Compositional Few-shot Learning of Movement Trajectories

Omkar Patil, Anant Sah, Nakul Gopalan

***Abstract*—Humans can perform various combinations of physical skills without having to relearn skills from scratch every single time. For example, we can swing a bat when walking without having to re-learn such a policy from scratch by composing the individual skills of walking and bat swinging. Enabling robots to combine or compose skills is essential so they can learn novel skills and tasks faster with fewer real world samples. To this end, we propose 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. Our method is based on probabilistically composing diffusion policies to better model the few-shot demonstration data-distribution than any individual policy. Our goal here is to learn robot motions few-shot and not necessarily goal oriented trajectories. 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.**

I. Introduction

For robots to be deployed in unstructured environments and interact with humans, they should be capable of learning new skills from very few demonstrations. For example, wiggling the end-effector while moving forward to clean a table is a combination of two independent motions. This wiggling motion can be combined with different primitive motions to clean floors, to wash dishes, to fit a bed-sheet, to iron a cloth, etc. These are not goal oriented trajectories, but continuous motions that are sometimes dynamical trajectories in configuration space where a robot follows a sequence of movements. Robots should not be expected to learn these composed motions one at time but rather combine 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 [1], or do not have a method to utilize residual information in the provided demonstrations [2, 3].

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.

Underlying our approach is the insight that composing diffusion models can result in novel motion generation that interpolates between the individual distributions. We leverage this insight to efficiently learn a novel skill by interpolating between the noisy distribution learned from the few demonstrations of a novel skill and the set of base policy distributions for minimizing the distance to the few-shot demonstration data-distribution. We show that by inferring the compositional weights by minimizing the Maximum Mean Discrepancy distance [4] over the Forward Kinematics (FK) kernel [5] (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). Our method does not rely on manually choosing which base policies to compose, and scales the performance with the number of demonstrations provided for the new skill. 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.
- We showcase that our approach is superior to simple probabilistic composition of base policies and even training a model on the demonstration data. We showcase our results on nine non-orthogonal base policy priors and with multi-modal priors for several new trajectories that the robot has not learned before. Moreover, our real world experiments to teach the robot novel trajectories from a few demonstrations showcase the robustness and utility of DSE with noisy real robot data to learn policy compositions.

II. Background

A. 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 [6] learn the reverse diffusion kernel $p_\theta(a_t|a_{t-1})$ for a fixed forward kernel that adds Gaussian noise at each step $q(a_t|a_{t-1}) =$

School of Computing and Augmented Intelligence, Arizona State University. Contact:{opatil3, asah4, ng}@asu.edu

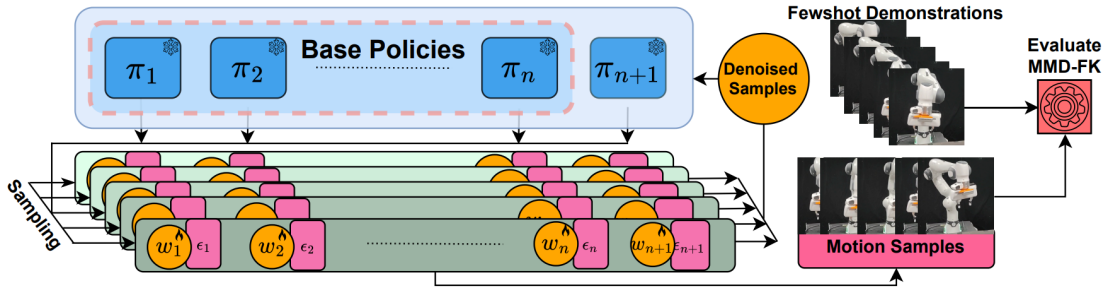

Fig. 1: An outline of our approach. We assume a set of base policies π_i, $i = 1..N$ and train another policy π_{N+1} on the provided demonstrations. We compose over these policies and infer the compositional weights using an optimization procedure with objective of Maximum Mean Discrepancy on the Forward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

$\mathcal{N}(a_t; \sqrt{\alpha_t}a_{t-1}, (1-\alpha_t)\mathcal{I})$, 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 generate trajectories from the learned data distribution $p_\theta(a_0)$, we sample at time step T from $\mathcal{N}(0, \mathcal{I})$ and apply the reverse diffusion kernel $p_\theta(a_t|a_{t-1})$ at each time step.

To sample from a product of distributions, we need the score of the composition at each noise scale of the ancestral sampling chain. Our product distribution can be expressed as $p^{comp}(a_0) = p^1_\theta(a_0) * p^2_\theta(a_0)$, 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. [7] 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.

III. Methodology

A. 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 at each time-step can then be written as:

$$\nabla_{a_t} log\ q^{comp}(a_t) \approx \sum_i^N w_i \left(\nabla_{a_t} log \left[\int q^i(\frac{a'_0}{\sqrt{\bar{\alpha}_t}}) \Phi\left(\frac{a_t - a'_0}{1-\bar{\alpha}_t}\right) da'_0\right]\right) \tag{1}$$

Where 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 $\nabla log p_{i;t}(a) = -\nabla E_{i;t}(a)$, 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 [7]. 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.

B. MMD-FK Metric

Several integral probability metrics have been proposed in the image generation literature such a FID [8] and Maximum Mean Discrepancy (MMD) [4] 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 for m and n samples from the two distributions respectively can be expressed as:

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

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 [5]. Here K_{FK} is the positive-definite Forward Kinematics kernel in Equation 3. Equation 3 sums over the m control points defined on the robot, typically associated with each link in the kinematic chain. For our purposes, one control point on each kinematic chain allows us to capture the movements of the links of the robot in the task-space. In Equation 3, 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 $\gamma > 0$.

$$K_{FK}(x, x') = \frac{1}{M} \sum_{m=1}^{M} K_{RQ}(FK_m(x), FK_m(x')) \quad (3)$$

C. 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 [9] 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 [10] to tune our weights with the objective function of MMD-FK. Our approach is described in Algorithm 1.

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 $j = 1$ to $NUM\_INFERENCE\_STEPS$ **do**
      for $i = 1$ to $N + 1$ **do**
        Obtain $\hat{\epsilon}_{i;\theta}(a_t, t)$
      end for
      $\hat{\epsilon}_{comp} = \sum_{i=1}^{N+1} w_i\hat{\epsilon}_{i;\theta}(a_t, t)$
    end for
  end for
  Calculate $MMD\text{-}FK(SAMPLES, D)$
end for
return w_i, $i = 1..N + 1$

IV. Experiments

A. Data Generation and Model Architecture

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, unless explicitly specified. We generate joint-position demonstration data using damped-least squares based differential inverse kinematics [11] for Franka Research-3 robot in Mujoco [12]. The 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 [13], 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 [14].

B. Few-shot learning

We present our few-shot learning results in this section. We utilize two baselines to compare against our approach. The first is the composition of diffusion policies as proposed by Du et al. [7, 15]. We find optimal compositional weights for this method using the optimization procedure similar to ours. The sample size for the optimization procedure is adjusted based on the number of demonstrations in the few-shot dataset. The second is a non-compositional baseline of a diffusion model trained on the demonstration data. A core element of our approach is the optimization procedure to evaluate the compositional weights. 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. 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. We also report the mean squared error values with the trajectories collected on the real robot for all the policies. Table I 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 the paper, our method optimizes the compositional weights for all the links of the robot. The video rollouts of the composed trajectories can also be viewed on our webpage[1].

- **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.

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

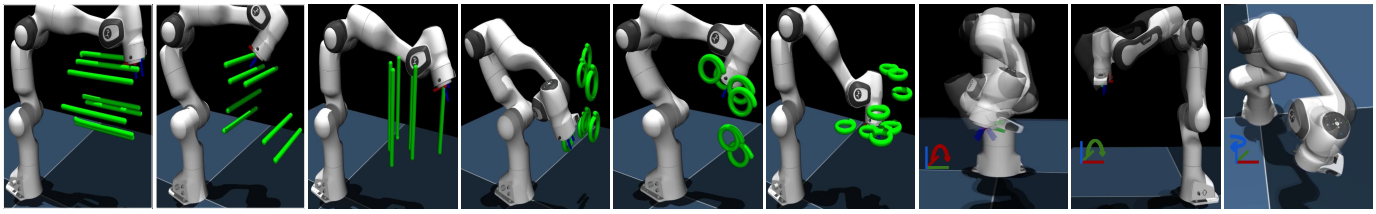

Fig. 2: Base policies in order: $LineX$, $LineY$, $LineY$, $CircleX$, $CircleY$, $CircleZ$, $OscX$, $OscY$, $OscZ$. The last three base policies Osc oscillate about the specified axis with fixed end-effector position.

- **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.

Trajectories	Number of demos	Vanilla Composition	Fine-tuned Policy	Diffusion Score Equilibrium
Step	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**

TABLE I: MMD-FK scores for 50 rollouts across skills and demonstrations counts. DSE out-performs both our baselines consistently.

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 II and visually represented in Figure 3. We also show the mean squared error(MSE) between the collected demonstrations on the real robot and the rolled out trajectory from the corresponding initial states. DSE also achieved lower MSE with the collected demonstrations than the baselines, confirming the utility of our metric MMD-FK for evaluating compositional weights.

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**

TABLE II: 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.

V. 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.

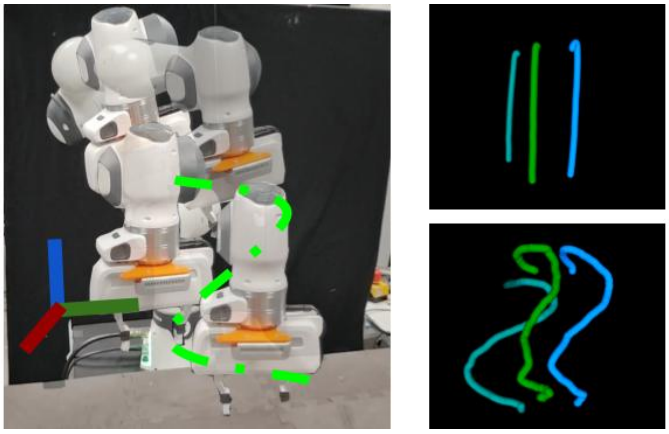

Fig. 3: This panel of figures shows Left: Overlay of real robot demonstration collection; Top-right: Policy rollout of vanilla composition with 15 demos; Bottom-right: Policy rollout of DSE trained on 5 demos.

Our results can also improve with more priors however this would lead to increased compute time to find optimal weights. 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 [1] 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. Finally, while we choose diffusion model priors for this work, the same can be achieved for different model families such as Gaussian.

VI. 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

[1] X. B. Peng, M. Chang, G. Zhang, P. Abbeel, and S. Levine, "Mcp: Learning composable hierarchical control with multiplicative compositional policies," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[2] J. Urain, A. Li, P. Liu, C. D'Eramo, and J. Peters, "Composable energy policies for reactive motion generation and reinforcement learning," *The International Journal of Robotics Research*, vol. 42, no. 10, pp. 827–858, 2023.

[3] 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.

[4] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[5] N. Das and M. C. Yip, "Forward kinematics kernel for improved proxy collision checking," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2349–2356, 2020.

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

[7] 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," 2023.

[8] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[9] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," in *International Conference on Learning Representations*, 2022.

[10] D. Kraft, "A software package for sequential quadratic programming," *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[11] S. R. Buss, "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods," *IEEE Journal of Robotics and Automation*, vol. 17, no. 1-19, p. 16, 2004.

[12] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[13] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[14] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[15] Y. Du, S. Li, and I. Mordatch, "Compositional visual generation with energy based models," in *Advances in Neural Information Processing Systems*, 2020.