that different observational modalities may have.

Diffusion models [3] have been extensively leveraged for learning robot skills from demonstrations [4]. The current methods for training diffusion policies, jointly condition the action diffusion process $\boldsymbol{x}$ on all M observational modalities $\boldsymbol{y}^{1:M}$ for every task [4]. This is a monolithic joint conditioning approach - "when all you have is a hammer, everything looks like a nail". Existing diffusion policies do not flexibly accommodate the differing influence of various observational modalities. We empirically show that this joint conditioning approach hurts the sample complexity of diffusion policies as it is difficult to learn the level of dependence that actions have on the observational modalities with limited data. Further, diffusion policies are sensitive to small distribution shifts in *any* modality $\boldsymbol{y}^{1:M}$ that they conditions upon, and cannot 'de-prioritize' the modality susceptible to noise, similar to humans [2]. Incorporating robustness to such shifts require a prohibitively large amount of data when the observation modalities are high-dimensional. To that end, we propose a novel policy formulation called Factorized Diffusion Policies (FDP) for enabling prioritization of observational modalities during policy learning.

At its core, FDP learns a diffusion *base policy* using k ($k < M$) input modalities $\boldsymbol{y}^{1:k}$ to be prioritized, followed by a diffusion *residual policy* that learns the noise or score residual conditioned on all modalities $\boldsymbol{y}^{1:M}$. We provide a probabilistic formulation to the residual from the first principles, and also develop a novel architecture for efficiently learning it. The base and residual models are then composed to obtain samples from the full conditional action distribution $p(\boldsymbol{x}|\boldsymbol{y}^{1:M})$. By enabling modality prioritization, FDP introduces flexibility in learning diffusion policies with the inclusion of prioritization order of observational modalities in the hyperparameter search space. Through extensive experiments across visual (`vision`), point-cloud (`pcd`), proprioceptive (`prop`), environment-state (`env-state`) and tactile (`tactile`) modalities, we show that leveraging this added flexibility results in more performant and robust diffusion policies. Our contributions are as follows.

- We propose FDP, a novel policy formulation that enables observational modalities to have differing influence on the action diffusion process. We mathematically derive a framework to split the jointly conditioned policy into a base policy learned with prioritized modalities and a residual policy learned with all modalities.
- We show the merits of modality prioritization through extensive **experimentation across visual, point-cloud, proprioceptive and tactile** observational inputs on several

[1]School of Computing and Augmented Intelligence, Arizona State University. Corresponding author emails- {opatil3, ng}@asu.edu

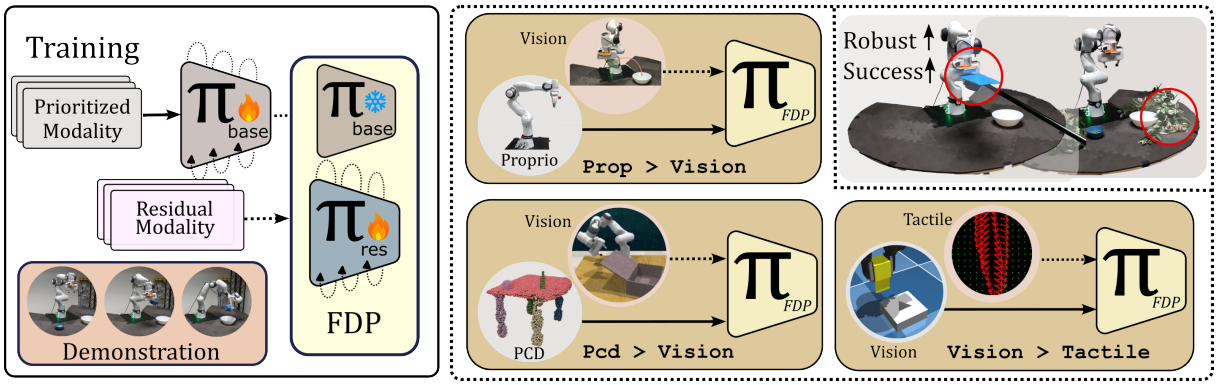


Fig. 1. Policy learning using FDP with different prioritization orders. In FDP, we train a base and a residual policy by prioritizing over different observation modalities for the same task. We demonstrate this approach with different combinations of observation modalities such as `prop>vision`, and `vision>tactile` among others. FDP results in more performant policies (15% ↑) that are robust to distractors and camera occlusions (40% ↑).

simulated benchmarks such as RLBench (15% ↑), Adroit (10% ↑), Robomimic (10% ↑) and M3L (20% ↑). We thoroughly analyze our method and present ablations.

- Our real-world experiments demonstrate the usefulness of FDP in learning policies that are robust to visual distribution shifts (40% ↑). Policies learned using the prioritization order of `prop>vision` were not only **robust to distractors but also to camera occlusions** (5× ↑), where diffusion policies failed catastrophically.

II. Related Work

Sample complexity and Generalization. Despite recent scaling efforts [5], [6], the collection of multimodal data is difficult in robotics and the number of variations of tasks is unbounded. Several works leverage compositionality for solving novel combinations of tasks with existing solutions, such as composing distributions across heterogeneous modalities for tool use [7] and sequencing skills for long horizon problems [8]. The most relevant to our work is PoCo [7], that composes single task policies conditioned on different modalities. However, PoCo composes pre-learned policies for the same task and requires manual tuning of the compositional weights. Instead FDP learns the residual to be composed with the base prioritized policy, using the same data and requiring no manual tuning. Augmentation [9] or retrieval-based [10] approaches of addressing sample-complexity add a substantial computational and data overhead and are orthogonal to our proposed algorithmic improvement.

Residual Learning and Adapters. Residual reinforcement learning has been used to improve the performance of behavior cloning policies through interaction with the environment [11]. Jiang et al. [12] show sim-to-real transfer by learning a supervised residual for human feedback on real world rollouts of policies learned in simulation, maximizing the likelihood of the correction applied. In FDP, the effect of less-influential modalities is captured by learning a residual over a policy trained on prioritized modalities. We theoretically derive this residual within the framework of diffusion and score-based models. Our work is also similar to Q-Adapter [13] in terms of learning the residual using an adapter, but does not necessitate a base foundation model.

Several works in robotics have used adapters such as LoRA [14] and ControlNet [15] for fine-tuning multi-task or foundation-models [5], [6] on downstream tasks. Diff-Control from Liu et al. [16] learns a ControlNet over a diffusion policy base to impart stateful behavior to the policy. Interestingly, FDP can be leveraged to reach a similar learning formulation as Diff-Control, with the residual learned for the previous action chunk instead of a modality.

III. Background

Diffusion Models. Gaussian diffusion models [17] learn the reverse diffusion kernel $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ for a fixed forward kernel that adds Gaussian noise at each step $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathcal{I})$, such that $p(\boldsymbol{x}_T) \approx \mathcal{N}(0, \mathcal{I})$. Here, $t <= T$ is the diffusion time step and α_t is the noise schedule. In practice, the models learn a reparametrized form corresponding to the noise added to the input $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ [3].

Score-based Models. Song et al. [18] presented a unified framework showing that both diffusion models [3], [17] and score-based models [19] can be interpreted as discretizations of different forward stochastic differential equations (SDEs). The latter learn the score $\nabla_{\boldsymbol{x}_t} \log p_{\sigma_t}(\boldsymbol{x}_t)$ at different noise scales σ_t required for sampling from the data distribution. Explicit Score Matching (ESM) [20], [21] was proposed to estimate the score by minimizing the Fisher divergence with the Gaussian-smoothed data distribution $p_{\sigma_t}(\boldsymbol{x}_t) = \int p(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}, \sigma_t^2 I)d\boldsymbol{x}$. Denoising Score Matching (DSM) alleviates the computational difficulties of ESM [21], [22], and is shown in Equation 1, where $s_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \sigma_t)$, abbreviated to $s_{\boldsymbol{\theta}}(\boldsymbol{x}_t)$, represents the learned score model.

$$\begin{aligned}\mathcal{J}_{\sigma_t}(\boldsymbol{\theta}) &\overset{\text{ESM}}{=} \mathbb{E}_{p_{\sigma_t}(\boldsymbol{x}_t)}\left[\frac{1}{2}\left\|\nabla_{\boldsymbol{x}_t}\log p_{\sigma_t}(\boldsymbol{x}_t) - s_{\boldsymbol{\theta}}(\boldsymbol{x}_t)\right\|_2^2\right] \\ &\overset{\text{DSM}}{=} \mathbb{E}_{p_{\sigma_t}(\boldsymbol{x},\boldsymbol{x}_t)}\left[\frac{1}{2}\left\|\nabla_{\boldsymbol{x}_t}\log p_{\sigma_t}(\boldsymbol{x}_t|\boldsymbol{x}) - s_{\boldsymbol{\theta}}(\boldsymbol{x}_t)\right\|_2^2\right] + C\end{aligned} \tag{1}$$

Diffusion models use a $(1-\bar{\alpha}_t)$ weighted DSM objective along with a forward transition kernel $p_{\bar{\alpha}_t}(\boldsymbol{x}_t|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}, (1-\bar{\alpha}_t)I)$ with discrete time

and $\bar{\alpha_i}=\prod_{j=1}^{i}\alpha_j$, yielding the simplified diffusion loss from Ho et al. [3]. Score-based model typically use $\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{x},\sigma_t^2 I)$, where α_t and σ_t are respective noise scales. In the simplified case, an optimal diffusion model is related to the score of the α_t-diffused data distribution by $-\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)/\sqrt{1-\bar{\alpha_t}} \overset{def}{=} \boldsymbol{s}^*_{\boldsymbol{\theta}}(\boldsymbol{x}_t)=\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)$ [18]. Typically, diffusion models generate samples via progressive denoising through the reverse diffusion process [3], while score-matching models sample from the data distribution using Langevin dynamics [23].

Classifier Guided Sampling. Dhariwal and Nichol [24] obtain conditional samples from an unconditional diffusion model trained on $\boldsymbol{x}$ using Bayes' theorem. We can sample from a class $\boldsymbol{y}$ by decomposing the conditional score at time step t into the unconditional score and the classifier gradient $\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t|\boldsymbol{y}) = \nabla_{x_t}\log p(\boldsymbol{x}_t;\boldsymbol{\theta}) + \nabla_{x_t}\log p(\boldsymbol{y}|\boldsymbol{x}_t;\boldsymbol{\phi})$. However, classifier guidance needs an explicit classifier trained on noisy samples to estimate the gradients [25].

IV. Methodology

Assume that we have robot demonstrations $D=\{(\boldsymbol{x},\boldsymbol{y})_i\}$ where $i = 1..N$, consisting of actions $\boldsymbol{x}$ and different observational modalities $\boldsymbol{y}^{1:M}$, such as images, point clouds, proprioception or tactile. Let $\boldsymbol{y}^{1:k}$ be the prioritized observational modalities of the M total modalities, where $\boldsymbol{y}^{1:k}\equiv\boldsymbol{y}^1,..,\boldsymbol{y}^k$ and $k<M$. To sample actions $\boldsymbol{x}$ conditioned on all $\boldsymbol{y}^{1:M}$, we need to estimate the score $\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t|\boldsymbol{y}^{1:M})$. Taking a cue from classifier guidance [24], we decouple the observational modalities utilizing Bayes' theorem to obtain Equation 2. For scores to be valid, observational modalities $\boldsymbol{y}^{1:M}$ can be noised with a Gaussian kernel $\mathcal{N}(\tilde{\boldsymbol{y}};\boldsymbol{y},\tau^2 I)$ of variance τ^2 that is small enough such that $p_\tau(\tilde{\boldsymbol{y}}) \approx p(\boldsymbol{y})$, where we drop the notation τ going forward. Actions $\boldsymbol{x}$ are noised with the kernel $p_{\bar{\alpha}_t}(\boldsymbol{x}_t|\boldsymbol{x})$.

$$\begin{aligned}&\boldsymbol{s}^*(\boldsymbol{x}_t,\boldsymbol{y}^{1:M}) = \nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t|\boldsymbol{y}^{1:M})\\&= \nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t|\boldsymbol{y}^{1:k}) + \nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{y}^{k+1:M}|\boldsymbol{x}_t,\boldsymbol{y}^{1:k})\end{aligned} \tag{2}$$

The first score term $\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t|\boldsymbol{y}^{1:k})$ corresponds to a policy $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:k})$ that would be obtained on training with just the modalities $\boldsymbol{y}^{1:k}$, referred to as π_{base} going forward. The effect of the other modalities is captured in the second score term $\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{y}^{k+1:M}|\boldsymbol{x}_t,\boldsymbol{y}^{1:k})$. However, explicitly training a classifier $p(\boldsymbol{y}^{k+1:M}|\boldsymbol{x}_t,\boldsymbol{y}^{1:k})$ as suggested by Dhariwal and Nichol [24] is impractical due to the high dimensionality and continuity of multiple observational modalities $\boldsymbol{y}^{1:M}$, such as images and tactile. Hence, we employ explicit score matching $\mathcal{J}_{\alpha_t}(\boldsymbol{\phi})$ [20], [21] as shown in Equation 3.

$$\mathbb{E}_{p_{\alpha_t}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})}\left[\frac{1}{2}\left\|\begin{array}{l}\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{y}^{k+1:M}|\boldsymbol{x}_t,\boldsymbol{y}^{1:k})\\-\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})\end{array}\right\|_2^2\right] \tag{3}$$

Due to the high computational complexity of estimating the empirical scores such as $\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{y}^{k+1:M}|\boldsymbol{x}_t,\boldsymbol{y}^{1:k})$, Chao et al. [22] derive the denoising likelihood score matching (DLSM) objective for conditional distributions, which forms the basis for our next result.

Theorem 1: Explicit score matching for $\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})$ in Equation 3 is equivalent to the objective $\mathcal{J}^{res}_{\alpha_t}(\boldsymbol{\phi})$:

$$\mathbb{E}_{\substack{p_{\alpha_t}(\boldsymbol{x},\boldsymbol{x}_t,\\ \boldsymbol{y}^{1:M})}}\left[\frac{1}{2}\left\|\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{x})- \begin{array}{l}\boldsymbol{s}^*(\boldsymbol{x}_t,\boldsymbol{y}^{1:k})\\-\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})\end{array}\right\|_2^2\right]$$

Here $\boldsymbol{s}^*(\boldsymbol{x}_t,\boldsymbol{y}^{1:k})$ is the the frozen optimal score model for $\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t|\boldsymbol{y}^{1:k})$, approximated using a learned model $\pi_{\text{base}}:\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:k})$ in practice.

We believe FDP to be the first work to prove this equivalence for an arbitrary number of conditionals and directly learn the classifier guidance in a high-dimensional setting for the purposes of policy learning. Interestingly, comparison of $\mathcal{J}^{res}_{\alpha_t}(\boldsymbol{\phi})$ with DSM shown in Equation 1 reveals that the effect of $\boldsymbol{y}^{k+1:M}$ captured through $\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})$ can be learned as a residual to $\pi_{\text{base}}:\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:k})$, the policy learned with just the modalities $\boldsymbol{y}^{1:k}$. Thus we refer to $\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})$ as π_{res}. **Essentially, by factorizing the score of the full-conditional action, FDP learns π_{base} using $\boldsymbol{y}^{1:k}$, and then learns π_{res} using $\boldsymbol{y}^{1:M}$ and a frozen π_{base}.** This two-phase training prioritizes modalities $\boldsymbol{y}^{1:k}$ over $\boldsymbol{y}^{k+1:M}$.

A concise proof of Theorem 1 is presented, and a detailed version can be found on our website. We substitute $\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})$ as $\bullet$ for brevity. The inner product obtained by opening the square in Equation 3 can be simplified as-

$$\begin{aligned}&\mathbb{E}_{p_\alpha(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})}\left[\left\langle\bullet,\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{y}^{k+1:M}|\boldsymbol{x}_t,\boldsymbol{y}^{1:k})\right\rangle\right] =\\&\mathbb{E}_{\substack{p_{\alpha}(\boldsymbol{x}_t,\\ \boldsymbol{y}^{1:M})}}\left[\left\langle\bullet,\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{y}^{1:M})-\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{y}^{1:k})\right\rangle\right]\end{aligned} \tag{5}$$

Further simplifying the inner product with the first term-

$$\begin{aligned}&\mathbb{E}_{p_{\alpha_t}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M})}\left[\left\langle\boldsymbol{s}_{\boldsymbol{\phi}}(\boldsymbol{x}_t,\boldsymbol{y}^{1:M}),\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{y}^{1:M})\right\rangle\right]\\&=\mathbb{E}_{p(\boldsymbol{y}^{1:M})}\int_{\boldsymbol{x}_t}p_\alpha(\boldsymbol{x}_t|\boldsymbol{y}^{1:M})\langle\bullet,\frac{\nabla_{\boldsymbol{x}_t}p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{y}^{1:M})}{p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{y}^{1:M})}\rangle d\boldsymbol{x}_t\\&=\mathbb{E}_{p(\boldsymbol{y}^{1:M})}\int_{\boldsymbol{x}_t}\langle\bullet,\nabla_{\boldsymbol{x}_t}\int_{\boldsymbol{x}_0}p_0(\boldsymbol{x}_0|\boldsymbol{y}^{1:M})p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{x}_0,\boldsymbol{y}^{1:M})d\boldsymbol{x}_0\rangle d\boldsymbol{x}_t\\&=\mathbb{E}_{p_{\alpha_t}(\boldsymbol{x}_0,\boldsymbol{x}_t,\boldsymbol{y}^{1:M})}\left[\left\langle\bullet,\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{x}_0)\right\rangle\right]\end{aligned} \tag{6}$$

Note that $\mathbb{E}_{\substack{p_{\alpha_t}(\boldsymbol{x}_t,\\ \boldsymbol{y}^{1:M})}}\left[\frac{1}{2}||\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{y}^{k+1:M}|\boldsymbol{x}_t,\boldsymbol{y}^{1:k})||_2^2\right]$ is a constant. Substituting results obtained in Equations 5 and 6 back in Equation 3, and adding the constant term $\mathbb{E}_{\substack{p_{\alpha_t}(\boldsymbol{x}_t,\\ \boldsymbol{y}^{1:k})}}\left[\frac{1}{2}||\nabla_{\boldsymbol{x}_t}\log p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{y}^{1:k})-\nabla_{\boldsymbol{x}_t}\log p_\alpha(\boldsymbol{x}_t|\boldsymbol{x}_0)||_2^2\right]$ we complete the proof for Theorem 1. Hence, we prove that the ESM $\mathcal{J}_{\alpha_t}(\boldsymbol{\phi})$ in Equation 3 is equivalent to minimizing the objective $\mathcal{J}^{res}_{\alpha_t}(\boldsymbol{\phi})$ presented in Theorem 1, differing up to a constant.

Theorem 1 implies that the effect of de-prioritized modalities $\boldsymbol{y}^{k+1:M}$ can be learned as a residual over the prioritized modalities $\boldsymbol{y}^{1:k}$. From a different lens, π_{res} effectively learns the classifier guidance required to sample from π_{base} further conditioned on $\boldsymbol{y}^{k+1:M}$. **Learning π_{res} as a residual of π_{base}**

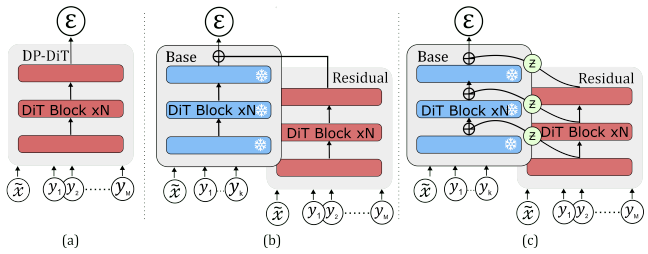


Fig. 2. Architectural representations for [a] diffusion policy that jointly conditions on all observational modalities, [b] straightforward composition of the score outputs from π_{base} and π_{res} and [c] FDP architecture with block-wise composition with a layer $\mathcal{Z}$ applied on π_{res}.

ensures that the policy does not overfit modalities $\boldsymbol{y}^{k+1:M}$, but only learns correlations to bridge the error arising from π_{base} trained on the prioritized modalities $\boldsymbol{y}^{1:k}$. Hence, policies learned in this factorized way are naturally robust to distribution shifts in $\boldsymbol{y}^{k+1:M}$. Moreover, explicit prioritization of $y^{1:k}$ by training π_{base} prior to learning the residual leads to sample efficiency, as the model learns correlations with the stronger modality without having to attend to other modalities.

Factorizing Diffusion Policies. As developed in Section III, Theorem 1 applies to diffusion models. Resolving $\nabla_{\boldsymbol{x}_t} \log p_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{x})$ to $-\boldsymbol{\epsilon}_0/\sqrt{1-\bar{\alpha}_t}$ and replacing $\boldsymbol{s}^*_{\boldsymbol{\theta}}(\boldsymbol{x}_t)$ with $-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)/\sqrt{1-\bar{\alpha}_t}$, we get the simplified diffusion losses for π_{base} and π_{res}.

$$\mathcal{L}^t_{base}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x}_0,\boldsymbol{y}^{1:k})\mathcal{N}(\boldsymbol{\epsilon}_0;0,\mathcal{I})}\left[\left\|\boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:k}, t)\right\|^2_2\right]$$
$$\mathcal{L}^t_{res}(\boldsymbol{\phi}) = \mathbb{E}_{\substack{\boldsymbol{x}_0,\boldsymbol{y}^{1:M}\sim p \\ \boldsymbol{\epsilon}_0\sim\mathcal{N}(0,\mathcal{I})}}\left[\frac{1}{2}\left\| \begin{array}{l} \boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:k}, t) \\ -\boldsymbol{\epsilon}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:M}, t) \end{array} \right\|^2_2\right] \tag{7}$$

From the perspective of diffusion models, π_{base} maximizes a reweighted lower bound on the data likelihood only considering the prioritized k modalities, while π_{res} learns a residual over π_{base} to maximize it for demonstration data with all the modalities included, thus learning their residual effect. Since diffusion models are trained on discrete time steps, π_{res} is learned on the same time discretization as used for π_{base}. Once trained, actions can be sampled from the conditional distribution $p(\boldsymbol{x}|\boldsymbol{y}^{1:M})$ using reverse diffusion [3] on the composition [26] of π_{base} and π_{res}:

$$\boldsymbol{x}_{t-1}\sim\mathcal{N}\Big(\boldsymbol{x}_t; \tfrac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \tfrac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}(\boldsymbol{x}_t, \boldsymbol{y}^{1:M}, t)\right), \sqrt{1-\alpha_t}\,\mathcal{I}\Big)$$
$$\boldsymbol{\epsilon}(\boldsymbol{x}_t, \boldsymbol{y}^{1:M}, t) = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:k}, t) + \boldsymbol{\epsilon}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:M}, t) \tag{8}$$

Architectural Implementation of FDP. The models π_{base} and π_{res} can be instantiated using standard architectures such as UNet [27] or DiT [28]. Inspired from the late stage score-composition, we propose a more integrated way to compose π_{base} and π_{res}, as shown in Figure 2. Instead of learning a residual for the final score output, π_{res} learns the blockwise residual over the intermediate outputs of the frozen π_{base}. Specifically, let $\mathcal{F}^i_{\text{base}}$ and $\mathcal{F}^i_{\text{res}}$ denote the i-th DiT block outputs of the base and residual models, respectively. Then the composed output at level i can be written as $\mathcal{F}^i_{base}(\boldsymbol{x}', \boldsymbol{y}'^{1:k}) + \mathcal{Z}(\mathcal{F}^i_{res}(\boldsymbol{x}', \boldsymbol{y}'^{1:M}))$, where x' and $y'^{1:M}$ are layer inputs. Similar to Zhang et al. [15], $\mathcal{Z}$ is a zero-initialized layer to avoid harmful updates at the start of the training and to ensure that gradient updates to the residual model improve the predictions of the composed model over π_{base}. This architecture enables a simplified training objective from Ho et al. [3] for the residual model. Our residual model is structured following the Vision Transformers architecture [29]. In π_{res}, all observational modalities are passed through self-attention layers after encoding into a single embedding.

All transformer-based models are trained over 2000 epochs with a batch size of 64 for visual tasks and for 3000 epochs with a batch size of 256 for low-dimensional tasks. All models except VLAs are trained on NVIDIA A5000 GPUs, with training times ranging from 6-12 hours depending on model size and the number of camera inputs. Model `prop` π_{base} consists of $\sim 30M$ parameters, while `vis` π_{res} with two camera image inputs is $\sim 55M$ parameters large. Our current implementations support an action prediction latency of $\sim$50ms for transformer-based diffusion policy baseline, $\sim$100ms for UNet [4] and output composition of models as shown in Figure 2 $[b]$, and $\sim$150ms for FDP model in $[c]$.

V. Simulation Experiments

We train and evaluate FDP and related baselines on 14 tasks from RLBench [30] along with their distractor variants, 4 tasks from Adroit [31], 4 tasks from Robomimic [32], and the visuo-tactile insertion task from M3L [33]. RLBench provides a diverse suite of visuomotor manipulation tasks with joint positions as the action space and a built-in planner for demonstration collection. Within RLBench, we evaluate prioritization orders `prop`$\lesssim$`vision` and `pcd`$\lesssim$`vision` for FDP. The Adroit benchmark contains high-dimensional hand manipulation tasks performed with a 24-DoF anthropomorphic hand. Environments in Robomimic use an action representation defined by changes in end-effector position and orientation (axis–angle). For both Adroit and Robomimic, we explore the prioritization orders `prop`$\lesssim$`env-state`. Finally, the M3L environment requires precise insertion of differently shaped pegs into holes randomly placed on a surface, using a single RGB camera and two tactile sensors for perception and Δxyz as the action representation. Demonstrations for M3L are collected using an expert RL policy proposed by Sferrazza et al. [33]. On M3L we evaluate the performance of `vision`$\lesssim$`tactile` prioritizations using FDP against jointly conditioned diffusion policy. Unless otherwise noted, reported results are averaged over 300 rollouts. Additional experimental details are provided on our project webpage. `https://fdp-policy.github.io/fdp-policy/`.

Baselines. For evaluation of sample efficiency in visuomotor tasks, we compare against several approaches that differ in the way in which they model generative policy learning. However, for all approaches, we choose DiT-small ($\sim$90M) [28] as our model architecture. We implement Diffusion Policy [4] using DiT, referred to as DP-DiT in our results. For comparison, we also include UNet [27]

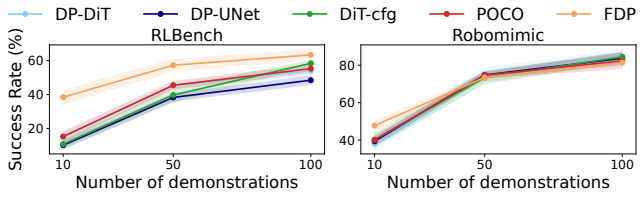


Fig. 3. Mean (with Std. Dev.) Performance for all models across 10-50-100 demonstrations on RLBench and Robomimic. Prioritization of modalities using FDP enables strong performance gains, especially in low-data regimes.

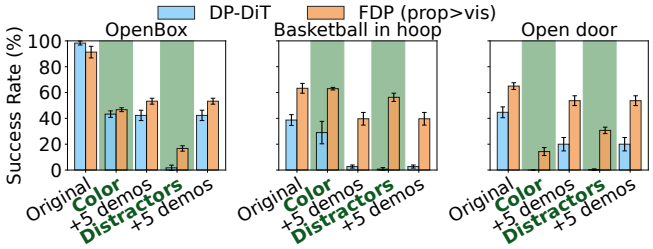


Fig. 4. Performance of FDP (`prop>vision`) and DP-DiT across original, color-variant and distractor environments. We also fine-tune on 5 additional demos collected in the modified environments. Note the strong performance of FDP (`prop>vision`) in the color-variant and distractor experiments.

implemented by Chi et al. [4] in our baselines as DP-UNet. We reformulate POCO [7] to compose observational modalities. We train π_{base} and π_{res} models independently, prior to sampling from the composed distribution [26] using $\boldsymbol{\epsilon}(\boldsymbol{x}_t, \boldsymbol{y}^{1:M}, t) = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:k}, t) + \lambda * \boldsymbol{\epsilon}_{\boldsymbol{\phi}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:M}, t)$. Here, $\lambda = 0.1$ based on POCO's ablations [7]. We also report results for classifier-free guidance [25] as CFG, where we train a single model and switch out the weaker modality with a probability of 0.2. We then sample using $\boldsymbol{\epsilon}(\boldsymbol{x}_t, \boldsymbol{y}^{1:M}, t) = \lambda_1 * \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:k}, \boldsymbol{y}^{k+1:M}, t) + \lambda_2 * \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}^{1:k}, \boldsymbol{\phi}, t)$, where we set $\lambda_1 = 1.1$ and $\lambda_2 = 0.1$, as suggested by [25]. We also fine-tune a 450M parameter vision-language action model SmolVLA [6] for at least 40k steps, evaluating its sample-efficiency and robustness on selected RLBench tasks. For real-world and distractor experiments in simulation, we compare against DP-DiT.

A. Sample Efficiency Gain Through Modality Prioritization.

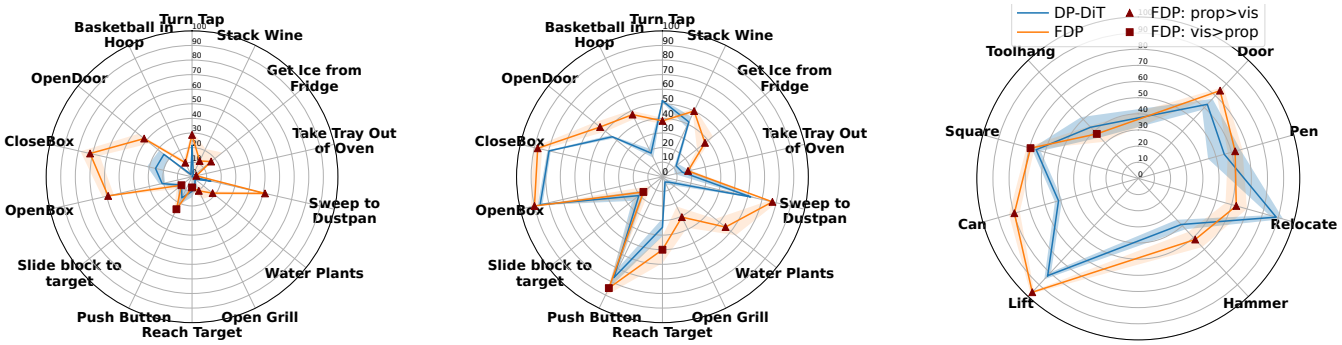
FDP enables us to include the prioritization order of modalities for tasks as a hyperparameter. Prioritizing either `vision`, `env-state`, or `prop` modalities using **FDP consistently outperforms joint conditioning across a wide range of visuomotor, state-based and tactile tasks** in RLBench, Robomimic, Adroit (Fig. 5) and M3L (Table 7) respectively. In RLBench, modality prioritization with FDP achieves on average a 15% higher success rate with 10 and 50 demonstrations, and a 10% higher success rate with 100 demonstrations, compared to the strongest baseline. We also observe clear gains with the prioritization order of `prop>env-state` in Robomimic and Adroit tasks. Fig. 3 illustrates how performance scales with increasing demonstrations in RLBench and Robomimic. Prioritization is **especially beneficial in low-data regimes**, where a jointly conditioned model lacks sufficient data to learn the correct modality weighting. FDP enforces conditioning on the most essential modality, leading to stronger overall performance. SmolVLA fails drastically at `SweepToDustpan` (Figure 6) that requires precise spatial motions to sweep all the dirt particles into the dustpan. This highlights that there is potential for VLAs to improve beyond table-top tasks where algorithmic approaches like FDP do better. For M3L, we observe in Table 7 that the prioritization order of `vision>tactile` outperforms the joint-conditioning approach by over 20% at 100 and 200 demonstrations. Several works [34], [35] have fine-tuned a BC policy using reinforcement learning, and FDP can serve as a performant prior policy for further fine-tuning. These results clearly show that FDP leads to more performant policies across various observational modalities, especially in low-data regimes.

B. Robustness Gain Through Modality Prioritization.

Prioritization prevents the model from developing spurious correlations by learning a residual policy for modalities with limited influence on robot actions. To test this, we evaluate FDP (`prop>vision`) against a jointly conditioned diffusion policy in environments with color variations and clutter. Both DP-DiT and FDP are trained on 100 demonstrations collected in the original environment and evaluated in three settings: the original, color-variant, and cluttered environments as shown in Figure 8. **FDP significantly outperforms DP-DiT in both color-variant and cluttered settings by more than** 40%. We further collect five demonstrations in each modified environment to study few-shot adaptation to out-of-distribution data. FDP adapts more effectively, improving its performance by 15% on average compared to 10% for DP-DiT. This is achieved by updating only the residual model π_{res} with new demonstrations, which adjusts the conditional distribution on visual modalities $p(y^{vis} \mid x, y^{prop})$ without modifying the full conditional action distribution $p(x \mid y^{prop}, y^{vis})$. We extend this setting to point clouds (`pcd>vision`), where FDP learns a `vision` π_{res} over DP3 used as π_{base} [36], and compare it to DP3 with RGB inputs. Point clouds are sample-efficient for policy learning since they encode scene geometry in a single modality [36]. However, our distractor experiments show that FDP with a visual residual over DP3 achieves $\sim 20\%$ higher performance than DP3 using RGB inputs. These results clearly outline the robustness gain on adopting FDP as the policy formulation.

VI. Effects of Prioritization

Intuition on the Order of Modality Prioritization As for most hyperparameters, intuition can be developed for the order of modality prioritization. To experimentally test this, we develop 3 variants of the block pick environments- `S`: $0.15 \times 0.2m$, `M`: $0.35 \times 0.45m$ and `L`: $0.55 \times 0.75m$, with increasing range of generalization in initial block-placement. The results are presented in Table I. We observe that prioritization of proprioception out-performs all other models for the `S` environment while prioritization of vision tends to do better for larger `L` environment. This also



(a) RLBench tasks with 10 demonstrations (b) RLBench tasks with 50 demonstrations (c) Robomimic & Adroit with 10-50 demos

Fig. 5. Radial plot showing performance of FDP in comparison to DP-DiT. For FDP, the best results obtained using ■ `vis>prop` and ▲ `prop>vision` are marked on the plots. These plots show that searching through the modality prioritization space yields improvements in a wide-spectrum of tasks.

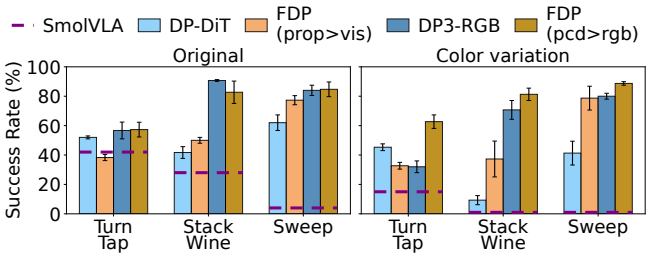


Fig. 6. Evaluation of robustness gained using `prop|pcd>vision`. Notably, we see a significant drop in performance of DP-DiT and DP3 on introducing color variations in the task. Unlike FDP, SmoLVLA despite having a strong VLM backbone, fails to adapt to color variations across tasks.

Task	Demos	DP-DiT	vision> tactile	tactile> vision
square peg	100	22	**48**	28
	200	52	**72**	24
triangle peg	100	14	**42**	10
	200	28	**50**	16
All pegs	200	6	**26**	4

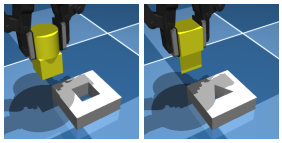


Fig. 7. Results from the visuo-tactile insertion tasks from M3L [33]. We see a clear benefit with the prioritization order of `vision>tactile`. The vision modality plays a crucial role in navigating the peg towards the hole [34]. Tactile input becomes useful once the peg is in contact with the contour of the hole, and is learned as a residual over vision π_{base}.

conforms to the intuition of vision playing a diminished role when object placement area is smaller and the motions are repetitive, and a more significant role when the motion varies significantly based on the placement of the object. Tasks that correlate heavily with robot proprioception are not uncommon as the robot is solving them in the first person view, and can move close to the object if required. Our results in the M3L visuo-tactile environment also conform with `localize-then-execute` strategy [34], where the visual modality plays a more important role in localizing the hole for peg insertion. Learning a `tactile` π_{res} over the `vision` π_{base} learns residual scores for states where tactile influences the actions of the robot. Finally, the chosen action representation also plays a role, with FDP (`prop>`) realizing higher success rate for joint-state actions.

Ablations. Table II presents ablations for our method. Results for DiT-Base show that lack of performance cannot be compensated for by increasing the model size. We also show that the integrated form of composition presented in Figure 2[c] outperforms the output score composition method [b] by a significant margin. Further, we find that preserving the diversity of π_{base} is essential: overfitting the base model leaves little residual signal to learn, reducing generalization, while stopping the training too early leaves an unstable base. Our ablations show that selecting the π_{base} checkpoint with the lowest validation loss (at 700 epochs) provides a good foundation for residual learning. Finally, the results in Table I show that learning the residual is crucial. Policy performance with π_{base} is unsatisfactory, and **our factored approach is able to improve the performance without diminishing the policy robustness given an additional modality**.

TABLE I

BLOCK PICK SUCCESS RATES FOR 10 AND 50 DEMONSTRATIONS.

10 demos	**S**	**M**	**L**
DP-DiT	29.7 ± 3.1	12.0 ± 1.0	3.3 ± 0.6
FDP (prop > vision)	**73.7** ± **3.8**	**21.3** ± **3.5**	6.3 ± 3.1
FDP (vision > prop)	18.7 ± 2.3	6.7 ± 1.2	3.3 ± 3.1
50 demos	**S**	**M**	**L**
DP-DiT	95.3 ± 3.2	**69.0** ± **7.0**	45.7 ± 7.1
FDP (prop > vision)	**98.7** ± **1.5**	55.0 ± 2.6	20.3 ± 3.5
FDP (vision > prop)	96.7 ± 1.2	**65.3** ± **8.1**	**60.0** ± **2.0**
prop π_{base}	47.3 ± 2.5	5.0 ± 1.7	1.3 ± 1.2
vision π_{base}	96.0 ± 2.0	**68.0** ± **6.0**	51.3 ± 8.1

TABLE II

ABLATION RESULTS ON THE OPEN DOOR TASK (10 DEMONSTRATIONS).

Model	**Succ. (%)**
DiT: small (∼33M)	24.0 ± 7.2
DiT: base (∼130M)	27.3 ± 5.0
Score Comp.: [b] Fig. 2	20.7 ± 8.3
FDP [c]: Conv	**42.0** ± **5.2**

π_{base} **ep**	**Succ. (%)**
100 ep	24.7 ± 6.1
700 ep	**42.0** ± **5.2**
1500 ep	40.0 ± 6.0
2000 ep	40.7 ± 3.1

VII. REAL-WORLD EXPERIMENTS

We evaluate FDP and the DP-DiT baseline across four real-world domains and report their task success rates. The

Fig. 8. We modify the original RLBench environments to introduce color variations and add distractors.

domains are – *Close Drawer* as a simple task where the robot has to push the drawer; *Put Block in Bowl* that assesses the policy's ability to perform precise pick-and-place actions; *Pour in Bowl* to evaluate the policy's dexterity in operating near joint limits and *Fold Towel* to assess effectiveness in manipulating deformable objects.

We collect 50 demonstrations per domain on a Franka FR3 robot using a 6D space mouse, recording both proprioceptive and visual observations from two cameras—one mounted on the gripper and a static camera covering the workspace. The trained policies are evaluated on four task variations in each domain: (a) `default`: an in-distribution setup matching the conditions used during demonstration collection; (b) `color`: the object's color is altered to test robustness to visual appearance changes; (c) `distractor`: novel, unseen objects such as vegetation props and soft toys are added to the scene to introduce clutter; and (d) `occlusion`: visual input is intermittently blocked during policy rollout to simulate partial observability. Figure 9 shows different task domains and their variations used in our experiments. We use 10 rollouts in each experiment and report the task success rate as shown in Table III.

TABLE III
SUCCESS RATES OF DP-DIT AND FDP ACROSS REAL-WORLD TASKS.

Task Domain	`default`		`color`		`dist.`		`occl.`	
	DP	**FDP**	**DP**	**FDP**	**DP**	**FDP**	**DP**	**FDP**
Close Drawer	**90**	**90**	**90**	**90**	10	**80**	0	**80**
Put Block in Bowl	**80**	**80**	0	**60**	0	**60**	10	**60**
Pour in Bowl	70	**80**	40	**80**	20	**60**	10	**50**
Fold Towel	40	**60**	40	**70**	30	**70**	10	**50**

Result Analysis. We find that FDP is robust to distribution shifts in the environment. DP-DiT regularly produces unachievable robot actions under `distractor` and `occlusion` settings, often triggering safety stops, resulting in task failure. In contrast, FDP guided by its `prop` π_{base}, consistently generates stable actions even under severe occlusions and cluttered scenes, yielding an average absolute performance improvement of 40%. In the `default` experiment we observe that the FDP policy outperforms DP-DiT in the pouring and towel-folding tasks, which require precise object manipulation. **Notably, FDP achieves $5\times$ success-rate than that of DP-DiT in the camera occlusion setting, highlighting it's practicality for robots that must operate reliably in visually degraded environments.**

VIII. Conclusion and Future Work

We present Factorized Diffusion Policies (FDP), a novel policy formulation that factorizes the joint conditioning in diffusion models so that observational modalities can have a differing influence on the action diffusion process by design. We derive a novel loss function to realize the prioritization order of modalities and propose a novel architecture for efficient training. Through extensive experiments across visual, point-cloud, proprioceptive and tactile modalities, we demonstrate several benefits of modality prioritization, including improved sample efficiency and increased robustness. Overall, we observe 15% absolute performance improvement on more than 20 tasks spread across several observational modalities after adopting FDP over jointly conditioned diffusion policy and even SmolVLA. **We believe that this novel paradigm of modality prioritization along with strong performance gains, especially in low-data regimes make FDP a valuable contribution to the robot learning community**. Finally, our real-world experiments highlight the practical value of FDP in being a safe-to-deploy policy in the face of visual disturbances and even *camera occlusions*, outperforming diffusion policies by over 40%.

FDP opens new avenues for future research, such as scalable integration of diversely sourced observational modality data for robot policy learning. FDP presents a computational overhead of searching through the prioritization order, and future work can develop a stronger framework that infers the modality to prioritize based on the task or the collected data. FDP prioritizes a single modality for the whole duration of the task, and dynamic prioritization of modalities also presents an interesting avenue for future work. Finally, we believe our approach will also have implications for fine-tuning of VLAs on new modalities not encountered in training. We believe FDP is a novel first-step towards the goal of observation modality prioritization.

IX. Acknowledgment

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0239.

References

[1] B. Wahn and P. König, "Is attentional resource allocation across sensory modalities task-dependent?" *Advances in cognitive psychology*, vol. 13, no. 1, p. 83, 2017.

[2] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv Preprint 2006.11239*, 2020.

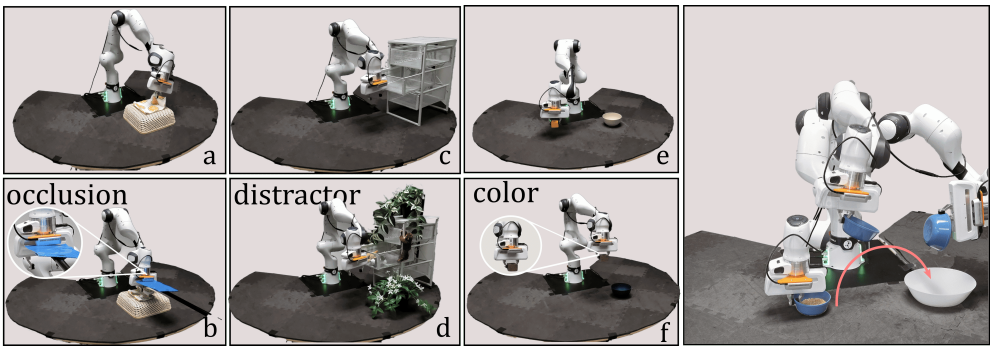


Fig. 9. Real robot task domains and their variations. We show that DP-DiT fails in the presence of visual distribution shifts, highlighting the utility of prioritization for tasks with repetitive motions. Furthermore, FDP (`prop>vision`) is robust to camera blinds which can cause safety risks for DP-DiT.

[4] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2023.

[5] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[6] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[7] L. Wang, J. Zhao, Y. Du, E. H. Adelson, and R. Tedrake, "Poco: Policy composition from and for heterogeneous robot learning," *arXiv preprint arXiv:2402.02511*, 2024.

[8] U. A. Mishra, Y. Chen, and D. Xu, "Generative factor chaining: Coordinated manipulation with diffusion-based factor graph," in *ICRA 2024 Workshop- Back to the Future: Robot Learning Going Probabilistic*, 2024.

[9] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter *et al.*, "Scaling robot learning with semantically imagined experience," *arXiv preprint arXiv:2302.11550*, 2023.

[10] L.-H. Lin, Y. Cui, A. Xie, T. Hua, and D. Sadigh, "Flowretrieval: Flow-guided data retrieval for few-shot imitation learning," *arXiv preprint arXiv:2408.16944*, 2024.

[11] M. Alakuijala, G. Dulac-Arnold, J. Mairal, J. Ponce, and C. Schmid, "Residual reinforcement learning from demonstrations," *arXiv preprint arXiv:2106.08050*, 2021.

[12] Y. Jiang, C. Wang, R. Zhang, J. Wu, and L. Fei-Fei, "Transic: Sim-to-real policy transfer by learning from online correction," *arXiv preprint arXiv:2405.10315*, 2024.

[13] Y.-C. Li, F. Zhang, W. Qiu, L. Yuan, C. Jia, Z. Zhang, Y. Yu, and B. An, "Q-adapter: Customizing pre-trained llms to new preferences with forgetting mitigation," *arXiv Preprint 2407.03856*, 2025.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[15] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv Preprint 2302.05543*, 2023.

[16] X. Liu, Y. Zhou, F. Weigend, S. Sonawani, S. Ikemoto, and H. B. Amor, "Diff-control: A stateful diffusion-based policy for imitation learning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7453–7460.

[17] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *arXiv Preprint 1503.03585*, 2015.

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[19] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," 2020.

[20] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[21] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[22] C.-H. Chao, W.-F. Sun, B.-W. Cheng, Y.-C. Lo, C.-C. Chang, Y.-L. Liu, Y.-L. Chang, C.-P. Chen, and C.-Y. Lee, "Denoising likelihood score matching for conditional score-based data generation," *arXiv Preprint 2203.14206*, 2022.

[23] G. O. Roberts and R. L. Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, pp. 341–363, 1996.

[24] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[25] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[26] Y. Du, C. Durkan, R. Strudel, J. B. Tenenbaum, S. Dieleman, R. Fergus, J. Sohl-Dickstein, A. Doucet, and W. Grathwohl, "Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc," *arXiv Preprint 2302.11552*, 2023.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[28] W. Peebles and S. Xie, "Scalable diffusion models with transformers," *arXiv Preprint 2212.09748*, 2023.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[30] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *CoRR*, vol. abs/1909.12271, 2019.

[31] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *arXiv preprint arXiv:1709.10087*, 2017.

[32] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," *arXiv preprint arXiv:2108.03298*, 2021.

[33] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, "The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 9698–9705.

[34] Z. Zhao, S. Haldar, J. Cui, L. Pinto, and R. Bhirangi, "Touch begins where vision ends: Generalizable policies for contact-rich manipulation," 2025.

[35] L. Ankile, A. Simeonov, I. Shenfeld, M. Torne, and P. Agrawal, "From imitation to refinement-residual rl for precise assembly," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 01–08.

[36] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.