Denoising Diffusion Probabilistic Models in Robotics: Theory and Applications

Bryce Grant[1] and Fabio Isaza[2]

Abstract— **Denoising Diffusion Probabilistic Models (DDPMs) represent a powerful class of generative models that has gained significant attention in recent years for high-quality image synthesis and other generative tasks. This project explores DPPMs through the lens of Markov chains and stochastic processes, with a specific focus on their emerging applications in robotics. We establish the foundations that make these models effective for both general generation tasks and robot control problems. DDPMs operate by defining a forward diffusion process as a Markov chain that gradually adds Gaussian noise to data, and then learning a reverse process that iteratively denoises random noise into samples resembling the original data distribution. In the robotics domain, this framework enables us to formulate robot policies by denoising actions. We will examine the mathematical framework of DDPMs, particularly focusing on their connection to stochastic differential equations and how they leverage Markov chains. Additionally, we will implement a DDPM model, NoProp, for image classification on the Fashion MNIST dataset, demonstrating how these models can generalize on state of the art datasets. This project aims to bridge the gap between recent advances in deep generative modeling, stochastic processes, and practical robotics applications.**

I. INTRODUCTION

Generative modeling has evolved significantly with the emergence of diffusion models, which provide a unique approach to generating complex data distributions. Denoising Diffusion Probabilistic Models (DDPMs) represent a significant advancement in generative modeling, combining principles from stochastic processes with deep learning architectures. Since their introduction by [1], diffusion models have rapidly gained prominence in image synthesis, demonstrating remarkable generation quality that rivals or exceeds previous approaches such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).The core innovation of diffusion models lies in their formulation of the generative process as the reversal of a Markov chain that gradually adds noise to data. This mathematical framework has proven particularly powerful due to its stable training dynamics and expressivity, enabling diffusion models to capture complex, multimodal distributions.

As we will demonstrate in this paper, the Markov chain structure makes diffusion models not only valuable for traditional generative tasks but also suitable for sequential decision processes in robotics. By viewing a robot's policy as a process of gradually denoising random actions into purposeful movements, diffusion models provide a novel approach to robot control that leverages their ability to model complex distributions.

This paper explores the mathematical foundations of diffusion models through the lens of Markov chains and stochastic processes, with a specific focus on their emerging applications in robotics. We first establish the foundational theory of DDPMs, highlighting their connections to Markov processes, and then examine their application to robotics through the lens of Diffusion Policy [2] and related frameworks.

II. RELATED WORKS

Denoising Diffusion Probabilistic Models represent an important advancement in generative modeling, building upon several research threads and inspiring numerous applications. This section explores the evolution of diffusion models and their applications in various domains, with a particular focus on robotics.

A. Diffusion Models in Generative Modeling

Diffusion models have emerged as powerful tools in generative modeling, evolving from earlier approaches. The foundational work by [3] introduced the concept of non-equilibrium thermodynamics for deep unsupervised learning, which laid the mathematical foundation for diffusion models. Building on this foundation, [1] formalized the Denoising Diffusion Probabilistic Models (DDPMs) framework, establishing the now-standard formulation of the forward and reverse processes as Markov chains.

Further refinements were introduced by [4], who proposed improvements to the noise scheduling and architectural design that significantly enhanced the sample quality. The connection between diffusion models and score-based generative models was established by [5], who showed that both approaches can be unified under a common framework of stochastic differential equations (SDEs). This theoretical bridge has enriched our understanding of the mathematical properties of diffusion processes.

The computational efficiency of diffusion models was addressed by [6], who developed Denoising Diffusion Implicit Models (DDIM) that enable fewer sampling steps with minimal loss in generation quality. This advancement has been crucial for practical applications where inference speed is important, such as robotics.

B. Applications in Vision and Language Domains

The versatility of diffusion models has led to their successful application across multiple domains. In computer vision, [7] demonstrated that diffusion models can outperform

Credit: The author from [1] made the slides and worked on the introduction, related works, theory and methodology, application as well as the conclusion [2] worked on the methodology section and related works

Code available at: `https://github.com/bryceag11/NoProp`

GANs in image synthesis, establishing a new state-of-the-art in generation quality. The capacity of diffusion models to handle multimodal distributions has made them particularly effective for tasks with inherent ambiguity.

Recent work has extended diffusion models to text generation [8], audio synthesis [9], and video generation [10]. The application of diffusion models to 3D content generation, as explored by [11], demonstrates their adaptability to different data modalities and dimensional spaces.

In the domain of vision-language models, diffusion approaches have been used for text-to-image generation [12] and text-guided image editing [13]. These applications showcase the ability of diffusion models to effectively condition generation on complex inputs, a capability that has direct implications for robotics applications.

C. Diffusion Models in Robotics

The application of diffusion models to robotics represents a significant new direction with promising results. [2] introduced Diffusion Policy, which reformulates robot control as a conditional denoising diffusion process. This approach offers several advantages for robotics, including the ability to model multimodal action distributions, handle high-dimensional action spaces, and maintain stable training dynamics.

Further exploration of diffusion models in robotics includes the work by [14], who proposed 3D Diffusion Policy (DP3) to improve generalization in visuomotor policy learning. Their approach leverages simple 3D representations to enhance robustness to varying viewpoints and environmental conditions, addressing a key challenge in robot learning.

Recent work by [15] introduces TinyVLA, a compact vision-language-action model that combines pretrained multimodal models with a diffusion policy decoder. This approach achieves fast inference speeds while maintaining high performance, eliminating the need for extensive pretraining on robotic datasets. TinyVLA demonstrates the synergy between diffusion-based policy learning and multimodal representation learning, offering a data-efficient path for developing capable robotic systems.

D. Alternative Training Methods

An interesting extension of diffusion models is their application to neural network training methodologies. [16] introduced NoProp, a novel training approach inspired by diffusion models that eliminates the need for backward or forward propagation. This method represents a radical departure from traditional gradient-based learning, potentially opening new avenues for neural network training with different computational characteristics.

The connection between NoProp and diffusion models highlights the broader impact of the diffusion paradigm beyond generative modeling. By viewing neural network training through the lens of denoising processes, NoProp establishes a conceptual bridge between diffusion models and fundamental aspects of machine learning.

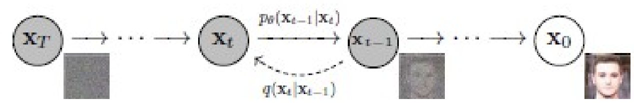

Fig. 1. Example of a diffusion model

E. Theoretical Connections to Markov Processes

The mathematical foundation of diffusion models is deeply rooted in the theory of Markov processes. The forward diffusion process in DDPMs defines a Markov chain that gradually transforms data into noise according to a predefined schedule. Similarly, the reverse process forms another Markov chain that progressively removes noise to generate samples.

The connection to continuous-time processes has been explored by [5], who showed that as the number of diffusion steps approaches infinity, the discrete Markov chain converges to a continuous stochastic differential equation (SDE). This connection provides additional theoretical tools for analyzing diffusion models and opens possibilities for more efficient sampling procedures.

The Markovian structure of diffusion models has been particularly beneficial for robotics applications, as it aligns well with the sequential nature of decision-making in robotic control. By leveraging this property, approaches like Diffusion Policy and TinyVLA have demonstrated superior performance in robot learning tasks, particularly in scenarios with complex, multimodal action distributions.

III. Theoretical Foundations

A. Markov Chains and Diffusion Processes

At their core, diffusion models establish a principled framework rooted in Markov chains. To understand this connection, we begin by defining a Markov chain as a stochastic process $\{X_t\}$ where the conditional probability distribution of future states depends only on the present state and not on the sequence of events that preceded it:

$$\begin{aligned} &P(X_{t+1} = x_{t+1}|X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) \\ &= P(X_{t+1} = x_{t+1}|X_t = x_t) \end{aligned} \tag{1}$$

Diffusion models leverage this property by defining two Markov processes: a forward process that progressively adds noise to data, and a reverse process that systematically removes noise.

B. Forward Process as a Markov Chain

The forward process in a DDPM defines a sequence of latent variables $x_1, x_2, \ldots, x_T$ obtained by gradually adding Gaussian noise to an initial data point x_0 according to a predefined variance schedule $\{\beta_t\}_{t=1}^{T}$. Mathematically, this process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

This is clearly a Markov process since x_t depends only on x_{t-1} and not on any earlier states. An important property of this formulation is that we can sample x_t directly from x_0 using a closed-form expression:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

This closed-form sampling is possible due to the Markovian nature of the process and the properties of Gaussian distributions. It allows us to directly sample from any step of the forward process without having to sequentially apply the transition kernel for each intermediate step.

C. Reverse Process and Denoising

The reverse process in diffusion models aims to invert the forward process, starting from pure noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ and progressively denoising until reaching a sample x_0 from the data distribution. While the forward process is fixed, the reverse process must be learned.

The true reverse process would follow:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t\mathbf{I}) \tag{4}$$

However, since x_0 is unknown during generation, we approximate this process with a learned model:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5}$$

This learned model is typically parameterized as a neural network that predicts the noise component of x_t, effectively learning to denoise the corrupted sample.

D. Connection to Stochastic Differential Equations

As the number of steps T approaches infinity, the discrete Markov chain of the diffusion process can be viewed as a discretization of a continuous stochastic differential equation (SDE). Specifically, the forward process converges to an Ornstein-Uhlenbeck process, while the reverse process can be expressed as a controlled SDE.

This connection to SDEs provides additional theoretical tools for analyzing diffusion models and opens possibilities for designing more efficient sampling procedures, such as those based on numerical SDE solvers.

E. Variational Lower Bound Objective

The training objective for diffusion models is derived as a variational lower bound on the log-likelihood of the data. This can be expressed as:

$$\mathbb{E}_{q(x_0)}[-\log p_\theta(x_0)] \leq \mathbb{E}_{q(x_0,x_1,\ldots,x_T)} \left[-\log \frac{p_\theta(x_0, x_1, \ldots, x_T)}{q(x_1, \ldots, x_T|x_0)}\right] \tag{6}$$

After mathematical manipulation and using the Markov properties of the processes involved, this simplifies to a weighted sum of denoising score matching objectives across different noise levels:

Fig. 2. Interpolations with 500 timesteps of diffusion

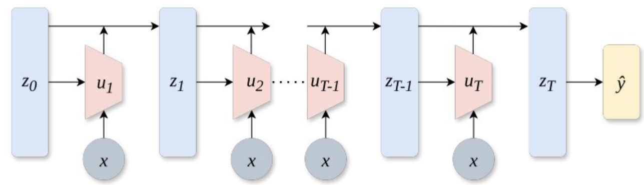

Fig. 3. Architecture of NoProp. z_0 represents Gaussian noise, while z_1, ..., z_T are successive transformations of z_0 through the learned dynamics u_1, ..., u_T with each layer conditioned on the image x, ultimately producing the class prediction $\hat{y}$.

$$L_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \tag{7}$$

where ϵ is the noise added to obtain x_t from x_0, and ϵ_θ is the neural network's prediction of this noise. This objective has the interpretation of training the model to predict the noise that was added during the forward process, enabling the model to iteratively denoise during generation.

IV. Methodology

During the inference phase, the sample input-label pair x and y are drawn from the target distribution $q_0(x, y)$ and the Gaussian noise z_0 goes through the diffusion process, where at each step in the process, the latent variable z_t is transformed by the diffusion block u_t (which is conditioned on the previous latent state z_{t-1} and the input x) producing the sequence $(z_t)_{t=0}^{T}$, where each term in the sequence corresponds to the stochastic activation functions of a neural network (used in the intermediate layers/blocks) with T number of blocks, which are trained to estimate $q_0(y|x)$. Next, we consider the decompositions of a stochastic forward propagation process p and the target distribution q:

$$p((z_t)_{t=0}^{T}, y|x) = p(z_0)\left(\prod_{t=1}^{T} p(z_t|z_{t-1}, x)\right) p(y|z_T), \tag{8}$$

$$q((z_t)_{t=0}^{T}|y, x) = q(z_T|y)\left(\prod_{t=T}^{1} q(z_{t-1}|z_t)\right). \tag{9}$$

It turns out that p can be expressed as a residual network with Gaussian noise added to its activation functions:

$$z_t = a_t\hat{u}_{\theta_t}(z_{t-1}, x) + b_t z_{t-1} + \sqrt{c_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}_d(\epsilon_t|0, 1) \tag{10}$$

where $\mathcal{N}_d(\epsilon_t|0, 1)$ is a d-dimensional Gaussian with mean vector 0 and identity covariance matrix, and a_t, b_t, c_t are scalars given by:

$$a_t = \frac{\sqrt{\overline{\alpha}_t}(1-\alpha_{t-1})}{1-\overline{\alpha}_{t-1}},\ b_t = \frac{\sqrt{\alpha_{t-1}}(1-\overline{\alpha}_t)}{1-\overline{\alpha}_{t-1}},$$
$$c_t = \frac{(1-\overline{\alpha}_t)(1-\alpha_t)}{1-\overline{\alpha}_{t-1}}. \tag{11}$$

Note that the $b_t z_{t-1}$ represents a weighted skip connection (the lines in the diagram that connect each z_{t-1} to the subsequent z_t, bypassing the corresponding u_t diffusion block), and $\hat{u}_{\theta_t}(z_{t-1}, x)$ is a residual block with parameterized by θ_t (this differs from typical deep neural networks in that we allow direct connections from the input x into each block.) We can also interpret p as a conditional latent variable model for y given x in that using variational formulation, we can learn the forward process p with the q distribution being the variational posterior. Then, the lower bound of the log likelihood $\log p(y|x)$ is

$$\log p(y|x) \geq E_{q((z_t)_{t=0}^T)|y,x} \left[\log p((z_t)_{t=0}^T, y|x) - \log q((z_t)_{t=0}^T, y|x)\right]. \tag{12}$$

Then, using Orenstein-Uhlenbeck process (it is variance preserving) gives us

$$q(z_T|y) = \mathcal{N}_d(z_T|\sqrt{\alpha_T}u_y, 1-\alpha_T),$$
$$q(z_{t-1}|z_t) = \mathcal{N}_d(z_{t-1}|\sqrt{\alpha_{t-1}}z_t, 1-\alpha_{t-1}) \tag{13}$$

where u_y is an embedding of the class label y in $\mathbb{R}^d$. After using properties of Gaussians, we have

$$q(z_t|y) = \mathcal{N}_d(z_t|\sqrt{\overline{\alpha}_t}u_y, 1-\overline{\alpha}_t),$$
$$q(z_t|z_{t-1}, y) = \mathcal{N}_d(z_t|\mu_t(z_{t-1}, u_y), c_t) \tag{14}$$

where $\overline{\alpha}_t = \prod_{s=t}^T \alpha_s$ and $\mu_t(z_{t-1}, u_y) = a_t u_y + b_t z_{t-1}$.Then, to optimize the lower bound of the log likelihood, we parametrize p to have the same form as q:

$$p(z_t|z_{t-1}, x) = \mathcal{N}_d(z_t|\mu_t(z_{t-1}, \hat{u}_{\theta_t}(z_{t-1}, x)), c_t),$$
$$p(z_0) = \mathcal{N}_d(z_0|0, 1) \tag{15}$$

where $p(z_0)$ is stationary. After plugging the parametrization into (12), we get

$$\mathcal{L}_{\text{NoProp}} = E_{q(z_T|y)}[-\log \hat{p}_{\theta_{\text{out}}}(y|z_T)]$$
$$+ D_{\text{KL}}(q(z_0|y)\|p(z_0))$$
$$+ \frac{T}{2}\eta E_{t\sim\mathcal{U}_{\{1,T\}}}[\text{SNR}(t) - \text{SNR}(t-1))\|\hat{u}_{\theta_t}(z_{t-1}, x) - u_y\|^2] \tag{16}$$

where $\text{SNR}(t) = \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}$ is th signal-to-noise ratio, η is a hyperparameter, and $\mathcal{U}_{\{1,T\}}$ is the uniform distribution on $1, ..., T$.

V. Application

The NoProp ResNet50 and ResNet18 model was implemented for our project for evaluation on the Fashion MNIST dataset. This dataset has the same format as MNIST (28x28 grayscale images, 60,000 training examples, 10,000 test examples), allowing us to adapt the existing code with minimal challenges. However, while structurally similar to MNIST, Fashion-MNIST presents a more challenging classification task, as clothing items have more complex features than digits

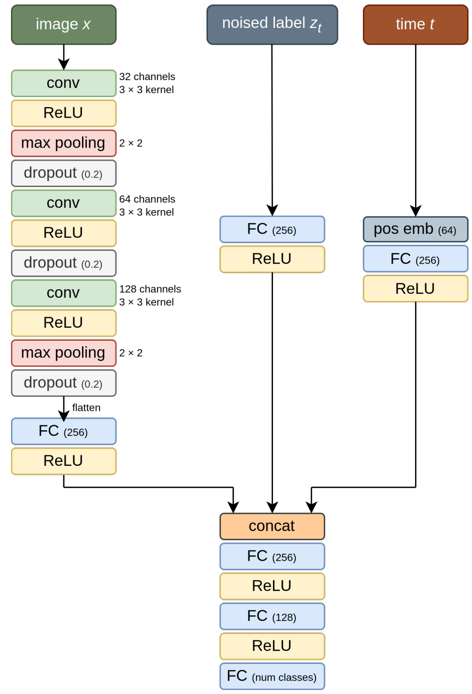

Fig. 4. NoProp Architecture for Resnet50

Fig. 5. Latent samples from linear (top) and cosine (bottom) schedules respectively at linearly spaced values of t from 0 to T.

A. Implementing NoProp for Fashion MNIST Classification

1) Architecture Details: Our implementation follows the architecture shown in Figure 4, where the diffusion blocks are structured as follows:

- Each input image x is processed through a convolutional backbone with 32, 64, and 128 channels respectively
- The noised label embedding z_t is processed through a fully connected network
- Time embedding t is encoded using positional encoding and processed through fully connected layers
- These three streams are concatenated and passed through final FC layers to produce class predictions

2) Experimental Setup: We trained the model for X epochs using the Adam optimizer with a learning rate of Y. The diffusion process used a cosine noise schedule with Z diffusion steps. We compared our NoProp approach against standard backpropagation training of an equivalent ResNet50 architecture.

B. Results and Analysis

As shown in Table I, NoProp achieves competitive accuracy compared to standard backpropagation while requiring

TABLE I
CLASSIFICATION ACCURACY ON FASHION MNIST TEST SET

Model	Test Accuracy (%)	Time per epoch (s)
ResNet18 (Backprop)	90.7	11.7
ResNet50 (BackProp)	92.5	21.2
NoProp ResNet18	90.3	3.1
NoProp ResNet50	91.8	8.7

less training and inference time. This demonstrates that diffusion-based training can be effective for classification tasks without requiring gradient backpropagation.

VI. CONCLUSION

This paper has explored denoising diffusion probabilistic models through the lens of Markov chains and stochastic processes, with a focus on their applications to robotics and neural network training. Our analysis highlights several key insights:

First, the Markovian structure of diffusion models provides a mathematically principled framework for generative modeling. By formulating generation as the reversal of a noise-adding Markov chain, diffusion models achieve both expressivity and stability, enabling them to capture complex, multimodal distributions.

Second, the application of diffusion models to robot control demonstrates the power of this framework beyond traditional generative tasks. The Diffusion Policy framework effectively leverages the properties of diffusion models to address challenges in visuomotor policy learning, including multimodal action distributions, high-dimensional action spaces, and stable training.

Third, extensions such as NoProp show how diffusion-inspired approaches can lead to fundamentally new paradigms in machine learning, challenging established practices like backpropagation and raising questions about the necessity of learned versus designed representations.

Future work in this area could explore several promising directions:

- Further theoretical analysis of the connection between diffusion models and stochastic processes, particularly non-Gaussian noise distributions
- Extensions of Diffusion Policy to more complex robotics scenarios, including multi-agent settings and long-horizon planning
- Exploration of hybrid approaches that combine the strengths of diffusion models with other learning paradigms
- Investigation of the computational efficiency of diffusion-based methods, which currently require multiple denoising iterations during inference

In conclusion, diffusion models represent a significant advancement in machine learning, with their foundation in Markov chains and stochastic processes. Their application to robotics and neural network training demonstrates their versatility and suggests that the full potential of these models is still being discovered.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[2] 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," 2024. [Online]. Available: https://arxiv.org/abs/2303.04137

[3] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," 2015. [Online]. Available: https://arxiv.org/abs/1503.03585

[4] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," 2021. [Online]. Available: https://arxiv.org/abs/2102.09672

[5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2021. [Online]. Available: https://arxiv.org/abs/2011.13456

[6] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2022. [Online]. Available: https://arxiv.org/abs/2010.02502

[7] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2105.05233

[8] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," 2022. [Online]. Available: https://arxiv.org/abs/2205.14217

[9] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2009.09761

[10] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2204.03458

[11] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," 2022. [Online]. Available: https://arxiv.org/abs/2209.14988

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[13] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10741

[14] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," 2024. [Online]. Available: https://arxiv.org/abs/2403.03954

[15] J. Wen, Y. Zhu, J. Li, M. Zhu, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, Y. Peng, F. Feng, and J. Tang, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2409.12514

[16] Q. Li, Y. W. Teh, and R. Pascanu, "Noprop: Training neural networks without back-propagation or forward-propagation," 2025. [Online]. Available: https://arxiv.org/abs/2503.24322