CoMo: Learning Continuous Latent Motion from Internet Videos for Scalable Robot Learning Supplementary Material

Jiange Yang[1] Yansong Shi[2,3] Haoyi Zhu[2,3] Mingyu Liu[2,4]
Kaijing Ma[2,5] Yating Wang[2,6] Gangshan Wu[1] Tong He[2] Limin Wang[1,2,✉]
[1]Nanjing University, [2]Shanghai AI Lab, [3]University of Science and Technology of China
[4]Zhejiang University, [5]Fudan University, [6]Tongji University
jiangeyang.jgy@gmail.com, lmwang@nju.edu.cn

A. Appendix

A.1. Real-World Experiments Details

In this section, we detail the specifics of our real-world experiments. Specifically, our experiments setup is illustrated in Fig. 1, which comprises a single Franka Emika Research 3 robot arm, equipped with a UMI [2] gripper, and utilizes a statically positioned RealSense D435 camera (with a resolution of 640×480 pixels) from a third-person view to acquire real-time RGB visual observations. Following publicly available code[1], we employ a 3D mouse for teleoperation data collection. The robot system operates at 20 Hz (moderately reduced from the native 100 Hz control frequency to balance training efficiency and motion continuity), with actions defined as relative end-effector pose changes in SE(3) space (3D position change + quaternion orientation change + gripper state).

For five real-world tasks we evaluated—picking up corresponding toy and placing it into the basket, opening the drawer, closing the drawer, inserting the bread into the container, and pouring the balls into the basket—they respectively require the robot arm to perform basic picking-and-placing, fine-grained and contact-rich opening, contact-rich closing, fine-grained picking-inserting, and picking-pouring capability. During evaluation, the initial pose of the robot arm was set to a fixed home position. The initial poses of the objects to be interacted with were significantly varied. A special case is the opening-drawer and closing-drawer task, where adhesive was applied to the bottom of the drawer to mitigate significant sliding during opening and closing. Consequently, in this task, the placement pose of the drawer was slightly perturbed, within a range of approximately 8 cm in the lateral and longitudinal directions.

As for the policy of our real-world experiments, we adopt a diffusion-based policy architecture. Specific training and architecture details can be found in Section A.3. Finally, we jointly train the policy using collected robot data and human hand video data labeled with the corresponding latent motion IDM.

A.2. CoMo Details

In this section, we describe the specifics of our CoMo. As shown in Tab. 1, we report the training and architectural details of our CoMo. We aim to learn a generalizable latent motion IDM that can extract latent motion representing any form of inter-frame changes. To this end, we uniformly sample a total of 120,000 videos from SAM-V [19], EgoVid [20], and Droid [7], with each dataset contributing 40,000 videos. These datasets collectively cover both egocentric and fixed-camera viewpoints, and encompass a wide range of motions, including those of robotic arms, humans, and various objects in the wild. Importantly, all of our baselines employ the same model architecture, training data, and hyperparameters as CoMo, which ensures the strict fairness of our comparisons.

Specifically, for the discrete latent motion baseline, there is a trade-off regarding the choice of codebook size. A larger codebook size typically enables more comprehensive capture of motion information, but also increases the risk of encoding action-irrelevant background noise. Conversely, a smaller codebook size may limit the captured motion details but reduces such noise. In our experiments with the discrete latent motion baseline, we compare a codebook size of 8 (following LAPA [22]) in the LIBERO and real-world settings, and a codebook size of 128 (following Moto-GPT [1]) in CALVIN. In all cases, the results consistently demonstrate the superiority of our CoMo.

A.3. Diffusion-based Policy Details

In this section, we detail our unified diffusion-based policy. We primarily implement the diffusion-based policy

✉Corresponding author.
[1]https://github.com/UT-Austin-RPL/deoxys_control

Figure 1. **The real-world Franka robot arm experiments hardware platform.**

for the LIBERO [10] simulation and real-world experiments. Specifically, we jointly learn the unified diffusion-based policy from video data with pseudo action labels constructed using the corresponding latent motion IDM, and continuous robot action data.

In Tab. 2, we report the training and architectural details of our diffusion-based policy. Specifically, we employ BERT [3] and ViT [4] to extract language instructions and visual observations features, respectively. Following RDT-1B [11], we utilize a more scalable DiT [16] block as the backbone. The extracted language and visual features are incorporated as conditioning through cross-attention layers within the DiT block. To perform joint learning of action-less video data and robot data within a unified policy model, we construct two sets of MLP networks to map latent motion and robot actions into a shared embedding space, and back to their respective original spaces. In the training phase, we adopt the DDPM scheduler with a glide cosine scheduling scheme (specifically, the squaredcos cap v2 variant) across a diffusion process of 1000 steps. Conversely,

Table 1. The training and architectural hyperparameters for our CoMo learning.

Hyperparameter	Value
CoMo training	
Optimizer	AdamW [9]
Base learning rate	0.0001
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.99$
Effective batch size	256
Total training steps	50,000
Frame interval on SAM-V [19]	10
Frame interval on EgoVid [20]	10
Frame interval on Droid [7]	20
Inverse dynamics Model	
Feature extractor	MAE [5] ViT-L
Codebook size of discrete baseline	8 and 128
Number of motion queries	8
Latent motion embedding dimensionality	16
#layers	4
#MHSA heads	12
Hidden dim	768
Forward dynamics Model	
#layers	12
#MHSA heads	12
Hidden dim	768

for inference, we leverage the DPM-Solver++ [12] in conjunction with an analogous glide cosine scheduler, albeit with a substantially reduced sampling budget of 5 steps. Finally, to capture the temporal dependencies of actions and ensure real-time dynamic adaptability during policy execution, we set an action / motion chunk size of 8 in both the training and inference phases.

A.4. Auto-regressive based Policy Details

In this section, we detail the specifics of our auto-regressive based policy, as shown in Tab. 3. We primarily implement this policy for the CALVIN [13] simulation environment experiments. Specifically, we employ T5 [18] and ViT [4] to extract token-level textual and visual features, respectively. Following [1, 8], we adopt a GPT-style [17] auto-regressive backbone and append two additional MLP networks at the output layer to predict continuous robot actions and latent motion separately using the MSE loss. For the discrete latent motion baseline, in accordance with Moto-GPT [1], we utilize the cross-entropy loss function to optimize latent motion prediction. Additionally, we incorporate an action MLP network, consistent with the continuous approach, and employ an MSE loss to learn the robot action.

Specifically, for motion prediction, we auto-regressively predict latent motion with a chunk size of 2. For action prediction, we parallelly decode actions with a chunk size of 5 based on a set of learnable action query tokens. Furthermore, to ensure a fair comparison with the discrete baseline in Moto-GPT [1], we first perform a round of latent motion prediction pre-training using action-less video data before conducting joint training on robot action data and action-less video data.

A.5. FDM future frame prediction visualization

In this section, we present further visualizations of FDM future frame predictions to qualitatively assess the advantages of CoMo compared to the naïve continuous baseline, as shown in Fig. 2. Specifically, given two frames from a prompt video, we extract the latent motion between them. This extracted motion is then used to predict the subsequent frame via FDM in a new environment. The red rectangles highlight that the naïve continuous baseline tends to incorporate significant static background noise from the prompt video. In contrast, our CoMo effectively avoids this issue. Notably, as indicated by the orange rectangles, CoMo produces more precise latent motion representations, resulting in predictions that more accurately reflect the fine-grained foreground motions present in the prompt video.

A.6. Dual-arm and humanoid MSE results details

In the main text, we report ablation results of MSE on more complex robotic platforms, including dual-arm and humanoid robots equipped with dexterous hands. Specifically, for the dual-arm robot, we use RoboTwin [14] dataet,

Table 2. The training and architectural hyperparameters for our diffusion-based policy learning.

Hyperparameter	**Value**
Diffusion-based policy training	
Optimizer	AdamW [9]
Base learning rate	0.0005
Effective batch size	256
Total training epochs	100
Diffusion-based policy architecture	
Vision feature extractor	DINOv2 [15] ViT-B [4]
Language feature extractor	BERT [3]
#layers	12
#MHSA heads	16
Hidden dim	768
Action / motion chunk size	8
Action projector	(7, 768)
Latent motion projector	(128, 768)
Action head	(768, 7)
Latent motion head	(768, 128)
Noise scheduler	
Type	DDPM [6]
Prediction type	sample
Training step number	1000
Sampling step number	5
Solver	DPM-Solver++ [12]

Table 3. The training and architectural hyperparameters for our auto-regressive based policy learning.

Hyperparameter	**Value**
Auto-regressive based policy training	
Optimizer	AdamW [9]
Base learning rate	0.0005
weight decay	0.0001
Effective batch size	512
Total training epochs	20
Auto-regressive based policy architecture	
Vision feature extractor	MAE [5] ViT-B [4]
Language feature extractor	T5 [18]
#layers	12
#MHSA heads	12
Hidden dim	768
Action chunk size	5
Motion chunk size	2

where the action space consists of the absolute joint states of both arms (Aloha AgileX), totaling 14 dimensions. For the humanoid robot platform, we utilize the early open-source EgoVLA [21] dataset, which contains both human motion capture data and humanoid robot data. The EgoVLA adopts a shared action space for humans and humanoid robots, comprising the absolute wrist pose and MANO hand parameters, for a total of 128 dimensions. Overall, the experimental results underscore the effectiveness of CoMo as a unified action space for cross-embodiment data. Notably, it demonstrates robust versatility across both relative and absolute action spaces, particularly within high-dimensional contexts that require the capture of fine-grained motions.

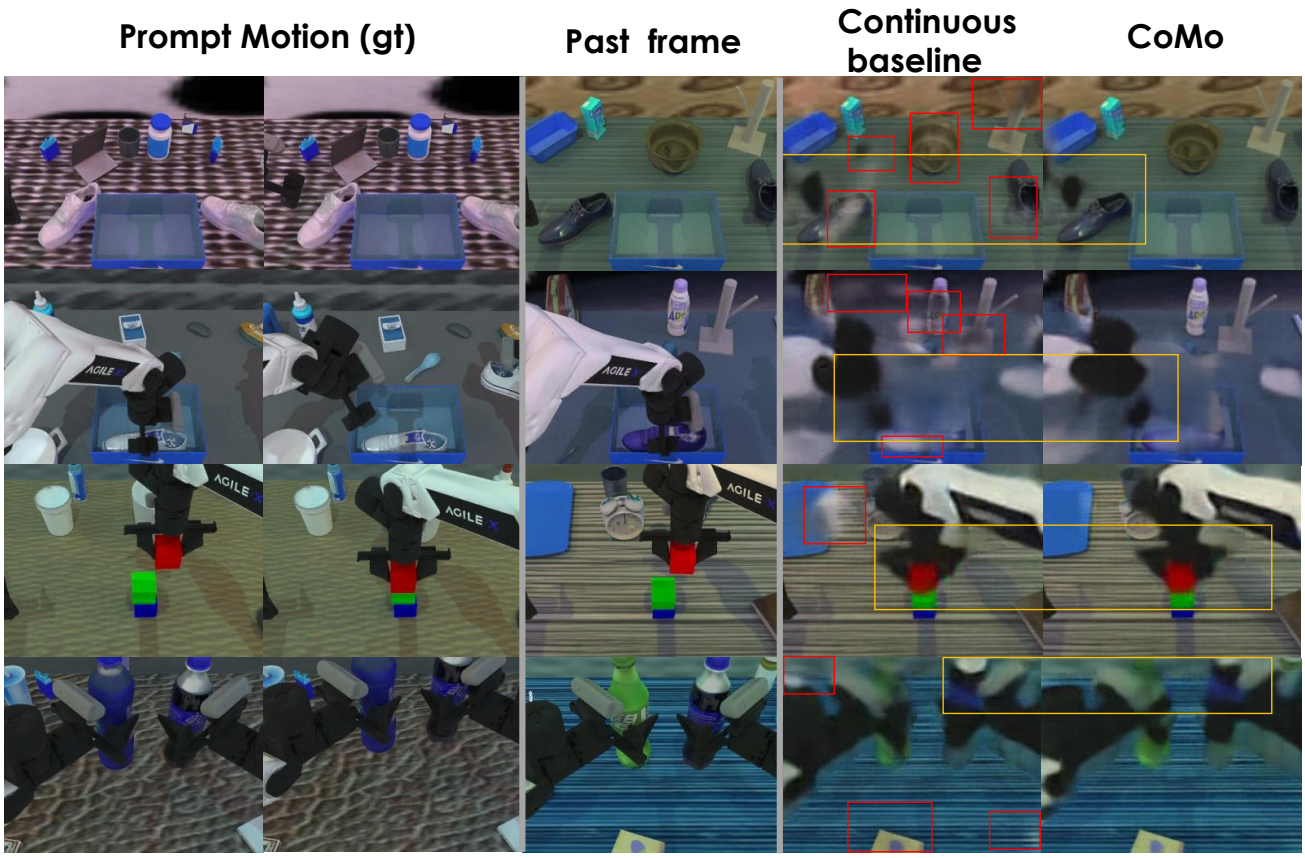

Figure 2. **The FDM future frame prediction visualization.**

References

[1] Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19752–19763, 2025. 1, 3

[2] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2, 4

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 3, 4

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3, 4

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[7] 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. 1, 3

[8] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 3

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for

stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3, 4

[10] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 2

[11] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: a diffusion foundation model for bimanual manipulation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2

[12] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 3, 4

[13] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3): 7327–7334, 2022. 3

[14] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). In *European Conference on Computer Vision*, pages 264–273. Springer, 2024. 3

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[16] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3, 4

[19] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 3

[20] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*, 2024. 1, 3

[21] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 4

[22] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Yuchen Lin, et al. Latent action pretraining from videos. In *ICLR 2025*, pages 90629–90655. International Conference on Learning Representations, ICLR, 2025. 1