Diffusion Models for Robotics

Jessica E. Liang

Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104-6309 USA
jeliang@seas.upenn.edu

Abstract

Diffusion Models (DMs) offer robust tools for addressing uncertainty and enhancing adaptability in robotics. This work explores their application to trajectory generation, 3D image synthesis, and interpretable scene understanding. For trajectory planning, we propose using colored Gaussian noise to improve robustness and temporal coherence. In 3D image generation, Transfer Entropy enhances information flow between textual and visual modalities for more coherent outputs. Partial Information Decomposition (PID) is leveraged to improve model interpretability and efficiency in scene generation. Rigorous evaluation will assess trajectory quality, robustness, and real-world transferability, aiming to advance autonomous decision-making and scene understanding in robotics.

1 Introduction

Diffusion Models (DMs) operate by modeling data as a noisy process and learning to reverse that process to generate new samples (Ho, Jain, and Abbeel 2020). The forward diffusion process gradually corrupts data with Gaussian noise, while the reverse process learns the conditional probability distribution $p(x_{t-1}|x_t)$ to progressively remove the noise. This approach has the potential to handle uncertainty more robustly in robotics tasks such as 3D object and image generation, as well as trajectory planning.

This proposal aims to explore how DMs can be applied to enhance robotics by addressing uncertainty in trajectory generation and its surrounding scenes. By leveraging the inherent stochasticity of DMs, we seek to improve robots' adaptability and robustness in dynamic environments.

2 Background

2.1 Diffusion Models in Robotics

Diffusion Models have demonstrated increasing potential in addressing uncertainty and generating complex behaviors in various robotics tasks. They have been applied to robot morphology generation, as seen in DiffuseBot, which uses a physics-augmented diffusion model to design soft robotic forms (Wang et al. 2024), and to scene rearrangement through web-scale models that manipulate and understand environments (Kapelyukh, Vosylius, and Johns 2023). DMs have also been utilized for robot motion learning and planning, enabling robots to learn feasible motions from data (Carvalho et al. 2023), as well as for imitating human behavior in sequential tasks, which is crucial for natural human-robot interaction (Pearce et al. 2023). Hierarchical latent DMs have been proposed to generate scenes for robotic simulations (Kim et al. 2023), and the Diffusion Policy model has demonstrated their use in generating robot behaviors via denoising (Chi et al. 2023). DMs have further been applied to trajectory denoising for robot planning (Janner et al. 2022), parallel sampling from pretrained models for efficiency (Shih et al. 2024), and 3D scene generation, facilitating perception and interaction in robotics (Ze et al. 2024; Huang et al. 2023). Additionally, latent 3D DMs have been proposed for generating static and articulated assets (Ntavelis et al. 2023), point-voxel diffusion has been used for 3D shape generation (Zhou, Du, and Wu 2021), and text-guided voxel editing has enabled intuitive robot programming (Sella et al. 2023). These applications demonstrate the broad applicability and versatility of DMs in robotics.

2.2 Prior Work by the Author

In (Liang 2024a), we proposed a cross-modal information recovery and enhancement method using Multiple-Input Multiple-Output (MIMO) Variational Autoencoders (VAEs) for multimodal IoT systems, which is applicable to robots processing diverse sensor data. Additionally, in our work on causal discovery (Liang 2024b), we recognized the limitations of correlation-based approaches and developed necessary conditions for causal discovery using Partial Information Decomposition (PID).

3 Approach

3.1 Diffusion Models for 3D Image Generation

Developing DMs for 3D image generation can open up new possibilities in fields such as virtual reality, gaming, and medical imaging. Building upon existing work on text-guided voxel editing for 3D object generation (Sella et al. 2023), which introduced a volumetric regularization loss, we propose to incorporate *Transfer Entropy* into the loss function to enhance information flow between modalities.

Transfer Entropy is a measure of the directional information transfer between two random processes (Schreiber 2000; Barnett, Barrett, and Seth 2009). It quantifies how knowledge of one process X can reduce the uncertainty of another process Y beyond the information contained in Y itself. In the context of 3D image generation using latent DMs, let X represent the fixed text embeddings obtained from the CLIP text encoder, and let Y_t represent the latent variables at time t in the diffusion process, then the Transfer Entropy from X to Y is defined as:

$$T_{X \to Y} = \sum_{y_{t+1}, y_t, x} p(y_{t+1}, y_t, x) \log \frac{p(y_{t+1}|y_t, x)}{p(y_{t+1}|y_t)} \quad (1)$$

In this application, X is time-invariant, so x has no index t. By maximizing $T_{X \to Y}$, we encourage the model to effectively utilize textual information in generating the 3D images. This approach aims to improve the coherence and relevance of the generated content to the textual prompts.

3.2 Interpretable Diffusion Models for Robotics

To enhance interpretability and increase processing speed in stable diffusion models, we propose using PID (Williams and Beer 2010; Dutta, Venkatesh, and Grover 2022) to the DM design. Although PID was applied to interpret diffusion models in (Dewan et al. 2024), it was used solely for interpretation and not for the design of the models. PID allows us to decompose the mutual information between multiple variables into redundant, unique, and synergistic components. In this context: X_i and X_j represent the paired images and texts, respectively, and Y represents the generated images. The mutual information $I(X_i, X_j; Y)$ can be decomposed into Redundant Information, Unique Information, and Synergistic Information. The goal is to maximize redundant information and minimize synergistic information in the loss function during UNet training in stable diffusion. This promotes the generation of images that are closely aligned with both the text and image inputs, increasing training speed and enhancing the interpretability. Applying this model to scene generation for robots can improve their understanding and interaction with human and the environment and increase robots responding speed in real world applications.

3.3 Robotic Trajectory Generation

For robotic trajectory generation, the diffusion model will be trained to generate trajectories $T = \{x_0, x_1, ..., x_T\}$ conditioned on noisy sensor inputs in dynamic environments. Given a current state x_0 and a goal state x_T, the model will generate a distribution over feasible intermediate points that satisfy task-specific constraints such as obstacle avoidance.

While prior work (Janner et al. 2022) used white Gaussian noise in DMs for trajectory generation, we propose using *colored Gaussian noise*. Colored noise, which has a non-flat power spectrum, can capture temporal dependencies and directional preferences in trajectories. We observed that colored Gaussian noise can improve the performance of denoising diffusion probabilistic models.

By incorporating colored noise into the diffusion process, we aim to enhance the model's ability to generate smooth and realistic trajectories that respect dynamic constraints, improve convergence speed and stability in the denoising process, and increase robustness to sensor noise and environmental uncertainties.

4 Evaluation

The proposed models will be rigorously evaluated through a combination of simulated experiments and real-world trials.

- Trajectory Quality: Evaluate the generated trajectories for smoothness, efficiency, and safety. Metrics include path length, curvature, energy consumption, obstacle avoidance, and the number of collisions.
- Robustness to Uncertainty: Compare the performance of the diffusion-based models to deterministic approaches and reinforcement learning-based methods under varying levels of sensor noise and environmental unpredictability.
- Simulation-to-Real Transfer: Assess the models' ability to transfer knowledge from simulation to real-world tasks. Evaluate performance degradation, adaptability, and learning efficiency in real-world deployments.
- Interpretability: Analyze the interpretability of the models by examining the information decomposition and the alignment between inputs and generated outputs.

5 Discussion

The proposed diffusion-based framework has the potential to significantly advance robotics by:

- Enhancing Adaptability: By explicitly modeling uncertainty, robots can generate diverse sets of feasible trajectories, improving their ability to adapt to unexpected changes.
- Improving Human-Robot Interaction: Predicting human movements and intentions enables robots to respond more effectively, fostering safer and more intuitive interactions.
- Advancing Scene Understanding: Interpretable models for scene generation can enhance robots' perception and decision-making capabilities in complex environments.
- Contributing to Theoretical Foundations: Incorporating measures like Transfer Entropy and PID introduces new theoretical insights into how information flows in generative models, potentially influencing broader AI research.

Potential challenges include the computational demands of training and deploying diffusion models, and ensuring that the models generalize well across different environments.

6 Conclusion

In summary, this research proposes integrating Diffusion Models into robotics to address uncertainty in trajectory generation, 3D image generation, and scene understanding. By leveraging advanced information-theoretic measures and innovative noise modeling, we aim to enhance robots' autonomy and performance in dynamic environments. The anticipated outcomes include improved adaptability, safer human-robot interactions, and contributions to both the practical and theoretical aspects of AI and robotics.

References

Barnett, L.; Barrett, A. B.; and Seth, A. K. 2009. Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical review letters*, 103(23): 238701.

Carvalho, J.; Le, A. T.; Baierl, M.; Koert, D.; and Peters, J. 2023. Motion planning diffusion: Learning and planning of robot motions with diffusion models. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1916–1923. IEEE.

Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E.; Burchfiel, B.; and Song, S. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*.

Dewan, S.; Zawar, R.; Saxena, P.; Chang, Y.; Luo, A.; and Bisk, Y. 2024. DiffusionPID: Interpreting Diffusion via Partial Information Decomposition. *arXiv preprint arXiv:2406.05191*.

Dutta, S.; Venkatesh, P.; and Grover, P. 2022. Quantifying feature contributions to overall disparity using information theory. *arXiv preprint arXiv:2206.08454*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, S.; Wang, Z.; Li, P.; Jia, B.; Liu, T.; Zhu, Y.; Liang, W.; and Zhu, S.-C. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16750–16761.

Janner, M.; Du, Y.; Tenenbaum, J. B.; and Levine, S. 2022. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*.

Kapelyukh, I.; Vosylius, V.; and Johns, E. 2023. Dall-e-bot: Introducing web-scale diffusion models to robotics. *IEEE Robotics and Automation Letters*, 8(7): 3956–3963.

Kim, S. W.; Brown, B.; Yin, K.; Kreis, K.; Schwarz, K.; Li, D.; Rombach, R.; Torralba, A.; and Fidler, S. 2023. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8496–8506.

Liang, J. E. 2024a. Cross-Modal Information Recovery and Enhancement Using Multiple-Input Multiple-Output Variational Autoencoder. *IEEE Internet of Things Journal*, 11(15): 26470–26480.

Liang, J. E. 2024b. Partial Information Decomposition for Causal Discovery With Application to Internet of Things. *IEEE Internet of Things Journal*, 11(13): 24289–24299.

Ntavelis, E.; Siarohin, A.; Olszewski, K.; Wang, C.; Gool, L. V.; and Tulyakov, S. 2023. Autodecoding latent 3d diffusion models. *Advances in Neural Information Processing Systems*, 36: 67021–67047.

Pearce, T.; Rashid, T.; Kanervisto, A.; Bignell, D.; Sun, M.; Georgescu, R.; Macua, S. V.; Tan, S. Z.; Momennejad, I.; Hofmann, K.; et al. 2023. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*.

Schreiber, T. 2000. Measuring information transfer. *Physical review letters*, 85(2): 461.

Sella, E.; Fiebelman, G.; Hedman, P.; and Averbuch-Elor, H. 2023. Vox-e: Text-guided voxel editing of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 430–440.

Shih, A.; Belkhale, S.; Ermon, S.; Sadigh, D.; and Anari, N. 2024. Parallel sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Wang, T.-H. J.; Zheng, J.; Ma, P.; Du, Y.; Kim, B.; Spielberg, A.; Tenenbaum, J.; Gan, C.; and Rus, D. 2024. Diffusebot: Breeding soft robots with physics-augmented generative diffusion models. *Advances in Neural Information Processing Systems*, 36.

Williams, P. L.; and Beer, R. D. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Ze, Y.; Zhang, G.; Zhang, K.; Hu, C.; Wang, M.; and Xu, H. 2024. 3d diffusion policy. *arXiv preprint arXiv:2403.03954*.

Zhou, L.; Du, Y.; and Wu, J. 2021. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5826–5835.