Leverage Noise Coupling in Generative AI models for Robotic Actions

Junhee Park
Digital Convergence Research Laboratory
Electronics and Telecommunications
Research Institute
Daejeon, Korea
juni@etri.re.kr

Yoosung Bae
Digital Convergence Research Laboratory
Electronics and Telecommunications
Research Institute
Daejeon, Korea
bys724@etri.re.kr

Hyeyoung AN
Digital Convergence Research Laboratory
Electronics and Telecommunications
Research Institute
Daejeon, Korea
hya@etri.re.kr

***Abstract*—A recent trend in robotics AI is to use generative AI modeling techniques to inference about robot behavior such as diffusion models and flow matching techniques. In this paper, we present a technique to give the meaningful starting noise of these generative AI models and a possible training algorithm for the generative models which have denoising steps in reasoning process. The performance demonstration based on large-scale data, the meaningful coupling of training data and noise, and mixture of experts architectures will be demonstrated in further research.**

***Keywords*—Robot Action AI, Generative AI Models, Diffusion, Flow Matching, Denoising, Noise Coupling**

I. INTRODUCTION

AI models for generating robot actions are evolving from training the explicit patterns that directly map observations to actions to training implicit patterns that are hidden between observations and actions.

Typical prior art applications of implicit pattern training methods to robot actions inference models include Diffusion Policy[1] and π_0 [2], both of which take a random value sampled from a Gaussian normal distribution as the starting actions and use a denoising process to find the correct actions. These methods have no relationship between the starting actions (noise) and the correct action data you are looking for, so the inference results are likely to be scattered and inaccurate over the range of the estimated distribution of the correct actions data. To solve this problem, it is necessary to assume a more accurate distribution of correct actions data, which in turn requires a large amount of training data.

To address these shortcomings in the training and inferencing robot actions, this paper proposes a noise coupling mechanism that applies a meaningful noise which is related to the correct actions.

II. BASELINE TECHNOLOGIES

Recently generative AI concept and technologies which have denoising steps for inferencing[3,4] are utilized in robot action reasoning AI models[1,2]. Fig. 1 and 2 [5] briefly describe the diffusion model [3] and the flow matching [4] techniques, which are the mainstreams of denoising based generative AI theories that are trending in the latest robot AI.

The sampled data from the Gaussian normal distribution becomes a starting noise in the denoising inference process that finally finds the inferred data which are in the scope of the correct data distribution. The starting noises in those methods [1,2,3,4], then, are not associated with the training data. They just are just randomly sampled from the Gaussian normal distribution, i.e. they are decoupled.

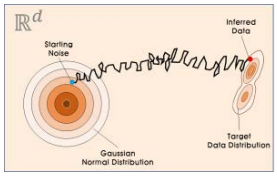


Fig.1 Diffusion Model[5]

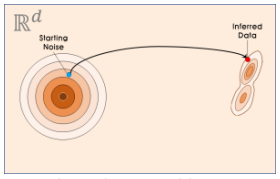


Fig.2 Flow Matching[5]

This paper proposes a method of replacing the starting noise with the one that reflects meaning in various ways – especially the latent variable from the conditional VAE(variational auto encoder)[6] algorithm. The result of this idea is expected to increase the correct answer rate as shown in Fig. 4[7] compared to the case where the starting noise and the correct answer data are not related at all (Fig. 3[7]).

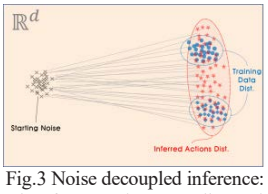


Fig.3 Noise decoupled inference: starting from random sampling noise, the inferred values are distributed over a large area encompassing the entire set of correct data[7]

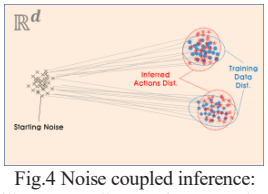


Fig.4 Noise coupled inference: utilizes sampling noise associated with the correct data, results in an inferred value that approximates the family of correct data[7]

III. A NOISE COUPLING BASED ROBOT ACTION TRAINING ALGORITHM

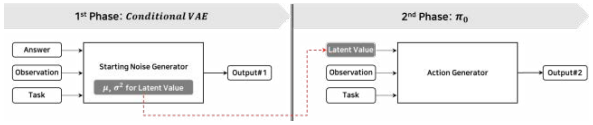
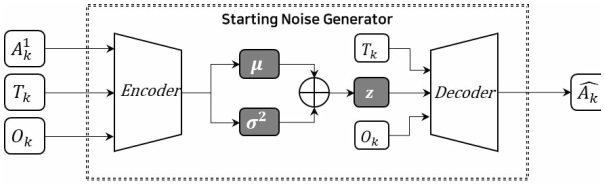


Fig.5 Brief Training Architecture



※ A_k^1: Answer action, $\widehat{A_k}$: Inferred action, T_k: Task, O_k: Observation, μ: mean, σ^2: variabnce, z: latent variable

Fig.6 Training Structure in 1st Phase : the same as cVAE[6]

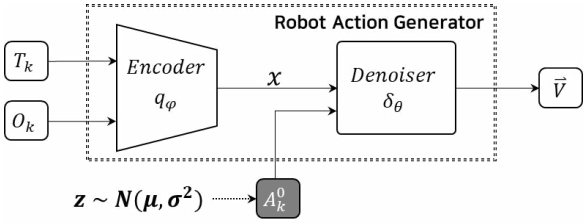


Fig.7 Training Structure in 2nd Phase: the same as [2]

Fig. 5 illustrates the overall brief structure of the proposed training process. The training is performed in two phases. In phase 1(Fig.6), the Starting Noise Generator(SNG) is trained. The loss functions used to train the SNG is the same as conditional VAE algorithm[6]. After the training of the SNG, we can have a pair of mean(μ) and variance(σ^2) which the latent variable(z) is sampled from. The mean and variance pair is the only information used for phase 2(Fig.7). The latent variable(z) is used as an input starting noise(A_k^0) to the Denoiser in training stage.

The Robot Action Generator(RAG) in phase 2 could be any robot action AI model which has the denoising process such as Diffusion Policy[1] and π_0[2]. In this paper, we are using π_0 as the RAG. The following algorithm is a supplement to the π_0 algorithm with the noise coupling method proposed in this paper. This algorithm is identical to π_0 except that it uses a z value sampled(line 6) from a distribution following the mean(μ) and variance(σ^2) values generated by the SNG at the initial noise setting (line 10) and inputs the set value into the learning process (line 12).

Algorithm : π_0[2] based Noise Coupled Training

1: Given: Training data set $\mathcal{D}$.
2: Let $\mathbb{E}, \mathcal{E}$ represent # of episode, # of epoch, μ, σ^2represent mean and variance from SNG, T_k, O_k represent a task, observations at step k.
3: Let A_k^t represents an action sequence at step k & time t, where A_k^0 is a starting noise action and A_k^1 is a final answer action at step k.
4: Initialize encoder $q_\varphi(x|T_k, O_k)$
5: Initialize denoising process $\delta_\theta(\vec{V}|x, z)$
6: Sample z from $N(\mu, \sigma^2)$
7: **for** iteration steps $k = 1, 2, \ldots, \mathbb{E} \times \mathcal{E}$ **do**
8: Sample x from $q_\varphi(x|T_k, O_k)$, where x is an embedding value of the task and observations at step k
9: Sample $t = random(0,1)$
10: Let $A_k^0 = z$, *the starting noise is coupled*
11: Get $A_k^t = t \cdot A_k^1 + (1-t)A_k^0$
12: Predict $\vec{V}$ with $\delta_\theta(\vec{V}|x, A_k^0)$
13: $Loss = MSE(\vec{V}, Vector(A_k^1, A_k^0))$
14: Update θ with $Loss$

IV. Future Works

The approach in this paper is expected to be very useful when training a separate data set specialized for unit actions of robots, and the following future tasks are to be necessary. If we can obtain refined training data by categorizing robot actions, the training data for each categorized action is predicted to distributed closer to each other, as shown in Fig. 4. By separating the data in this way and focusing on training, the noise coupling approach is expected to be even more effective. In such a case, it is anticipated that a technique for selectively utilizing initial noise according to the task to be performed will be required during the inference stage. It is also anticipated that a MOE(Mixture of Expert) structure that constructs and utilizes more specialized action expert models for each task will be effective.

ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [25ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling and evolving ways].

References

[1] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," Mar. 2023, arXiv:2303.04137v4. [Online] Available: https://arxiv.org/abs/2303.04137v4

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky, "π_0: A Vision-Language-Action Flow Model for General Robot Control," Oct. 2024, arXiv: 2410.24164. [Online] Available: https://arxiv.org/abs/2410.24164

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," 2020, arXiv:2006.11239. [Online] Available: https://arxiv.org/abs/2006.11239

[4] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. "Flow matching for generative modeling," 2022 arXiv:2210.02747. [Online] Available: https://arxiv.org/abs/2210.02747

[5] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, Itai Gat, "Flow Matching Guide and Code," 2024, arXiv:2412.06264v1. [Online] https://arxiv.org/abs/2412.06264

[6] Tony Z. Zhao, Vikash Kumar, Sergey Levine, Chelsea Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," 2023, arXiv:2304.13705. [Online] https://arxiv.org/abs/2304.13705

[7] Gianluigi Silvestri, Luca Ambrogioni, Chieh-Hsin Lai, Yuhta Takida, Yuki Mitsufuji, "VCT: Training Consistency Models with Variational Noise Coupling," 2025, arXiv:2502.18197v2. [Online] https://arxiv.org/abs/2502.18197