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

Abstract: 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.

Datasets: The dataset used in the computational experiment is sourced from open-source projects, which provide data in standardized formats, ready for modeling.

1. **Open X Embodiment Dataset:** The dataset consists of contributions from 22 different robots, collected through a collaboration between 21 institutions, demonstrating 527 skills (160,266 tasks). It includes diverse robotic manipulation tasks, enabling training and evaluation of generalist policies adaptable to new robots, tasks, and environments. The dataset link and details are available here.

2. **RLBench Dataset:** RLBench provides a suite of robotic tasks captured in a multimodal format, including vision, language, and action data, suitable for training and evaluating generalist robotic policies. This dataset is designed for benchmarking imitation learning and reinforcement learning methods. Details are available here.

3. **ManiSkill2 Benchmark Dataset:** ManiSkill2 offers a large-scale dataset for multimodal, multi-task robotic manipulation. It supports diverse evaluation protocols and comparisons with prior works. More information can be found here.

References: Papers with a fast intro and the basic solution to compare.

1. Open X Embodiment research paper on robotic manipulation policies [1].

2. OpenVLA [2] Large policies pretrained on a combination of Internet-scale vision-language data and diverse robot demonstrations have the potential to change how we teach robots new skills: rather than training new behaviors from scratch, we can fine-tune such vision-language-action (VLA) models to obtain robust, generalizable policies for visuomotor control.

3. TinyVLA [3] A new family of compact vision-language-action models, called TinyVLA, which offers two key advantages over existing VLA models: (1) faster inference speeds, and (2) improved data efficiency, eliminating the need for pre-training stage..

Basic solution: A link [1] to the code of the baseline algorithm and the implementation of the RT-X model. This serves as the state of the art for comparison with the proposed solution.

Authors: Expert and Consultant Kazybek Askarbek, Ruslan Rakhimov

Supplementary: This section provides a comprehensive overview of additional resources supporting this research:

- **Problem Statement:** Robotic systems often lack the capability to generalize across diverse tasks and environments, requiring extensive retraining for each new scenario. This research aims to address these limitations by leveraging multi-modal datasets and a unified Diffusion Transformer policy that enables cross-platform adaptability.
- **Methodology:** The project integrates a latent diffusion process to encode diverse modalities (vision, language, and action) into a shared latent space. Preprocessing includes normalizing visual data, tokenizing language inputs, and standardizing action representations. Policy training uses a imitation learning on curated benchmarks.
- **Training Setups:** Experiments are conducted on NVIDIA A100 GPUs, with a batch size of 64 and a learning rate of 1e-4. The training pipeline incorporates distributed training for scalability, with an average runtime of 72 hours per model on datasets with 500+ tasks.
- **Evaluation Metrics:** Performance is measured using task success rates, action consistency scores, and transfer learning efficiency. Additional metrics include the scalability of the model to new environments and the reduction in fine-tuning time for unseen tasks.
- **Extended Documentation:** Comprehensive resources, including annotated datasets, pre-trained models, and codebases, are provided here. Tutorials detail the implementation of baseline policies and adaptation techniques.
- **Experimental Materials:** Video demonstrations showcase the execution of manipulation tasks such as object stacking, tool use, and environment navigation. Benchmarks compare the performance of Diffusion Transformer policies against state-of-the-art models.
- **Reproducibility Resources:** Open-source code for model training, evaluation scripts, and configuration files are included to facilitate reproducibility. Pre-trained models for common tasks are made available to accelerate experimentation.

References

[1] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Haoshu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeff Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishan Rana, Krishna Parasuram Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Manfred Otto Heess, Nikhil J. Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R. Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Ho Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan C. Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. Open x-embodiment: Robotic learning datasets and rt-x models. *ArXiv*, abs/2310.08864, 2023.

[2] 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.

[3] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, and Jian Tang. Tinyvla: Towards

fast, data-efficient vision-language-action models for robotic manipulation. *ArXiv*, abs/2409.12514, 2024.