Learning a Unified Latent Action Space from Videos with Action-centric Cycle Consistency

Supplementary Material

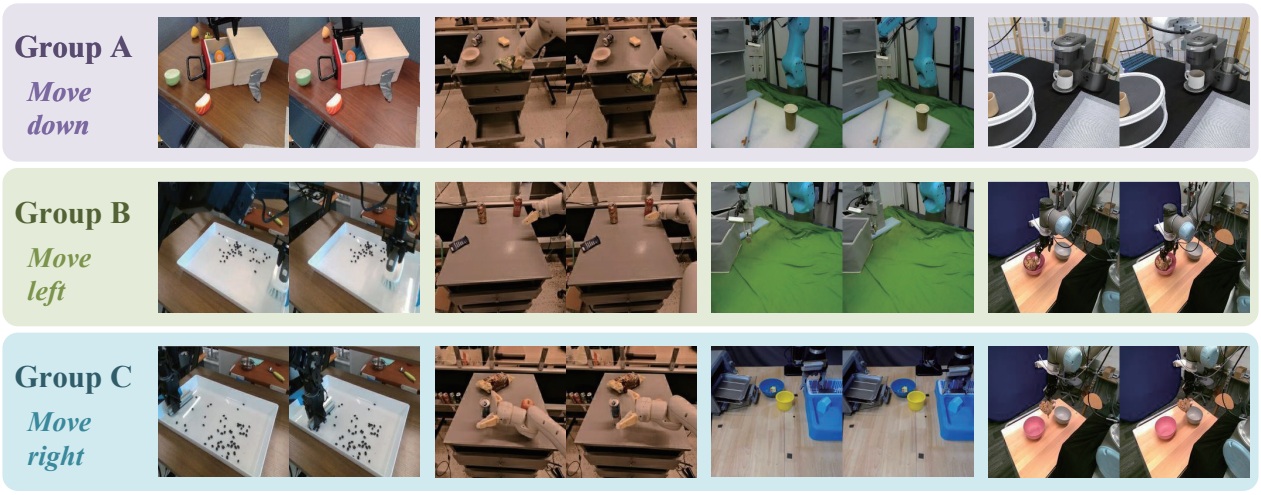

Figure 1. Latent action visualizations. We visualize image pairs sharing identical latent action labels across multiple data sources and embodiments. Each latent action group exhibits semantic consistency across different embodiments.

A. Notations

We provide a summary of the key notations in the full text in the Table 1.

B. Latent Action Visualizations

To intuitively illustrate the latent actions from our approach, we visualize image pairs sharing identical latent action labels across multiple data sources and embodiments. The results, shown in Figure 1, demonstrate that our method learns consistent latent action representations across diverse datasets and embodiments, thereby enabling effective policy pretraining on video datasets.

C. Implementation Details

Following UniVLA [1], our training data comprises three primary categories: robotic manipulation data, navigation data, and human video data. For robotic manipulation, we incorporate a subset from the Open X-Embodiment dataset [2], specifically targeting single-arm end-effector control tasks. The navigation data consists of a subset from the GNM dataset [3], encompassing both indoor and off-road scenarios captured via an ego-view fisheye camera. The human video data are sourced from the Ego4D dataset [4], which provides ego-centric recordings of daily human activities from a first-person perspective. The detailed dataset composition and corresponding mixture weights are presented in Table 2.

For latent action learning with AC3. The encoder utilizes a pre-trained ViT-base DINOv2 model to extract image embeddings, which are subsequently processed through a 12-layer spatial-temporal transformer with 768-dimensional hidden states. The resulting latent action embeddings are discretized via vector quantization using a codebook of size 16. Both the decoder and discriminator employ spatial transformers with identical architectural specifications: 12 layers and 768 dimensions. The latent action buffer is constructed by accumulating the encoded latent actions from the previous 4 batches. The latent action tokenizer is optimized using the AdamW optimizer, with a learning rate of $1e-4$ and weight decay of $1e-2$.

For latent action pretraining on the full dataset, we employ a global batch size of 512 with 32 samples per GPU and apply a constant learning rate of $1e-5$ for $50,000$ optimization steps. For pretraining on the Bridge datasets, we utilize a global batch size of 256 distributed across 8 GPUs with a learning rate of $2e-5$ for $30,000$ optimization steps.

During pretraining, we exclude actions and proprioceptive states from the robot datasets, utilizing only visual frames and textual instructions from each episode. We jointly optimize all components of our generalist policy end-to-end, including the visual encoders, the large language model (LLM) backbone, and the token prediction head.

Table 1. Notation Table

Notation	Description
Frame and Observation Variables	
o_t	Current observation/frame at time t
o_{t+H}	Future observation/frame at time $t+H$
$\hat{o}_{t+H}$	Reconstructed future frame
o_c	video frame from dataset (for cycle consistency)
$\hat{o}_g$	Generated frame from o_c (for cycle consistency)
Latent Action Representations	
z_t^e	Encoded latent action embedding at time t (before quantization)
z_t^q	Quantized latent action at time t
z_s^q	Sampled latent action from latent action buffer
$\hat{z}_s^q$	Predicted sampled latent action
$\hat{z}_s^e$	Latent action embedding of $\hat{z}_s^q$ (before quantization)
$\mathcal{Z}$	Latent action buffer
Codebook and Quantization	
e_k	k-th codebook vector
y_k	One-hot target vector for codebook index
$d(\cdot,\cdot)$	Distance metric for codebook matching
Model Components	
$\mathcal{E}$	Encoder network
$\mathcal{D}$	Decoder network
Ψ	Discriminator network
π_ϕ	Policy model with parameters ϕ
Action Tokens and Policy	
$z_{<i}^q$	Prefix of latent action tokens (tokens before index i)
$\hat{z}_i^q$	Predicted latent action token at position i
N	Total length of action token sequence
`LACT_k`	Specialized vocabulary token for k-th latent action
`ACT`	Action query tokens for robot action prediction
Loss Functions	
$\mathcal{L}_C$	Cycle consistency loss
$\mathcal{L}_{\text{GAN}}^{\Psi}$	Adversarial loss for discriminator
$\mathcal{L}_{\text{GAN}}^{\mathcal{D}}$	Adversarial loss for decoder
$\mathcal{L}_\pi$	Latent action pre-training loss

Table 2. We utilize the training data mixture from UniVLA [1], covering datasets from the OXE [2], GNM [3] and Ego4D [4].

Training Dataset Mixture	
Fractal [5]	13.9%
Kuka [6]	6.3%
Bridge [7]	6.8%
Taco Play [8]	3.5%
Jaco Play [9]	0.6%
Berkeley Cable Routing [10]	0.3%
Roboturk [11]	2.8%
Viola [12]	1.1%
Berkeley Autolab UR5 [13]	1.4%
Toto [14]	2.4%
Language Table [15]	5.2%
Stanford Hydra Dataset [16]	5.3%
Austin Buds Dataset [17]	0.3%
NYU Franka Play Dataset [18]	1.0%
Furniture Bench Dataset [19]	2.9%
UCSD Kitchen Dataset [20]	<0.1%
Austin Sailor Dataset [21]	2.6%
Austin Sirius Dataset [22]	2.0%
DLR EDAN Shared Control [23]	0.1%
IAMLab CMU Pickup Insert [24]	1.1%
UTAustin Mutex [25]	2.6%
Berkeley Fanuc Manipulation [26]	0.9%
CMU Stretch [27]	0.2%
BC-Z [28]	8.8%
FMB Dataset [29]	8.4%
DobbE [30]	1.7%
RECON [31]	8.9%
CoryHall [31]	2.3%
SACSoN [31]	3.5%
Ego4D [31]	3.0%

D. Discussions

Spatial information. While our CycleMimic exhibits promising performance, the latent actions are derived from 2D image sequences, thereby lacking explicit three-dimensional spatial information that could be crucial for complex manipulation tasks. This limitation opens several promising avenues for future work. One direction involves integrating depth estimation foundation models [32, 33] to provide geometric priors during action encoding. Alternatively, end-to-end visual geometry grounded transformers [34] could be incorporated to directly embed spatial geometric cues within the learned latent action representations, potentially enabling more spatially-aware action predictions.

Training data. Our current training paradigm, consistent with prior work, relies primarily on robot manipulation datasets across different embodiments. However, the availability of large-scale human demonstration datasets presents a compelling opportunity to enhance the generalization capability of latent action tokenizers. By training on human activities, the model could learn to generate latent action annotations for the vast repository of human activity videos available on the internet. This would represent a step toward bridging the gap between internet-scale video data and embodied AI training, potentially unlocking more diverse behavioral data for policy learning.

References

[1] Qingwen Bu et al. "Univla: Learning to act anywhere with task-centric latent actions". In: *arXiv preprint arXiv:2505.06111* (2025).

[2] Abhishek Padalkar et al. "Open x-embodiment: Robotic learning datasets and rt-x models". In: *arXiv preprint arXiv:2310.08864* (2023).

[3] Dhruv Shah et al. “Gnm: A general navigation model to drive any robot”. In: *arXiv preprint arXiv:2210.03370* (2022).

[4] Kristen Grauman et al. “Ego4d: Around the world in 3,000 hours of egocentric video”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18995–19012.

[5] Anthony Brohan et al. “Rt-1: Robotics transformer for real-world control at scale”. In: *arXiv preprint arXiv:2212.06817* (2022).

[6] Dmitry Kalashnikov et al. “Scalable deep reinforcement learning for vision-based robotic manipulation”. In: *Conference on robot learning*. PMLR. 2018, pp. 651–673.

[7] Homer Rich Walke et al. “Bridgedata v2: A dataset for robot learning at scale”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 1723–1736.

[8] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. “Grounding language with visual affordances over unstructured data”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 11576–11582.

[9] Shivin Dass et al. “Clvr jaco play dataset, 2023”. In: *URL https://github. com/clvrai/clvr_jaco_play_dataset* ().

[10] Jianlan Luo et al. “Multistage cable routing through hierarchical imitation learning”. In: *IEEE Transactions on Robotics* 40 (2024), pp. 1476–1491.

[11] Ajay Mandlekar et al. “Roboturk: A crowdsourcing platform for robotic skill learning through imitation”. In: *Conference on Robot Learning*. PMLR. 2018, pp. 879–893.

[12] Yifeng Zhu et al. “Viola: Imitation learning for vision-based manipulation with object proposal priors”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 1199–1210.

[13] Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. *Berkeley UR5 demonstration dataset*. https://sites.google.com/view/berkeley-ur5/home.

[14] Gaoyue Zhou et al. “Train offline, test online: A real robot learning benchmark”. In: *arXiv preprint arXiv:2306.00942* (2023).

[15] Corey Lynch et al. “Interactive language: Talking to robots in real time”. In: *IEEE Robotics and Automation Letters* (2023).

[16] Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. “Hydra: Hybrid robot actions for imitation learning”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 2113–2133.

[17] Yifeng Zhu, Peter Stone, and Yuke Zhu. “Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation”. In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 4126–4133.

[18] Zichen Jeff Cui et al. “From play to policy: Conditional behavior generation from uncurated robot data”. In: *arXiv preprint arXiv:2210.10047* (2022).

[19] Minho Heo et al. “Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation”. In: *The International Journal of Robotics Research* 44.10-11 (2025), pp. 1863–1891.

[20] Ge Yan, Kris Wu, and Xiaolong Wang. *ucsd kitchens Dataset*. Aug. 2023.

[21] Soroush Nasiriany et al. “Learning and retrieval from prior data for skill-based imitation learning”. In: *arXiv preprint arXiv:2210.11435* (2022).

[22] Huihan Liu et al. “Robot learning on the job: Human-in-the-loop autonomy and learning during deployment”. In: *The International Journal of Robotics Research* 44.10-11 (2025), pp. 1727–1742.

[23] Gabriel Quere et al. “Shared control templates for assistive robotics”. In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 1956–1962.

[24] Saumya Saxena, Mohit Sharma, and Oliver Kroemer. “Multi-resolution sensing for real-time control with vision-language models”. In: *2nd Workshop on Language and Robot Learning: Language as Grounding*. 2023.

[25] Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. “Mutex: Learning unified policies from multimodal task specifications”. In: *arXiv preprint arXiv:2309.14320* (2023).

[26] Xinghao Zhu et al. *Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200iD robot*. 2023.

[27] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. “Structured world models from human videos”. In: *arXiv preprint arXiv:2308.10901* (2023).

[28] Eric Jang et al. “Bc-z: Zero-shot task generalization with robotic imitation learning”. In: *Conference on Robot Learning*. PMLR. 2022, pp. 991–1002.

[29] Jianlan Luo et al. “Fmb: a functional manipulation benchmark for generalizable robotic learning”. In: *The International Journal of Robotics Research* 44.4 (2025), pp. 592–606.

[30] Nur Muhammad Mahi Shafiullah et al. “On bringing robots home”. In: *arXiv preprint arXiv:2311.16098* (2023).

[31] Alexander Khazatsky et al. “Droid: A large-scale in-the-wild robot manipulation dataset”. In: *arXiv preprint arXiv:2403.12945* (2024).

[32] Lihe Yang et al. “Depth anything: Unleashing the power of large-scale unlabeled data”. In: *arXiv preprint arXiv:2401.10891* (2024).

[33] Sili Chen et al. “Video depth anything: Consistent depth estimation for super-long videos”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 22831–22840.

[34] Jianyuan Wang et al. “Vggt: Visual geometry grounded transformer”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 5294–5306.