Opinion: Small VLAs Self-Learn Consistency

Anonymous Author(s)
Affiliation
Address
email

Abstract

Robotics increasingly leverages behavioral cloning for contact-rich tasks where accurate simulators are infeasible and dense reward functions difficult to define. Collected by humans sequentially, input trajectories are non-i.i.d. data and thus randomized to mitigate non-stationarity, and more closely adhere to the fundamental theoretical assumptions underlying statistical learning. Rather than modeling single actions, modern visuomotor policies are trained to model action chunks, which are crucially considered in complete isolation during training. However, empirical evidence suggests that powerful visuomotor policies seem to pick up on the sequential nature of the input trajectories provided during training, reproducing increasingly more consistent chunks, despite not being instructed to do so. In this opinion piece, we present initial empirical evidence substantiating the claim that, when fine-tuned on extra demonstrations, small-size VLAs might learn to exploit aspects of the input data self-learning consistency, conversely to larger models which in the same setting become less self-consistent.

1 Introduction

Learning policies from collections of human demonstrations is an increasingly popular approach in robotics [Brohan et al., 2022, Zhao et al., 2023b, Chi et al., 2024, Kim et al., 2024, Li et al., 2024, Black et al., 2024, O'Neill et al., 2024, Shukor et al., 2025]. Learning from real-world demonstration—reward-free—data proves particularly effective in highly dexterous tasks, where (1) simulation may prove expensive and (2) defining a reward function is non trivial.

Expert demonstrations are typically recorded via *tele-operation*, a process consisting of a human expert controlling symbiotic robot platforms while performing a task, all while recording the visuomotor data associated to its commands over time (an *expert trajectory*). Then, learning a desired behavior can be reduced to learning to reproduce these trajectories, approximating the mapping between visuomotor inputs—*(i)* camera views and *(ii)* robot's proprioperception—and the control applied by the human demonstrator. Learning from (potentially, large-scale) tele-operation data [Khazatsky et al., 2024, Collaboration et al., 2023] also appears to be uniquely positioned to benefit from the recent advancements in developing multi-modal foundation models [Beyer et al., 2024, Hurst et al., 2024], combining advancements in perception and visual reasoning with traditional planning.

Given an observation o_t of the environment, modern robotics policies π are trained to reproduce the expert demonstration by outputting *sequences* of H actions $\mathbf{A}^H$—action *chunks*—rather than a single action a_t drawn from $\pi(\bullet|o_t)$. Indeed, Zhao et al. [2023a] argue providing a controller with multiple actions to be enacted sequentially not only proves effective in mitigating catastrophic error compounding, but also aligns with the psychological understanding of how individual actions are grouped and executed as an atomic unit [Lai et al., 2022]. The prevalent technique considered is thus to learn multiple actions originating from a single, input observation of the environment, modifying accordingly the dataset to exhibit this chunk-level structure. Critically, during training action chunks $\mathbf{A}_t^H = \pi(o_t)$ are considered in isolation. That is, action chunks $\mathbf{A}_t^H$ are not compared to neighboring

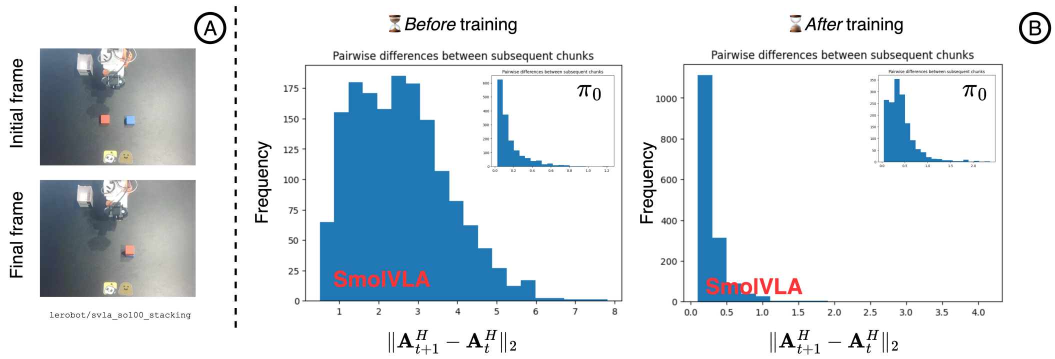

Figure 1: (A) Initial (top) and final (bottom) camera frames of the cube-stacking demonstration. Demonstrations start with cubes in arbitrary positions on a plane, and terminate with the two cubes stacked. (B) Histograms of the L2-norm differences $\|\mathbf{A}^H_{t+1} - \mathbf{A}^H_t\|_2$ between successive action chunks before (left) and after (right) training on the demonstrations, illustrating the marked improvement in temporal consistency during training (π_0 for control on the top-right of each visualization).

chunks $\left[k \in \mathbb{N} : \mathbf{A}^H_{t-k}, \mathbf{A}^H_{t-(k+1)}, ..., \mathbf{A}^H_{t+(k-1)}, \mathbf{A}^H_{t+k}\right]$ while learning from reward-free data. For similar, successive observations, one would naturally expect well-performing policies to produce similar actions, assuming generally unimodal demonstrations for a given task. Yet, such expectation seems to be only partially met by empirical evidence: in a small scale experiment assessing the similarity of successive chunks for similar observations for (1) a small, light-weight Vision-Language-Action model (VLA) and (2) a large, state-of-the-art VLA model, we found the discrepancy between the corresponding chunks to (1) increase and (2) decrease when fine-tuning.

Motivated by analyzing this phenomenon, we investigate the evolution of the similarity of neighboring action chunks over fine-tuning for the two different models. In particular, we assess the similarity of successive action chunks—i.e., *chunks' consistency*—for SmolVLA [Shukor et al., 2025], a compact VLA designed for deployment on low-end hardware platforms, trained on small-scale crowd-sourced and open-source robotics dataset. We evaluate the similarity of action chunks obtained for subsequent observations before, during and after further-training SmolVLA on a specific dataset, and observe that chunks become more and more temporal consistent as training proceeds. Conversely, when reproducing the same procedure with the same fine-tuning demonstrations on π_0 [Black et al., 2024], we found chunks to not increase in similarity, and in fact to widen as fine-tuning progresses—an observation we believe could prove interesting in understanding the training dynamics of VLAs.

Our experiments indicate further-training SmolVLA on a task-specific dataset seem to biases the model towards becoming more and more self-consistent, while π_0 exhibits the opposite behavior.

2 Background

Taken together, *(i)* multi-modal backbones for semantic reasoning over multi-modal input streams, *(ii)* reward-free learning via imitation, and *(iii)* chunk-level consistency mechanisms define the landscape within which our analysis is positioned.

2.1 Multi-modal Foundation Models for Robotics.

The recent advent of large-scale *Vision-Language Models* (VLMs) [Alayrac et al., 2022, Beyer et al., 2024] has provided robotics with precisely the kind of rich, general-purpose perception required to model potentially-noisy human demonstrations. By pretraining on billions of image-text pairs, VLMs acquire semantic representations that transfer remarkably well to downstream tasks and domains, including robotics [Brohan et al., 2023, Kim et al., 2024, Black et al., 2024, Shukor et al., 2025]. A common recipe for VLMs training couples a vision encoder with a pretrained language model (LM), trained solely on text [Radford et al., 2021, Zhai et al., 2023, Fini et al., 2024]. The merged system is subsequently exposed to multi-modal data through a sequence of increasingly supervised stages: (i) large-scale caption corpora [Schuhmann et al., 2022, Byeon et al., 2022], (ii) interleaved image-text documents [Laurençon et al., 2023, Zhu et al., 2023], and (iii) instruction-tuning collections to elicit conversational skills [Tong et al., 2024, Laurençon et al., 2024]. Besides semantic understanding, efficiency also emerged as an equally prominent objective in training VLMs. Computational budgets can be reduced by designing more compact backbones [Marafioti et al., 2025, Korrapati, 2024, Yao et al., 2024], or adopting parameter-efficient techniques to draw inference, or even update the model weights specifically for inference [Shukor et al., 2023, Vallaeys et al., 2024, Tsimpoukelli et al., 2021].

Robotics Transformer 2 (RT-2) [Brohan et al., 2023] demonstrated the connection between pre-trained VLMs and robotics explicitly: in their method, Brohan et al. [2022] present a frozen, internet-scale VLM used as perceptual backbone, while a task-specific action head is fine-tuned on the tele-operation data collected. Subsequent work has embraced the same recipe, giving rise to *Vision-Language-Action* (VLA) models, jointly processing language instructions, visual observations, and proprioceptive inputs to output series of actions [Kim et al., 2024, Wen et al., 2024].

2.2 Imitation Learning for Robotics

Learning control policies directly from human demonstrations [Brohan et al., 2022, Zhao et al., 2023b, Chi et al., 2024, Kim et al., 2024, Black et al., 2024, Shukor et al., 2025] has emerged as a powerful alternative to Reinforcement Learning (RL), especially in the context of dexterous manipulation where specifying dense reward functions is notoriously difficult, and high-fidelity simulation proves expensive or even unfeasible. In the standard tele-operation setting, an expert controls the robot while the system records synchronized streams of visual observations, proprioceptive readings, and the control commands actually executed. A policy π is then trained *without any task rewards* to reproduce the expert behavior by mapping an observation o_t to an *action chunk* $\mathbf{A}_t^H \in \mathbb{R}^{H \times D}$ specifying H consecutive low-level actions in the D-dimensional robot joint space. Predicting temporally extended sequences—that is, H actions—not only reduces error compounding but also mirrors the hierarchical structure of human motor control [Zhao et al., 2023b, Lai et al., 2022].

$\boldsymbol{\pi_0}$ Recent work by Black et al. [2024] leverages the idea of action chunking in the context of developing a foundation model for robotics. In particular, Black et al. [2024]'s π_0 architecture grafts a flow-matching diffusion head onto a pretrained VLM [Beyer et al., 2024], enabling control while inheriting internet-scale semantic understanding of the image data coming from camera streams. After pre-training on expert trajectories collected across diverse embodiments, pi_0 exhibits task-generalization by proving to be a *single* policy that can *zero-shot* perform highly dexterous tasks like folding shirts or bussing tables, receiving instructions in pure natural language.

SmolVLA While very effective, models like π_0 can prove to be difficult to deploy in resource-constrained scenarios. SmolVLA [Shukor et al., 2025] focuses precisely on resource-constrained deployment, developing a compact robotics model trained without rewards. In particular, SmolVLA couples a lightweight SigLIP vision encoder with a sub-400M parameter vision-language model backbone, and adds an action head as action expert, yielding a model with a total of sub-500M parameters. Despite its size, SmolVLA still retains the VLA recipe—joint image-language conditioning and chunked action prediction—and Shukor et al. [2025] report competing scores against baselines including both ACT [Zhao et al., 2023b] and π_0 [Black et al., 2024]. Crucially for fine-tuning and inference, the authors report SmolVLA can be fine-tuned and run on consumer-grade GPUs and even CPUs.

3 Analysis

In this study, we assess SmolVLA's [Shukor et al., 2025] internal consistency when producing action chunks for a cube-stacking manipulation task, where the robot must *(i)* grasp a cube from an arbitrary location and *(ii)* place it in stable equilibrium atop a second cube in a different location 1. In our work, we resort to the openly available implementation of SmolVLA provided with LeRobot [Cadene et al., 2024]. Importantly, we evaluate the model consistency *before* and *after* fine-tuning SmolVLA on a dataset of *cube stacking* demonstrations[1]. Our findings hint reward-free training does impact the

[1] huggingface.co/lerobot/datasets/svla_so100_stacking

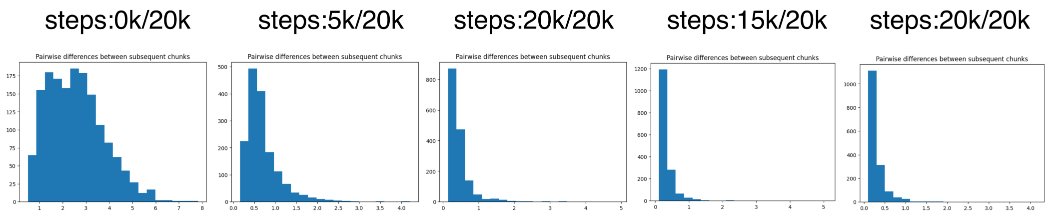

Figure 2: Empirical histograms of the L2-norm differences $\|\mathbf{A}^H_{t+1} - \mathbf{A}^H_t\|_2$ between successive action chunks at early, intermediate, and final stages of training. The increasingly narrow distributions indicate reduced temporal variability for successive chunks (with `Chunk-0` $\leftarrow \mathbf{A}^H_t$ and `Chunk-1` $\leftarrow \mathbf{A}^H_{t+1}$).

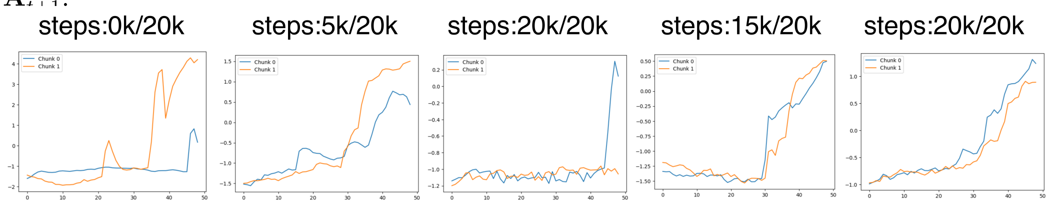

Figure 3: 1D PCA projection of successive action chunks `Chunk-0` and `Chunk-1` visualized over $H = 50$ timesteps over the course of training. Visualizations illustrate the pair of chunks scoring the median value for L2 difference over the course of training. In the worst case, PCA explains 60%+ of the total variance.

inner consistency of the model on overlapping action chunks. In particular, as training progresses the model becomes more and more consistent across chunks obtained for successive observations despite not having been explicitly instructed nor influenced to. Conversely, a control-experiment using π_0 does not result in the same behavior, and in fact π_0's consistency decreases as fine-tuning progresses (Figure 1).

Figure 1(B) confirms reward-free training induces SmolVLA to generate internally coherent action chunks over successive timesteps, capturing smooth and semantically consistent transitions *without* explicit temporal regularization at training time—this seems to be indicating consistency emerges from reproducing human demonstrations. Importantly, Figure 2 shows empirical distributions of $\|\mathbf{A}^H_{t+1} - \mathbf{A}^H_t\|_2$ over training, underscoring how the narrowing dynamics matches the progress of the training process, and that task-specific training results in improvements in temporal consistency.

To further validate this claim, we visualize representative chunk pairs $p = \{\mathbf{A}^H_t, \mathbf{A}^H_{t+1}\}$ whose L2-norm difference corresponds to the distribution's median value during training, and present a 1D-projection of the otherwise 6D joint representation through PCA (Figure 3). The PCA projection onto the principal component reveals progressively tighter alignment between successive chunks as training proceeds. Additionally, overlaying the joint-space trajectories of these median chunk pairs empirically demonstrates the reduction in drift over training the downstream task space, further validating the impact of training on execution consistency for successive action chunks.

4 Conclusions

Our findings highlight a divergence in temporal consistency across VLA models: while fine-tuning SmolVLA increases the similarity of successive action chunks, π_0 exhibits the opposite trend. This contrast suggests that model scale and pretraining may differently shape chunk-level training dynamics, and motivates further investigation into consistency as a key property of visuomotor policies.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascale, Jade Choghari, Jess Moss, and Thomas Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. `https://github.com/huggingface/lerobot`, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David

Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. `https://arxiv.org/abs/2310.08864`, 2023.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

Vik Korrapati. Moondream. Online, 2024. URL `https://moondream.ai/`. Accessed: 2025-03-27.

Lucy Lai, Ann Zixiang Huang, and Samuel J Gershman. Action chunking as policy compression. *PsyArXiv*, 2022.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=SKN2hflBIZ`.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *ICLR*, 2024.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

C Schuhmann, A Köpf, R Vencu, T Coombes, and R Beaumont. Laion coco: 600m synthetic captions from laion2b-en. *URL https://laion.ai/blog/laion-coco*, 2022.

Mustafa Shukor, Corentin Dancette, and Matthieu Cord. ep-alm: Efficient perceptual augmentation of language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22056–22069, 2023.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Théophane Vallaeys, Mustafa Shukor, Matthieu Cord, and Jakob Verbeek. Improved baselines for data-efficient perceptual augmentation of llms. *arXiv preprint arXiv:2403.13499*, 2024.

Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL `https://arxiv.org/abs/2408.01800`.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023a.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023b.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=tOd8rSjcWz`.