Compositional Few-shot Learning of Motions

Omkar Patil
SCAI*
Arizona State University
Tempe, AZ 85281
opatil3@asu.edu

Anant Sah
SCAI
Arizona State University
Tempe, AZ 85281
asah4@asu.edu

Nakul Gopalan
SCAI
Arizona State University
Tempe, AZ 85281
ng@asu.edu

Abstract

A novel compositional approach called DSE- Diffusion Score Equilibrium that enables few-shot learning for novel skills by utilizing a combination of base policy priors is presented. Our method is based on probabilistically composing diffusion policies to better model the few-shot demonstration data-distribution than any individual policy. By using our few-shot learning approach DSE, we show that we are able to achieve a reduction of over 30% in MMD distance across skills and number of demonstrations. Moreover, we show the utility of our approach through real world experiments by teaching novel trajectories to a robot in 5 demonstrations.

1 Introduction

For robots to be deployed in unstructured environments and interact with humans, they should be capable of combining previously learned skills along with utilizing any given demonstrations. However, finding the right skills to combine from a base set and the extent of their contributions in the resulting motion is non-trivial. Existing compositionality methods either directly pick and choose the priors to compose while only learning the ratios of the priors' contribution Peng et al. [2019], or do not have a method to utilize residual information in the provided demonstrations Urain et al. [2023], Wang et al. [2024].

To tackle these shortcomings, we propose Diffusion Score Equilibrium(DSE), a compositional method that works over a set of base policies by inferring the extent of their contribution given a few demonstrations. Importantly, our method does not assume the policies to compose for achieving the desired behavior, and scales the contribution of base policies based on the information available in the provided demonstrations. A core element of our approach is inferring the contribution of each base policy in the resulting behavior, which we refer to as compositional weights henceforth. We infer these weights by minimizing the distance between a proposed trajectory and the few-shot demonstration data-distribution.

We show that by inferring the compositional weights by minimizing the Maximum Mean Discrepancy distance Gretton et al. [2012] over the Forward Kinematics (FK) kernel Das and Yip [2020] (MMD-FK), our method DSE scales with the number of provided demonstrations and achieves superior performance in both low and high data regimes. DSE results in 30% to 50% lower MMD-FK error in different data regimes than a demonstration fine-tuned policy and is also superior to prior compositional approach using diffusion models. Our contributions in this work are as follows-

- We present a novel compositional approach for sample-efficient learning called Diffusion Score Equilibrium (DSE). To the best of our knowledge, our work is also the first to learn compositional weights over a set of diffusion policies from the target demonstrations.
- We propose MMD-FK to fill the gap of a task and action space agnostic metric. We use the novel combination of the distributional MMD measure with the Forward Kinematics kernel to calculate distances between two trajectory distributions over the whole body of the robot.

*School of Computing and Augmented Intelligence

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

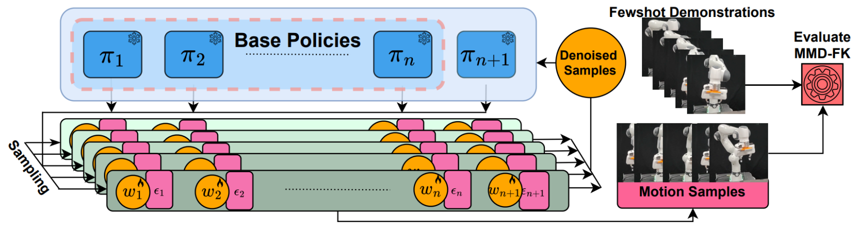

Figure 1: An outline of our approach. We assume a set of base policies π_i, $i = 1..N$ and train another policy π_{N+1} on the provided demonstrations. We compose over these policies and infer the compositional weights using quadratic optimization with the objective of MMD-FK. Only one optimization cycle is shown in the image.

2 Background

2.1 Policy Composition and Sampling

Our aim is to learn the action distribution a_0^L for a fixed trajectory length L from D demonstrations. Here, we use a to denote action for all the trajectory time-steps for brevity and drop the L notation. Gaussian diffusion models Sohl-Dickstein et al. [2015] learn the reverse diffusion kernel $p_\theta(a_t|a_{t-1})$ for a fixed forward kernel that adds Gaussian noise at each step $q(a_t|a_{t-1}) = \mathcal{N}(a_t; \sqrt{\alpha_t}a_{t-1}, (1-\alpha_t)\mathcal{I})$, such that $q(a_T) \approx \mathcal{N}(0, \mathcal{I})$. Here $t <= T$ represents the diffusion time-step and α_t the noise schedule. To sample from the product distribution, we need the score of the composition at each noise scale of the ancestral sampling chain. Our product distribution can be expressed as $p^{comp}(a_0) = p_\theta^1(a_0) * p_\theta^2(a_0)$, where a_0 has been specifically written to reflect that the distributions are composed in the data space. Then the score of the composed distribution $\nabla_{a_t} log\ q^{comp}(a_t)$ can be written as $\nabla_{a_t} log \left(\int \left[\prod q^i(a_0)\right] q(a_t|a_0) da_0\right)$. A long line of works instead add the individual scores of the distributions being composed $\sum_i \left(\nabla_{a_t} log \left[\int q^i(a_0) q(a_t|a_0) da_0\right]\right)$, since the former is not tractable. Du et al. [2023] bring this out as the reason for inferior quality of samples from composed image distributions and suggest Annealed MCMC samplers instead of ancestral sampling that does not result in the correct sequence of marginals expected by the reverse diffusion process. However, we utilize this sequence of marginals to interpolate between distributions.

3 Methodology

3.1 Novel Motion Generation by Composing Diffusion Models

To spatially blend between distributions for generating novel motion, we propose to sample from $q^{comp}(a_0) = \prod_{i=1}^N q_i(a_0)^{w_i}$, where $\sum_{i=1}^N w_i = 1$, where we have N base policies. The sum of scores of the composed distribution $\nabla_{a_t} log\ q^{comp}(a_t)$ at each time-step can then be approximated as $\sum_i^N w_i \left(\nabla_{a_t} log \left[\int q^i(\frac{a_0'}{\sqrt{\bar{\alpha}_t}}) \Phi\left(\frac{a_t - a_0'}{1-\bar{\alpha}_t}\right) da_0'\right]\right)$. Here Φ is the standard normal distribution. Here, we have split the mean and variance effects of the forward diffusion transition kernel $q(a_t|a_0)$ to suggest that the individual distributions being composed are not invariant across time-steps.

Expressing the i^{th} base policy distribution at diffusion time-step t as an EBM $p_{i;t}(a) = exp(-E_{i;t}(a))/Z_\theta$, we get its score as $\nabla logp_{i;t}(a) = -\nabla E_{i;t}(a)$, where $E_{i;t}$ represents the noisy shifted energy function. The gradient of the energy function $\nabla E_{i;t}(a)$ is proportional to the output of diffusion models $\hat{\epsilon}_{i;\theta}(a_t, t)$, both of which estimate the score of the data distribution corresponding to the i^{th} base policy Du et al. [2023]. Thus a weighted addition of the diffusion model outputs $\sum_{i=1}^N w_i \hat{\epsilon}_{i;\theta}(a_t, t)$ where $\sum_{i=1}^N w_i = 1$ is proportional to the gradient of the weighted energy function $\nabla\left(\sum_{i=1}^N w_i E_{i;t}(a)\right)$ at diffusion time-step t. Hence, this enables sampling from regions that are not minimums in any of the individual energy functions or distributions being composed, while also lending some control over it's placement.

3.2 MMD-FK Metric

Several integral probability metrics have been proposed in the image generation literature such a FID Heusel et al. [2017] and Maximum Mean Discrepancy (MMD) Gretton et al. [2012] to quantitatively evaluate the generated samples with respect to the data distribution. Moreover, we would like our metric to measure the distance in the task space where the effect of motion composition is apparent, and not be limited to the end-effector actions. With these requirements in consideration, we propose MMD-FK, a metric that uses the MMD distance on the FK kernel to evaluate the distance between two robot-link trajectory distributions. Our metric $\hat{dist}^2_{MMD-FK}(X, Y)$ for m and n samples from the two distributions respectively can be expressed as:

$$\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} K_{FK}(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_{FK}(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} K_{FK}(x_i, y_j) \tag{1}$$

It leverages MMD for it's kernel support that enables measurement of the distance between two distributions in terms of the distance between their feature means in a latent space. To evaluate task-space distances even with action space as the robot configuration, we use the positive-definite Forward Kinematics kernel as suggested in Das and Yip [2020]. Here $K_{FK}(x, x') = \frac{1}{M}\sum_{m=1}^{M} K_{RQ}(FK_m(x), FK_m(x'))$ is the positive-definite Forward Kinematics kernel in Equation 1. It sums over the m control points defined on the robot, typically associated with each link in the kinematic chain. K_{RQ} is a second-order rational quadratic kernel $K_{RQ}(x, x') = \left(1 + \frac{\gamma}{2}||x - x'||^2\right)^{-2}$, with the width of the kernel being $\gamma > 0$.

3.3 Diffusion Score Equilibrium

We present our few-shot learning approach DSE shown in Figure 1 in this section. Assuming M motion demonstrations D_j where $j = 1..M$, we want to learn the optimal policy, which we evaluate using the MMD-FK distance between the data-distribution and samples from the policy. Given the limited number of demonstrations, the policy trained on the few-shot data learns a very noisy estimate of the score function. Sampling from such a policy often results in incorrect motions as the energy function gradient estimates are not accurate. *Our main insight is to use gradient priors from the base set of policies to get a more accurate estimate of actual gradient towards the minimum.* We use this score estimate as a prior for our policy learned on the few-shot data $w_{comp}\hat{\epsilon}_{comp;\theta}(a_t, t) + w_{fs}\hat{\epsilon}_{fs;\theta}(a_t, t)$ where $w_{comp} + w_{fs} = 1$. This can be reformulated as $\sum_{i=1}^{N+1} w_i\hat{\epsilon}_{i;\theta}(a_t, t)$ where $\sum_{i=1}^{N+1} w_i = 1$, where the $(N+1)^{th}$ policy is trained on the few-shot demonstrations D. Finally, we estimate w_i by minimizing MMD-FK between the few-shot demonstration data and our composed policy samples.

Estimating w_i is challenging, but attempts have been made previously to estimate the sampling parameters in differentiable samplers for diffusion models Watson et al. [2022] with gradient based methods. These gradient based methods are computationally expensive due to multiple backward passes through the model. Instead, we utilize a non-gradient based quadratic optimizer Kraft [1988] to tune our weights with the objective function of MMD-FK. Our approach is described in Algorithm 1.

4 Experimental Details

4.1 Data Generation and Model Architecture

We generate 200 joint-position demonstrations using damped-least squares based differential inverse kinematics Buss [2004] for Franka Research-3 robot in Mujoco Todorov et al. [2012], as shown in Figure 2. These priors execute these trajectories in task space with random initial end-effector orientations and positions. All our policies are trained on the smallest variant of DiT Peebles and Xie [2023], conditioned on the initial state of the robot in configuration space. The model $\hat{\epsilon}_\theta(a_t, o, t)$ learns to predict the noise that was added to the input a_t, conditioned on the diffusion time-step t and the observation o using AdaLN Perez et al. [2018]. The models were trained using the standard hyper-parameter configuration as resented in the DiT paper. The training was performed on NVIDIA RTX A5000 GPUs and took approximately 2 hours for each model till 2000 epochs.

Algorithm 1 DSE: Compositional Weight Estimation

Input: Base policies p_i, $i = 1..N$; Demonstrations D
Output: Compositional weights w_i
Initialize : Train a diffusion model p_{N+1} on the demonstration data D
Minimize MMD-FK:
for $l = 1$ to OPT_ITER **do**
  Initialize : w_i, $\sum_{i=1}^{N+1} w_i = 1$
  for $k = 1$ to $NUM\_SAMPLES$ **do**
    for $t = 1$ to $NUM\_INFERENCE\_STEPS$ **do**
      $\hat{\epsilon}_{comp} = \sum_{i=1}^{N+1} w_i \hat{\epsilon}_{i;\theta}(a_t, t)$
    end for
  end for
  Calculate $MMD\text{-}FK(SAMPLES, D)$
end for
return $w_i, i = 1..N + 1$

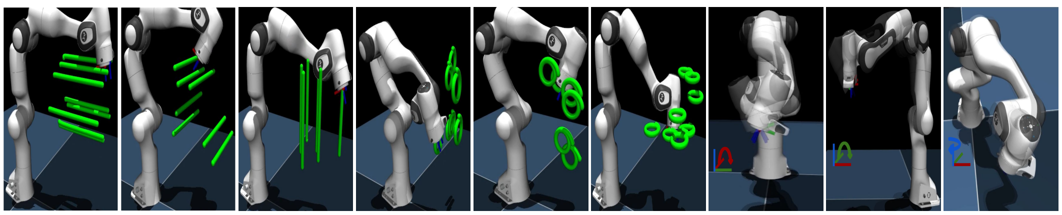

Figure 2: Base policies in order: $LineX$, $LineY$, $LineY$, $CircleX$, $CircleY$, $CircleZ$, $OscX$, $OscY$, $OscZ$. The last three base policies Osc oscillate about the specified axis with fixed end-effector position.

4.2 Sequential Quadratic Optimization

A core element of our approach is the optimization procedure to evaluate the compositional weights. The sample size for the quadratic optimizer is adjusted based on the number of demonstrations in the few-shot dataset. For all the experiments, we run the optimization procedure 4 times, where it is initialized with the normalized MMD-FK values between the prior motion datasets and the novel demonstration dataset, and three random initial values that sum to 1. We found that the optimization was also able to recover the base policies from corresponding demonstration data collected on the real robot. The optimization procedure took around 10-20 minutes depending upon the number of samples considered to evaluate MMD-FK on a single GPU.

5 Results

5.1 Few-shot learning

We use prior motions corresponding to a line, a circle and inverted pendulum along the X, Y and Z axis as base policies for most of our experiments, visually depicted in figure 2. We utilize two baselines to compare against our approach. The first is the composition of diffusion policies as proposed by Du et al. [2023, 2020]. We find optimal compositional weights for this method using the optimization procedure similar to ours. The second is a non-compositional baseline of a diffusion model trained on the demonstration data. We compare DSE against our baselines for 4 novel trajectories not seen by the robot, two in a simulated setting, and two collected on the real robot. We report MMD-FK values with the reference trajectory distribution wherever available, evaluated over 50 samples. Table 1 shows the results for the simulated experiments. DSE consistently achieves a lower or comparable MMD-FK score than both the baselines on all the tasks, for 5, 15 and 40 demonstrations. While we visually represent the end effector trajectories in Section 4, our method optimizes the compositional weights for all the links of the robot. Further experimental details can be found in Appendix A.1 and the rollout videos can be accessed on our project webpage [2].

[2] `https://sites.google.com/asu.edu/comp-fsl`

Table 1: MMD-FK scores for 50 rollouts across skills and demonstrations counts. Details on the few-shot trajectories provided for $StepX$ and $OscX + LineXZ$ can be found in the Appendix A.1.

Trajectories	Number of demos	Vanilla Composition	Fine-tuned Policy	Diffusion Score Equilibrium
StepX	5	0.79	0.50	**0.25**
	15	**0.18**	0.27	0.20
	40	**0.15**	**0.17**	**0.12**
OSC X + Line XZ	5	0.75	0.57	**0.32**
	15	0.30	0.25	**0.06**
	40	0.37	**0.14**	**0.12**

For our real world experiment, we collected 15 demonstrations resembling an S along the X-axis and Spring motion along X-axis. The MMD-FK results are shown in Table 2 and visually represented in Figure 5. DSE also achieved lower MSE with the collected demonstrations than the baselines, confirming the utility of our metric MMD-FK for evaluating compositional weights.

Table 2: Robot experiment results where we collected 15 demonstrations on Franka FR3 to train our policies. DSE achieves lower MMD-FK/MSE values exhibiting robustness to noise when learning.

Trajectories	Number of demos	Vanilla Composition	Fine-tuned Policy	Diffusion Score Equilibrium
S Motion	5	**0.50** / 0.0076	0.69 / 0.0034	**0.56 / 0.0019**
	15	1.70 / 0.0148	0.69 / 0.0023	**0.34 / 0.0015**
Spring Motion	5	1.65 / 0.016	4.28 / 0.0037	**0.37 / 0.0024**
	15	0.91 / 0.0110	5.10 / 0.0022	**0.47 / 0.0013**

6 Discussion and Limitations

As the number of training demonstrations are increased, the weight assigned by our approach DSE to the fine-tuned model increases. This is expected as if we have more demonstrations our model picks the true data distribution rather than the compositions over the base policies. However, as we observe more data vanilla composition models also perform better as they get a better estimate of the trajectory distribution. Further, our priors are not orthogonal, can be multi-modal and be chosen with a lot of freedom. This is unlike policy composition using multiplicative Gaussian policies Peng et al. [2019] which cannot handle multi-modality. Moreover, Gaussian Mixture Models face the challenge of exploding number of modes as the number of prior policies increase, further highlighting the efficiency of DSE. Our results can also improve with more priors however this would lead to increased compute time to find optimal weights. Finally, we do want to acknowledge that these compositions are in the state space of the robot rather than in the raw observation space such as the visual observations of the robot.

7 Conclusion

We present a novel compositional approach to few-shot learning called Diffusion Score Equilibrium (DSE) based on equilibrium of scores predicted by diffusion models. Our approach composes a policy trained on the target demonstrations with a set of base policy priors and infers the compositional weights by minimizing a measure of distance between the resulting composed distribution and the demonstration data distribution. Empirically, we observed that DSE will perform better than a policy simply trained on the data irrespective of the number of provided demonstrations on average by $30\% - 50\%$, while outperforming it by significant margins in the few-shot regime. We also propose a novel metric MMD-FK to measure the distance between two movement trajectory distributions for the whole body of the robot.

References

Samuel R Buss. Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods. *IEEE Journal of Robotics and Automation*, 17(1-19):16, 2004.

Nikhil Das and Michael C Yip. Forward kinematics kernel for improved proxy collision checking. *IEEE Robotics and Automation Letters*, 5(2):2349–2356, 2020.

Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. In *Advances in Neural Information Processing Systems*, 2020.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc, 2023.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Dieter Kraft. A software package for sequential quadratic programming. *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. Mcp: Learning composable hierarchical control with multiplicative compositional policies. *Advances in Neural Information Processing Systems*, 32, 2019.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

Julen Urain, Anqi Li, Puze Liu, Carlo D'Eramo, and Jan Peters. Composable energy policies for reactive motion generation and reinforcement learning. *The International Journal of Robotics Research*, 42(10):827–858, 2023.

Lirui Wang, Jialiang Zhao, Yilun Du, Edward H Adelson, and Russ Tedrake. Poco: Policy composition from and for heterogeneous robot learning. *arXiv preprint arXiv:2402.02511*, 2024.

Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*, 2022.

A Appendix

A.1 Detailed Results

A.1.1 Results with Multi-modal Priors

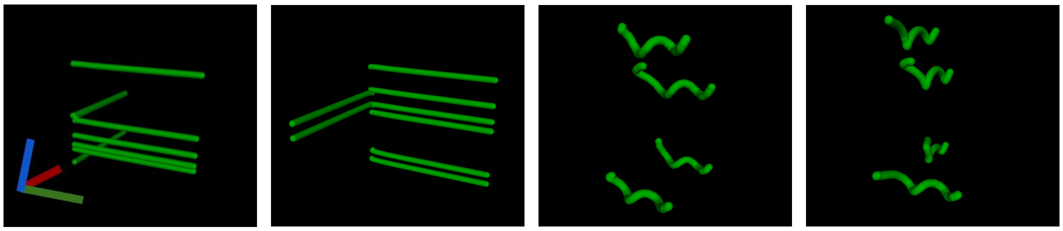

Figure 3: This panel of figures shows A: Demo data for $+X/+Y$ data. B: Demo data for $+X/-Y$ data. C: Policy rollout of composition of $LineX$ and $CircleX$. D: Composition of Line$+X/+Y$ and $CircleX$

We train multi-modal priors to test compositional approach's ability to sample from regions of high probability in both the distributions as shown in image A and B of Figure 3. We train policy A to reach towards the $+X$ or $+Y$ direction and policy B to reach towards the $+X$ or the $-Y$ direction. We expect the composed policy C with $w_1 = w_2 = 0.5$ to sample from the modes of reaching towards the $+X$ direction as the $+X$ behavior exists in both Policy A and B. We see exactly this behavior as the MMD-FK between Policy A and policy C is 0.58, between Policy B and Policy C is 0.27 and Policy C and a $+X$ direction policy is 0.11. Lower values of MMD-FK indicates lower errors or higher match between the two trajectory distributions. Composing policies to sample from the common regions of high probability was also shown for the reach and obstacle avoidance task by Urain et al. [2023]. However, their work used hand crafted potential functions to compose these distributions Urain et al. [2023]. We also showcase spatial blending where we compose a policy $CircleX$ and policy $LineX$ to create a spiral, as shown in image C Figure 3. The MMD-FK metrics obtained for both the cases are provided in Table 3. Finally, we showcase the result of composing the multi-modal policy Line$+X/+Y$ and policy $CircleX$ in image D in Figure 3. The composed policy is more dominant along the $+Y$ direction due to the directional similarity of motions.

Table 3: MMD-FK values between samples from the composed and the base policy distributions. The compositional weights are taken to be $w_1 = w_2 = 0.5$ for both cases. Self-Comparison implies that the MMD-FK is calculated between demonstration data and rollouts for the same policy.

	+X	CircleX
Spiral Vanilla Composition	0.92	0.87
Self-Comparison	0.03	0.01

We also present few-shot results in the multi-modal setting. We generate a spiral trajectory along the X-axis as the target policy. For this experiment, we consider only Line$+X/+Y$ and $CircleX$ as our prior policies. The vanilla composition method clearly struggles in this case due to the prior policy being multi-modal. DSE performs the best of the three approaches compared as shown in Table 4 and visually depicted in Figure 4.

Table 4: MMD-FK scores for 50 rollouts across skills and demonstrations counts for few-shot demonstrations in simulation for $SpiralX$. Vanilla composition allocates majority of the compositional weight to $LineX$, with DSE also using the residual information from the provided few-shot demonstrations. DSE out-performs both our baselines in terms of MMD-FK.

Trajectories	Number of demos	Vanilla Composition	Fine-tuned Policy	Diffusion Score Equilibrium
Spiral X	5	**0.58**	0.64	**0.51**
	15	0.58	0.26	**0.09**
	40	0.58	0.15	**0.09**

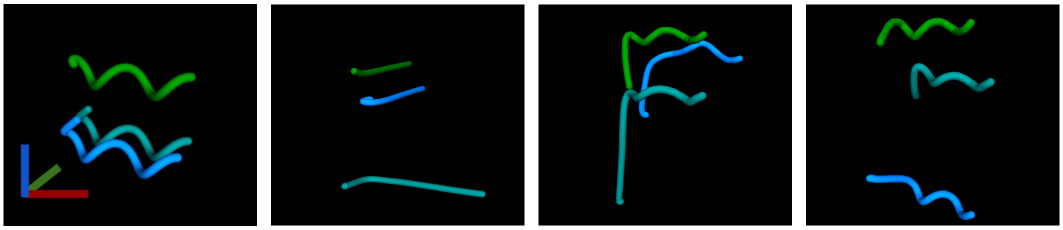

Figure 4: This panel of figures shows A: EEF few-shot demo data for spiral trajectory. B: Policy rollout of vanilla composition C: Policy rollout of the fine-tuned policy trained on 15 demos D: Policy rollout of DSE trained on 15 demos.

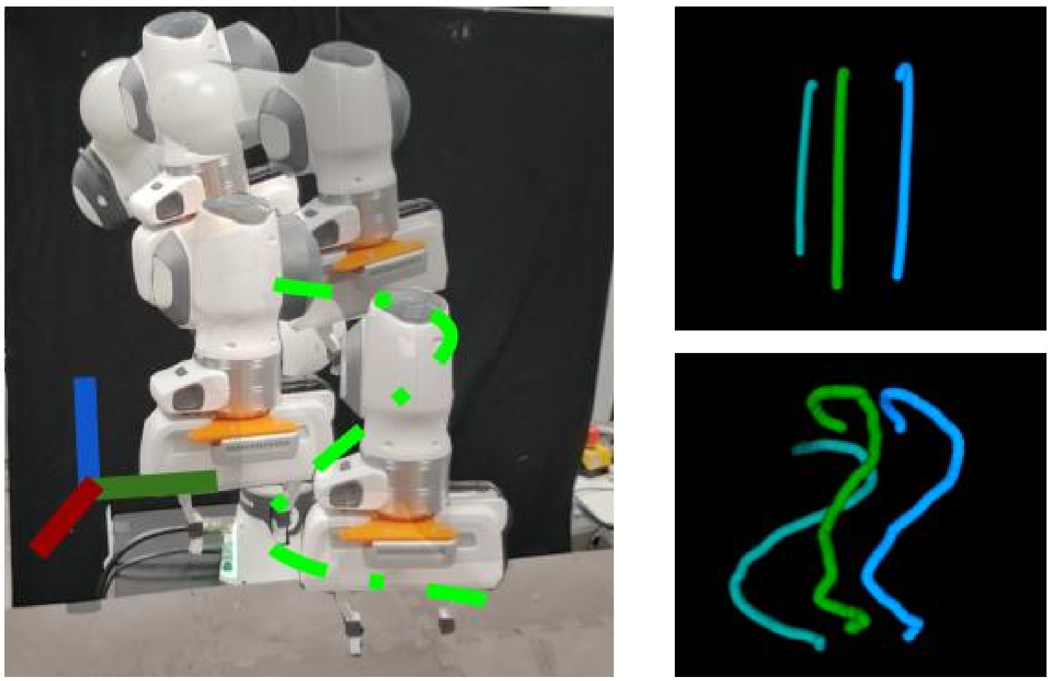

Figure 5: This panel of figures shows Left: Overlay of real robot demonstration collection for S-motion along the X-axis; Top-right: Policy rollout of vanilla composition with 15 demos; Bottom-right: Policy rollout of DSE trained on 5 demos.

A.1.2 Main Results

We provide details on the simulated few-shot demonstrations and analyze our results closely below. We also visually depict the end-effector trajectory resulting from the policy rollouts for the few-shot demonstrations of S-motion collected on the real robot in figure 5.

- **Step**: We generate a step trajectory in the XZ plane. We observe that DSE policy performs surprisingly well with just 5 demonstrations, largely due to the base policy gradient priors, while the fine-tuned policy does not perform well. As the number number of demonstrations is increased, the fine-tuned policy catches up to DSE in terms of MMD-FK.
- **OscX+LineXZ**: We create a difficult target distribution for the final case in the simulated setting. The robot end effector moves along a line while the robot body is oscillating about the X axis. We observe that the fine-tuned policy performance gets better with increasing number of demonstrations while compositional weight optimizer struggles due to the small oscillatory movements in the target.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The title and the introduction are not inflated and accurately reflect the method, domain and the results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discuss our limitations in section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theorem or proof has been provided in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We specify most experimental settings in the Appendix **??**.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the code after the camera-ready version of this paper is submitted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The main paper is brief due to space limitations. All the details are specified in Section **??**.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our results are evaluated over 50 rollouts of the model. We were not able to calculate the standard deviation in the results due to shortage of computational resources and time. However, our results are stronger than the next baseline by a large margin.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Provided in Appendix **??**.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to all the specified code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We believe there is no larger societal impact as the presented method is limited to efficient learning of robotic motions.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not required, as we do not train large, general-purpose models. In fact we do quite the opposite.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have utilized several open-source code-repositories in our work. We have appropriately credited and acknowledged them in our code, and the paper if necessary.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our research does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our research does not involve human subjects or crowd-sourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.