Research on effective methods for encoding modalities for VLA (vision-language-action) models in robotics

This research investigates the development of a Diffusion Transformer policy for Vision-Language-Action (VLA) models in robotics, focusing on integrating diverse modalities—text, images, 3D data, and video—into a coherent, shared representation. Using the latent diffusion process, this approach enables efficient encoding of multi-modal inputs, supporting robust policy learning through scalable stochastic action generation. By utilizing internet-scale multi-modal datasets, the Diffusion Transformer aims to establish a versatile policy that can be effectively adapted to downstream robotic tasks, thereby improving the generalization and adaptability of robotic control across varied environments and task types.

1 Introduction

Table 1: Comparative Analysis of Multi-Modal Integration Approaches for Robotic Policy Learning.

Solution	Strengths	Weakness
Diffusion Policy [1]	• Robust in modeling multi-modal action distributions • Exhibits stable training behavior • Scalable to high-dimensional action spaces	• Training relies on Stochastic Langevin Dynamics, which is computationally intensive • Lacks flexibility to adapt to novel modalities or arbitrary context input

Theia: Distilled Vision Foundation Model [2]	• Distills diverse VFMs for compact representations, reducing computational costs • Enhances visual knowledge for robot learning • Effective for downstream robot learning with less data	• Does not generalize well to unseen tasks outside the visual domain • Mainly limited to visual representations, lacking integration with action-specific modalities
Octo: Generalist Robot Policy[3]	• Pretrained on the largest multi-robot dataset • Provides flexible fine-tuning across different sensory inputs and action spaces • Enables efficient adaptation to new robotic platforms	• Restricted to robotic manipulation settings • Primarily focuses on visuomotor control • Lacks explicit integration of language-based commands
OpenVLA [4]	• Vision-Language-Action model capable of multi-robot control • Highly adaptable via parameter-efficient fine-tuning • Fully open-source for community use	• High model complexity with 7B parameters makes deployment challenging • Constrained by the amount of diverse data available in the Open X-Embodiment dataset

Transfusion [5]	• Seamlessly integrates text and image data by combining next token prediction and diffusion objectives • Scales well in cross-modal benchmarks	• Complexity in training with modality-specific encoding and decoding layers • High FLOPs requirement compared to discrete token-based approaches
RT-Affordance [6]	• Incorporates affordances as intermediate representations, offering efficient guidance for manipulation • Provides strong generalization across novel objects and scenes	• Limited to affordance-based control, may not handle arbitrary input-output mappings in diverse robotic tasks • Performance highly dependent on the quality of affordance data available

The section references contain the full list, collected for this project.

References

[1] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric A. Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *ArXiv*, abs/2303.04137, 2023.

[2] Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. *ArXiv*, abs/2407.20179, 2024.

[3] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Pannag R. Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. *ArXiv*, abs/2405.12213, 2024.

[4] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *ArXiv*, abs/2406.09246, 2024.

[5] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *ArXiv*, abs/2408.11039, 2024.

[6] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. 2024.