Humanoid Bimanual Dexterous Manipulation Driven by Egocentric Video

Hao Huang[1,3], Geeta Chandra Raju Bethala[1,3], Wentian Zhao[4], Shuaihang Yuan[1,2,3]
Congcong Wen[1,3], Mengyu Wang[5], Anthony Tzes[2,3], and Yi Fang[1,2,3]

***Abstract*—Bimanual dexterous manipulation lies at the core of human-level interaction with the physical world, enabling coordinated, contact-rich behaviors that single-arm systems cannot replicate. However, learning such skills for humanoid robots remains challenging: teleoperation—the dominant data-collection paradigm—requires expert operators and real-time execution, limiting the scale, diversity, and naturalness of available demonstrations. We introduce a label-free video-to-policy framework that reduces the need for teleoperation demonstrations by pretraining bimanual manipulation skills from large-scale, in-the-wild egocentric videos and then fine-tuning on a handful of teleoperation demonstrations. Our approach contains two key components: (1) an automatic wrist-finger pose extraction pipeline that reconstructs metrically consistent 3D wrist and fingertip trajectories from unconstrained egocentric videos by jointly recovering scene geometry and MANO-based hand kinematics; and (2) a temporal-scale world model that unifies visual dynamics prediction with action generation through temporal-scale autoregressive forecasting and sparse attention regularization. Experiments on real-world task evaluations using a Unitree H1-2 humanoid with dexterous hands demonstrate the ability to generalize across objects, backgrounds, and tasks. Our code is available at: https://github.com/hhuang-code/bi-dex-egovid.**

***Index Terms*—Dexterous Manipulation, MANO Hand, Scene Reconstruction, World Model.**

Fig. 1. (a) Internet-sourced egocentric videos showcasing diverse dexterous human manipulation. (b) Data collection via human demonstrations and robot teleoperation for manipulation tasks. (c) Policy rollout on a humanoid robot performing dexterous manipulation. (d) Our teleoperation hardware setup, including an Apple Vision Pro headset and a stereo camera.

I. Introduction

Recent advancements in robotics have begun to tackle the complex challenge of bimanual dexterous manipulation by drawing inspiration from human capabilities [1]–[6]. However, these approaches typically depend on either reinforcement learning in specialized simulations that require extensive trials or imitation learning guided by expert demonstrations. A key challenge in imitation learning is the need to collect a large number of expert demonstrations, which is often accomplished using a teleoperation system. While teleoperation systems such as GELLO [7], ALOHA [8], and immersive VR-based interfaces [9] enable the collection of high-quality demonstrations for complex manipulation tasks, they are costly, and the data collection process is notoriously time-consuming. Moreover, these teleoperation systems rely on trained operators who must be familiar with both the tasks and the teleoperation interface. As a result, teleoperated datasets are typically constrained in size and diversity.

[1]Embodied AI and Robotics (AIR) Lab, NYUAD.
[2]NYUAD Center for Artificial Intelligence and Robotics (CAIR).
[3]New York University Abu Dhabi.
[4]Adobe Inc.
[5]Harvard Ophthalmology AI Lab, Harvard University.

Learning from videos has recently emerged as a promising alternative to traditional teleoperation-based data collection in robot learning [10], [11]. The Internet already contains vast amounts of manipulation footage spanning diverse environments, objects, and task variations, while egocentric recordings from wearable devices offer natural first-person perspectives, *i.e.*, egocentric videos, that align well with robot visual perception [12]–[14]. Such video sources are orders of magnitude larger than teleoperation datasets. Moreover, videos inherently capture subtle aspects of human dexterity, *e.g.*, coordinated bimanual motions and adaptive finger articulations, that are difficult to reproduce consistently through teleoperation. Consequently, large-scale video-based learning is increasingly viewed as a foundation for scalable robot learning, with the potential to enable generalization at a scope closer to human skill acquisition than teleoperation alone.

In this work, we propose a label-free video-to-policy framework for bimanual dexterous manipulation that eliminates reliance on costly teleoperation and expert supervision. Our approach introduces an automatic egocentric hand-pose extraction pipeline that reconstructs metrically consistent wrist–fingertip trajectories from in-the-wild egocentric videos, along with a temporal-scale world model that jointly learns visuomotor dynamics and action generation within a shared

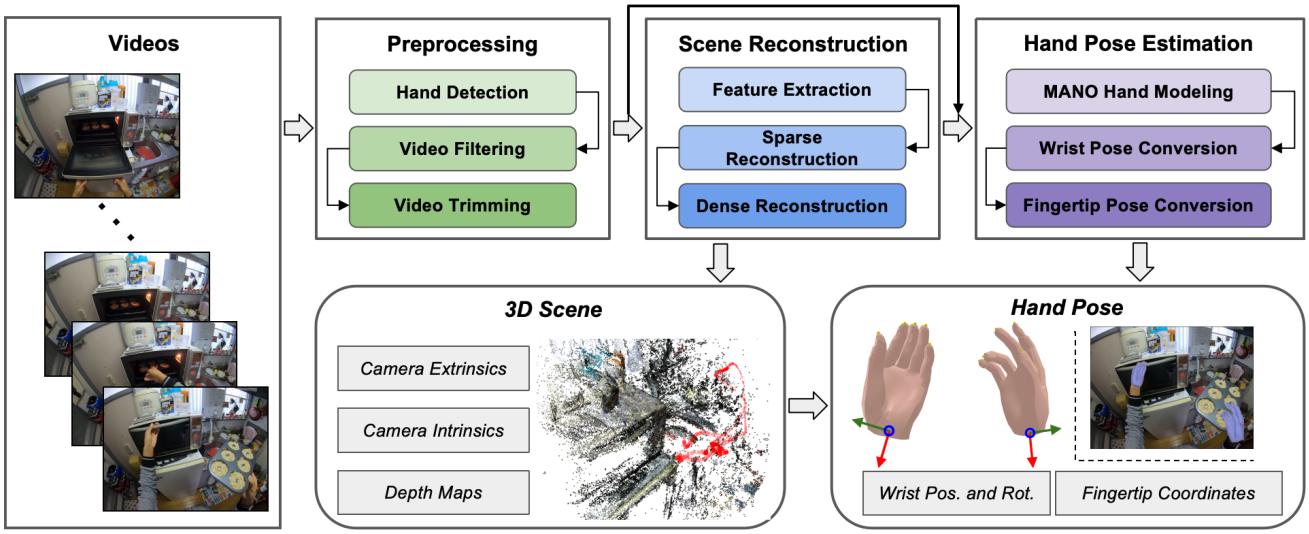

Fig. 2. Pipeline for processing raw egocentric videos to concurrently reconstruct the 3D scene and estimate metrically consistent 3D hand poses. The red sequence of prisms in the 3D scene represents the movement of the egocentric camera.

human–robot state–action space. Unlike prior methods that depend on manually annotated actions or simplified single-arm control, our model leverages large-scale egocentric video data to learn contact-rich bimanual manipulation directly from visual and pose cues, thereby unifying perception and control through predictive latent representations. Experiments using the Ego4D egocentric dataset [12] and real-world evaluations on a Unitree H1-2 humanoid platform demonstrate that our framework generalizes across different environments, object configurations, and manipulation tasks, narrowing the gap between human and robot dexterity in both in-distribution and out-of-distribution conditions. Our contributions are threefold:

- We present an automatic, label-free pipeline that reconstructs metrically consistent 3D wrist–fingertip poses from in-the-wild egocentric videos for policy learning.
- We introduce a world model with temporal-scale autoregressive prediction and sparse attention regularization to achieve visuomotor consistency.
- We demonstrate real-world performance on bimanual humanoid manipulation tasks, achieving generalization across different tasks and environments.

II. Method

We learn a visuomotor policy that maps stereo images I_t and bimanual hand state s_t to future bimanual actions $\{a_{t+1:t+H}\}_{H\geq 1}$ for contact-rich manipulation with a humanoid robot with a fixed lower body. A major challenge is the lack of expert demonstrations: prior methods [4], [9] usually collect real-world robot data via teleoperation, which is slow, expensive, and limits data scale and diversity; teleoperation can take 6–10$\times$ longer than direct human execution [4]. In contrast, in-the-wild human–object interaction videos are abundant. Recent work has used large-scale videos to train or pretrain robot policies [15]–[26], but these methods either require manually annotated action labels or focus on single-arm policies. We instead learn a policy for *humanoids with bimanual dexterous hands from unlabeled in-the-wild egocentric videos*, using an automatic pipeline to extract hand poses in the robot's camera frame and a *world model* that learns both manipulation dynamics and the corresponding hand actions.

A. Shared State-Action Space

To bridge the gap between human and robot hands, we use a shared 54D state–action space for both agents, following [4]:

- Head and wrist rotations: The orientations of the head, left wrist, and right wrist are each represented by 6D rotation vectors [27].
- Wrist translations: The positions of the left and right wrists are represented by 3D (x, y, z) vectors.
- Fingertips: The 10 fingertips are represented by 3D (x, y, z) keypoints for five-fingered dexterous hands.

This representation gives $3 \times 6 + 2 \times 3 + 10 \times 3 = 54$ parameters for both the action and state spaces, enabling a single network architecture to process egocentric videos, human demonstrations, and robot teleoperation data.

B. Hand Pose Estimation

The central challenge is to recover metrically consistent and temporally aligned wrist–fingertip states from unconstrained, label-free egocentric videos with unknown cameras, occlusions, and large appearance variations, while preserving the bimanual contact geometry required for dexterous control. To address this, as shown in Figure 2, we design an automatic pipeline that estimates bimanual 3D hand poses from unlabeled in-the-wild egocentric videos collected from the Internet by jointly reconstructing the scene. The pipeline first preprocesses each video to obtain usable clips and then processes each clip through two parallel streams: *scene reconstruction* and *hand pose estimation*. The scene reconstruction

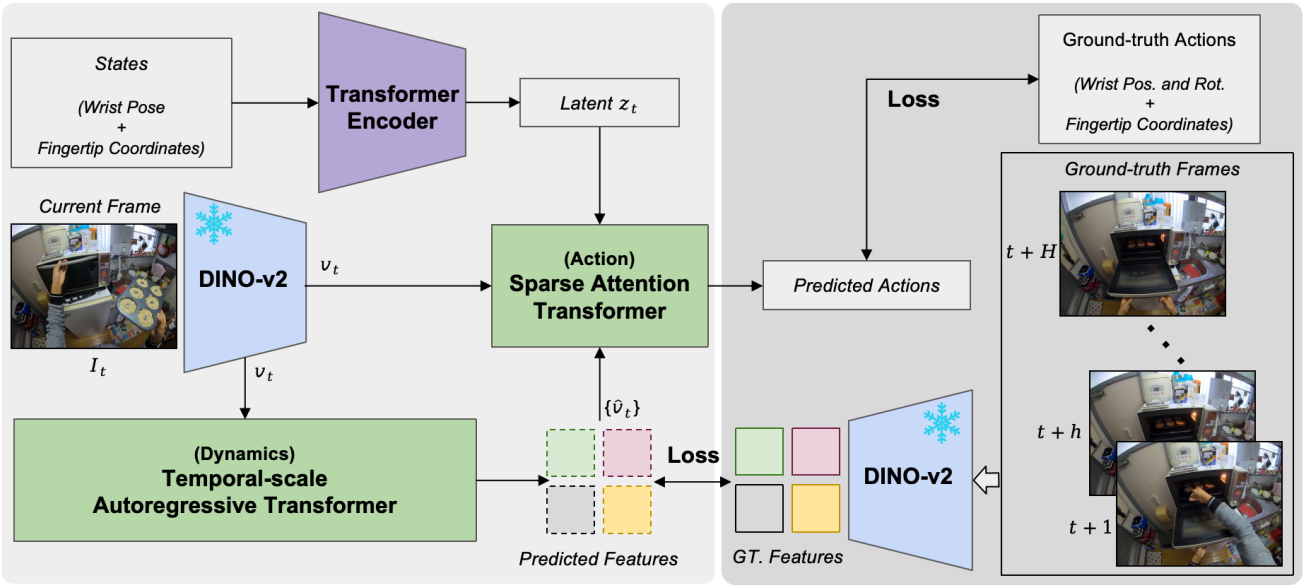

Fig. 3. Overview of our world model: a frozen DINO-v2 encoder encodes current frames, a Transformer encodes hand states, a temporal autoregressive Transformer predicts future visual features, a sparse-attention Transformer predicts bimanual actions, and training jointly matches actions to ground-truth trajectories and aligns predicted features with frozen DINO features from future frames.

stream builds a sparse-to-dense 3D scene and recovers camera intrinsics, extrinsics, and per-frame depth maps. In parallel, the hand-pose estimation stream uses MANO hand modeling [28] to recover wrist poses and fingertip positions, which are then converted to a consistent robot coordinate frame $\mathcal{F}_R$. We define the origin of $\mathcal{F}_R$ as the midpoint between the two eyes, motivated by the fact that both egocentric videos and humanoid observations are captured by head-mounted cameras. See Appendices II-A–II-C for the detailed video preprocessing procedure, camera-motion trimming criterion, scene reconstruction pipeline, and hand-pose recovery formulation.

Using this generic and automatic pipeline, we obtain metrically consistent wrist poses—including 3D positions and orientations in the robot coordinate frame—as well as 3D fingertip coordinates relative to the wrists from arbitrary in-the-wild egocentric videos *without any labels or annotations*. These recovered states serve as supervision for training the bimanual dexterous hand policy described next.

C. World Model

Unlike prior approaches [4], [16], which directly map the current state, optionally with history, to future actions, *i.e.*, $f : s_{t-h}t \geq h \geq 0 \mapsto at + H_{H\geq 0}$, we instead adopt a world model [29]–[33] for action prediction. A world model encourages the policy to learn environment dynamics rather than only imitating actions: in addition to predicting actions, it also forecasts future visual representations, forcing the policy to produce actions that are consistent with the predicted future. Specifically, as shown in Figure 3, at time t, the model takes an egocentric observation I_t and robot state s_t (wrist pose and fingertip coordinates) as input. A Transformer encoder maps s_t to a latent state z_t, while a frozen visual encoder Φ (*i.e.*, DINO-v2 [34]) extracts visual features $v_t = \Phi(I_t)$. Instead of predicting only the next observation, we forecast multiple future observations over a horizon of H steps at intervals of r using a *Temporal-scale Autoregressive Transformer*. Conditioned on the current state, current observation, and predicted future observations, a *Sparse Attention Transformer* then predicts the action chunk. Importantly, future dynamics are predicted in feature space rather than pixel space [30], [31]. See Appendix III-A for the temporal-scale autoregressive formulation; Algorithm 2 and Appendix III-B for the sparse-attention design and regularization details.

The Temporal-scale Autoregressive Transformer improves long-horizon prediction stability by generating future observations progressively from coarser to finer temporal scales, rather than predicting them strictly step by step. The Sparse Attention Transformer uses learnable queries to attend selectively to the most relevant regions in the current and predicted observations for action generation. To encourage compact and diverse attention patterns across heads, we further introduce regularization terms; their definitions are provided in Appendix III-B.

Training jointly optimizes an action objective and a dynamics objective, so that the policy is supervised not only by ground-truth actions but also through future visual consistency. The full objective, including the sparse-attention regularizers, is given in Appendix III-C.

III. Experiments

We evaluate our method on four bimanual dexterous manipulation tasks performed by a Unitree H1-2 humanoid robot with a fixed lower body: box opening, object passing, pick-and-place, and object pushing. The robot receives only egocentric visual input from a head-mounted stereo camera.

Fig. 4. Visualizations of four manipulation tasks performed by a humanoid robot. Each row shows one task. Row 1: Box Opening – the robot unfolds flaps to open a cardboard box. Row 2: Object Passing – the robot picks up a toy, transfers it from the right hand near the left hand, then picks it up with the left hand and places it near the right hand. Row 3: Pick-and-Place – the robot grasps an object from the table and places it inside the box. Row 4: Pushing – the robot pushes an object across the table from right to left. All objects are randomly placed, and demonstrations are performed under varied lighting conditions.

TABLE I

SUCCESS RATE OF EXECUTION ACROSS 10 TRIALS PER TASK. 'V. DATA' DENOTES EGOCENTRIC VIDEO PRETRAINING; 'H. DATA' DENOTES HUMAN DEMONSTRATIONS USED FOR CO-TRAINING WITH ROBOT TELEOPERATED DEMONSTRATIONS.

Method	V. Data	H. Data	Box Opening		Object Passing		Pick and Place		Object Pushing		Overall Success	
			ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
HAT [4]	✗	✗	0.80	0.70	0.30	0.10	0.30	0.10	0.80	0.60	0.55	0.38
Ours	✗	✗	0.90	0.80	0.50	0.20	0.30	0.20	0.90	0.50	0.65	0.43
Ours	✗	✓	0.70	0.50	0.10	0.10	0.30	0.20	0.80	0.40	0.48	0.30
Ours	✓	✗	0.90	0.90	0.60	0.30	0.60	0.40	1.00	0.60	0.78	0.55
Ours	✓	✓	0.90	0.60	0.30	0.10	0.30	0.20	0.70	0.40	0.55	0.33

Detailed descriptions of the video data, hardware platforms, and data collection procedure are provided in Appendix IV.

Experimental Protocol. Demonstrations were recorded across different backgrounds, objects, randomized object placements, and varying positions of both the human and the robot relative to the table. We evaluate policies on the four tasks shown in Figure 4. For each task, we consider both in-distribution (ID) and out-of-distribution (OOD) settings. The ID setting evaluates the learned policy under backgrounds and object arrangements similar to those seen in the robot and human training demonstrations, while the OOD setting tests generalization and robustness under new backgrounds and objects not present in any training demonstrations.

Results. The quantitative results are shown in Table I, revealing two key trends. First, pretraining on large-scale video data consistently improves performance, especially in out-of-distribution (OOD) settings, indicating that diverse egocentric footage provides generalization signals. Second, co-training with human demonstrations does not always help and can slightly reduce performance. This arises because the execution rates and motion characteristics differ between humans and the robot; naively mixing these two sources introduces temporal and dynamic inconsistencies that the policy must reconcile.

IV. CONCLUSION

In this work, we present a label-free video-to-policy framework for bimanual dexterous manipulation that combines automatic egocentric hand-pose extraction with a temporally structured world model for visuomotor prediction. Our approach enables pretraining of contact-rich bimanual skills directly from in-the-wild videos, substantially reducing reliance on large-scale teleoperation demonstrations. Real-world experiments on a humanoid robot demonstrate promising generalization across multiple tasks and environments, suggesting that video-driven learning provides a scalable and practical path toward more versatile humanoid manipulation.

ACKNOWLEDGMENT

Authors appreciate the support provided by the NYUAD Center for Artificial Intelligence and Robotics (CAIR), funded by Tamkeen under the NYUAD Research Institute Award CG010.

REFERENCES

[1] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang, "Towards human-level bimanual dexterous manipulation with reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5150–5163, 2022.

[2] Y. Lin, A. Church, M. Yang, H. Li, J. Lloyd, D. Zhang, and N. F. Lepora, "Bi-touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 9, pp. 5472–5479, 2023.

[3] M. Drolet, S. Stepputtis, S. Kailas, A. Jain, J. Peters, S. Schaal, and H. B. Amor, "A comparison of imitation learning algorithms for bimanual manipulation," *IEEE Robotics and Automation Letters*, 2024.

[4] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen *et al.*, "Humanoid policy ˜ human policy," in *Conference on Robot Learning*, 2025.

[5] K. Li, P. Li, T. Liu, Y. Li, and S. Huang, "Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6991–7003.

[6] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. J. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," in *IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 16 923–16 930.

[7] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2024, pp. 12 156–12 163.

[8] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *Robotics: Science and Systems*, 2023.

[9] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," in *Conference on Robot Learning*. PMLR, 2025, pp. 2729–2749.

[10] C. Eze and C. Crick, "Learning by watching: A review of video-based learning approaches for robot manipulation," *IEEE Access*, 2025.

[11] R. McCarthy, D. C. Tan, D. Schmidt, F. Acero, N. Herr, Y. Du, T. G. Thuruthel, and Z. Li, "Towards generalist robot learning from internet video: A survey," *Journal of Artificial Intelligence Research*, vol. 83, 2025.

[12] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *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. 18 995–19 012.

[13] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 720–736.

[14] T. Perrett, A. Darkhalil, S. Sinha, O. Emara, S. Pollard, K. K. Parida, K. Liu, P. Gatti, S. Bansal, K. Flanagan *et al.*, "Hd-epic: A highly-detailed egocentric video dataset," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 901–23 913.

[15] K. Shaw, S. Bahl, A. Sivakumar, A. Kannan, and D. Pathak, "Learning dexterity from human hand motion in internet videos," *International Journal of Robotics Research*, vol. 43, no. 4, pp. 513–532, 2024.

[16] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, X. Cheng, R.-Z. Qiu *et al.*, "Egovla: Learning vision-language-action models from egocentric human videos," *arXiv preprint arXiv:2507.12440*, 2025.

[17] H. Bi, L. Wu, T. Lin, H. Tan, Z. Su, H. Su, and J. Zhu, "H-rdt: Human manipulation enhanced bimanual robotic manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 22, 2026, pp. 18 135–18 143.

[18] M. Lepert, J. Fang, and J. Bohg, "Masquerade: Learning from in-the-wild human videos using data-editing," in *Conference on Robot Learning*, 2025.

[19] G. Li, Y. Lyu, Z. Liu, C. Hou, Y. Xu, J. Zhang, and S. Zhang, "H2r: A human-to-robot data augmentation for robot pre-training from videos," in *CVPR Workshop: Synthetic Data for Computer Vision*, 2025.

[20] H. G. Singh, A. Loquercio, C. Sferrazza, J. Wu, H. Qi, P. Abbeel, and J. Malik, "Hand-object interaction pretraining from videos," in *IEEE International Conference on Robotics and Automation*. IEEE, 2025, pp. 3352–3360.

[21] H. Luo, Y. Feng, W. Zhang, S. Zheng, Y. Wang, H. Yuan, J. Liu, C. Xu, Q. Jin, and Z. Lu, "Being-h0: vision-language-action pretraining from large-scale human videos," *arXiv preprint arXiv:2507.15597*, 2025.

[22] H. Chen, Y. Yao, Y. Ye, Z. Xu, H. Bharadhwaj, J. Wang, S. Tulsiani, Z. Erickson, and J. Ichnowski, "Web2grasp: Learning functional grasps from web images of hand-object interactions," *arXiv preprint arXiv:2505.05517*, 2025.

[23] L. Y. Zhu, P. Kuppili, R. Punamiya, P. Aphiwetsa, D. Patel, S. Kareer, S. Ha, and D. Xu, "Emma: Scaling mobile manipulation via egocentric human data," in *CoRL Workshop on Human to Robot: Sensorizing, Modeling, and Learning from Humans*, 2025.

[24] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "Okami: Teaching humanoid robots manipulation skills through single video imitation," in *Conference on Robot Learning*, 2025.

[25] R. Shah, S. Liu, Q. Wang, Z. Jiang, S. Kumar, M. Seo, R. Martín-Martín, and Y. Zhu, "Mimicdroid: In-context learning for humanoid robot manipulation from human play videos," *arXiv preprint arXiv:2509.09769*, 2025.

[26] S. Ye, J. Jang, B. Jeon, S. J. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin *et al.*, "Latent action pretraining from videos," in *International Conference on Learning Representations*, 2025.

[27] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.

[28] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: modeling and capturing hands and bodies together," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 1–17, 2017.

[29] Y. LeCun, "A path towards autonomous machine intelligence," *Open Review*, vol. 62, no. 1, pp. 1–62, 2022.

[30] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, "Dino-wm: World models on pre-trained visual features enable zero-shot planning," in *International Conference on Machine Learning*, 2025.

[31] F. Baldassarre, M. Szafraniec, B. Terver, V. Khalidov, F. Massa, Y. LeCun, P. Labatut, M. Seitzer, and P. Bojanowski, "Back to the features: Dino as a foundation for video world models," *arXiv preprint arXiv:2507.19468*, 2025.

[32] T. Yin, Z. Mei, T. Sun, L. Zha, E. Zhou, J. Bao, M. Yamane, O. Sho, and A. Majumdar, "Womap: World models for embodied open-vocabulary object localization," in *RSS Workshop: Mobile Manipulation: Emerging Opportunities & Contemporary Challenges*, 2025.

[33] J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang *et al.*, "Worldvla: Towards autoregressive action world model," *arXiv preprint arXiv:2506.21539*, 2025.

[34] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

[35] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," in *RSS Workshop: Data Generation for Robotics*, 2024.

[36] Y. Chen, C. Wang, Y. Yang, and K. Liu, "Object-centric dexterous manipulation from human motion data," in *Conference on Robot Learning*, 2024.

[37] Y. Chen, Y. Geng, F. Zhong, J. Ji, J. Jiang, Z. Lu, H. Dong, and Y. Yang, "Bi-dexhands: Towards human-level bimanual dexterous manipulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2804–2818, 2023.

[38] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 570–587.

[39] C. Bao, H. Xu, Y. Qin, and X. Wang, "Dexart: Benchmarking generalizable dexterous manipulation with articulated objects," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21 190–21 200.
[40] J. Ye, K. Wang, C. Yuan, R. Yang, Y. Li, J. Zhu, Y. Qin, X. Zou, and X. Wang, "Dex1b: Learning with 1b demonstrations for dexterous manipulation," in *Robotics: Science and Systems*, 2025.
[41] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong *et al.*, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning*, 2024.
[42] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations*, 2023.
[43] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 791–15 801.
[44] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
[45] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.
[46] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
[47] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.
[48] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *International Conference on Computer Vision*. IEEE, 2011, pp. 2564–2571.
[49] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2. IEEE, 1999, pp. 1150–1157.
[50] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[51] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.
[52] J. M. Coughlan and A. L. Yuille, "Manhattan world: Orientation and outlier detection by bayesian inference," *Neural Computation*, vol. 15, no. 5, pp. 1063–1088, 2003.
[53] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision*. Springer, 2016, pp. 501–518.
[54] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "Wilor: End-to-end 3d hand localization and reconstruction in-the-wild," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 242–12 254.
[55] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *Advances in Neural Information Processing Systems*, vol. 37, pp. 84 839–84 865, 2024.
[56] J. Yuan, H. Gao, D. Dai, J. Luo, L. Zhao, Z. Zhang, Z. Xie, Y. Wei, L. Wang, Z. Xiao *et al.*, "Native sparse attention: Hardware-aligned and natively trainable sparse attention," in *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 23 078–23 097.

APPENDIX

I. Related Work

Dexterous Hand Manipulation. Dexterous hand manipulation is a long-standing challenge in robotics, with recent work advancing data collection, benchmarking, and model design. DexCap [35] presents a portable motion capture system for the scalable acquisition of human hand trajectories, enabling imitation learning of fine-grained skills in real-world settings. Chen *et al*. [36] leverage human motion data to synthesize wrist trajectories adapted to robot embodiments. On the benchmarking side, Bi-DexHands [1], [37] establishes a suite of bimanual manipulation tasks in Isaac Gym to evaluate multi-task and multi-finger policies. DexMV [38] designs a simulation system to evaluate multi-finger robot hands completing complex dexterous manipulation tasks. Extending to articulated objects, DexArt [39] benchmarks dexterous manipulation of articulated objects within a physical simulation environment. Dex1B [40] generates one billion demonstrations with generative models for dexterous grasping and articulation tasks. Despite these advances, existing methods still face challenges in sample efficiency, sim-to-real gaps, and robustness across diverse real-world tasks.

Video-based Policy Learning. Early efforts demonstrate that large, unlabeled human videos can seed manipulation priors. VideoDex [15] shows that passive Internet footage encodes action and physics regularities that can improve robot-hand policy learning. HOP [20] extends this by reconstructing 3D hand–object trajectories from in-the-wild clips to train a robot arm policy. However, direct transfer between human hands in videos and robot hands is impeded by different hand embodiments and the lack of action labels. Latent Action Pretraining [26] addresses label scarcity by pretraining a Vision-Language-Action (VLA) model [41] to predict "latent actions" and later mapping these latent actions to robot controls using small-scale robot manipulation data. In parallel, H2R [19] and Masquerade [18] narrow the embodiment gap by robotizing egocentric human videos: the former composites simulated manipulators, while the latter removes human arms and overlays bimanual robots in videos. H-RDT [17] pretrains a diffusion Transformer [42] for bimanual grippers using videos containing human hands. EgoVLA [16] learns to predict wrist/hand actions from egocentric videos and employs a VLA model for bimanual dexterous manipulation policy learning. Pushing dexterity further, Being-H0 [21] treats the human hand as a "foundation manipulator", standardizing heterogeneous motion sources and tokenizing finger motions for instruction-tuned policies that span embodiments. For single-video open-world imitation, OKAMI [24] adopts object-aware retargeting to let humanoid robots mimic human motions in RGB-D videos, enabling robust, closed-loop humanoid manipulation policy learning without teleoperation. Beyond tabletop settings, EMMA [23] co-trains human mobile data from videos with static robot demonstrations and introduces phase identification to switch between navigation and manipulation, producing reliable mobile manipulation behaviors. MimicDroid [25] reframes video-to-policy transfer as meta-in-context learning from human play, enabling few-shot generalization to novel objects, scenes, and humanoid morphologies. Collectively, learning from video establishes a coherent pathway to scalable dexterous manipulation.

World Model. World models provide a predictive structure for planning in embodied control. JEPA/H-JEPA [29] places a configurable predictive world model at the core of an autonomous agent and argues for self-supervised, hierarchical representations that support gradient-based planning and intrinsic motivation, thereby decoupling perception from control while enabling test-time reasoning. Building on this vision, DINO-WM [30] learns dynamics directly in a pretrained

DINO-v2 [34] latent space and performs test-time model predictive control for zero-shot navigation and manipulation, avoiding pixel reconstruction and task-specific rewards. Extending this principle at scale, DINO-world [31] pretrains a generalist video world model in the DINO-v2 latent space on tens of millions of videos and then fine-tunes on observation–action data pairs so that imagined latent rollouts can score and select actions. Complementary to latent-feature predictors, Navigation World Models [43] treat controllable video generation as planning by adopting a Conditional Diffusion Transformer (CDiT) built upon DiT [44] to simulate candidate trajectories and evaluate goal similarity. For open-vocabulary active perception, WoMAP [32] combines Vision-Language-Model (VLM) proposals with a latent world model trained via a Gaussian Splatting [45] real-to-sim pipeline and reward distillation from open-vocabulary detectors. WorldVLA [33] unifies policy and prediction by interleaving an autoregressive action model with a visual world model, highlighting mutual benefits between action and world modeling.

II. Details of Hand Pose Estimation

A. Video Preprocessing

We apply YOLO-v3 [46], pretrained for hand detection, to detect hands frame by frame in a given video. We then apply a filtering operation: a frame is kept only if the number of detected hands is exactly two. Frames containing fewer or more than two hands, due to occlusion or multiple people, are discarded, yielding a temporally coherent version of the original video that contains valid bimanual hand–object interactions. We then trim the filtered videos into quasi-static video clips by detecting low inter-frame motion via homography geometry [47], as camera movement is unavoidable in these in-the-wild videos but detrimental to policy learning. Let $\{I_t\}_{t=1}^{T}$ be the frames of a given filtered video. For each adjacent pair (I_t, I_{t+1}), we detect keypoints and their corresponding descriptors using the ORB (Oriented FAST and Rotated BRIEF) [48] feature detector[1], find the best correspondences between the descriptors from frame I_t and frame I_{t+1} with a brute-force matcher, and estimate a planar homography H_t from the matched keypoints using RANSAC [50] with a minimum inlier count $n = 20$ and an inlier ratio $r = 0.3$. The homography H_t is a 3×3 matrix that describes the perspective transformation between the two adjacent frames.

We compute a motion score that measures the average pixel displacement induced by camera motion on a small set of anchor points P consisting of the four image corners and the image center:

$$s_t = \frac{1}{|P|} \sum_{p \in P} \|H_t p - p\|_2 \ , \tag{1}$$

where p denotes an anchor point and $H_t p$ denotes its transformed counterpart. A low motion score indicates small camera motion, whereas a high score indicates significant motion. A frame pair is regarded as *static* if $s_t < 10$ pixels. The homography estimate is considered valid only if the number of RANSAC inliers and the inlier ratio exceed the thresholds above; otherwise, we set $s_t = \infty$. We keep the longest contiguous sequence whose length, *i.e.*, the number of *static* frame pairs, exceeds a minimum temporal duration of $L = 90$ frames, *i.e.*, 3 seconds for a video at 30 FPS, as the final trimmed video clip.

[1]SIFT (Scale-Invariant Feature Transform) [49] can achieve more accurate results, but requires much longer processing time.

B. Scene Reconstruction

To estimate 3D hand poses from in-the-wild videos, it is necessary to recover the unknown camera intrinsics and extrinsics. The scene reconstruction stage transforms a trimmed video clip into a dense, gravity-aligned 3D scene, allowing us to estimate camera parameters and depth maps for each frame simultaneously. We first extract affine-invariant SIFT features from each frame. An initial sparse 3D scene is then reconstructed via Structure-from-Motion [51], assuming a single camera with fixed intrinsics across all frames. After the sparse scene is built, we apply the Manhattan-world assumption [52] to automatically rotate the scene so that its $+Z$ axis is antiparallel to gravity. Next, we incrementally register additional frames, triangulate 3D points, and refine the reconstruction through bundle adjustment. Finally, we apply Multi-View Stereo [53] to register all frames and generate a dense point cloud, while computing camera parameters and depth maps for each frame using an iterative propagation-based method.

C. Hand Pose Estimation and Coordinate Conversion

For MANO [28] hand modeling, we first use a YOLO-v3 hand detector to identify 2D bounding boxes for all hands in each frame of the trimmed video. All detected hand regions are then fed into WiLoR [54] to regress MANO hand model parameters [28], including a shape vector $\boldsymbol{\beta} \in \mathbb{R}^{10}$ representing 10 PCA coefficients that capture most interpersonal shape variations (*e.g.*, finger length and palm width), and a pose vector $\boldsymbol{\theta} \in \mathbb{R}^{48}$ parameterized in axis–angle form. The first three elements $\boldsymbol{\theta}_w$ describe the global wrist rotation with respect to the template/rest pose, while the remaining 15×3 parameters $\boldsymbol{\theta}_f$ describe finger joint rotations. WiLoR also outputs the projected 2D coordinates of the wrist and fingertips in each frame. We then transform the wrist rotation $\boldsymbol{\theta}_w$ from the MANO local space to the consistent robot coordinate frame $\mathcal{F}_R$ through a chain of homogeneous transformations. Meanwhile, we compute physically scaled 3D wrist and fingertip positions by integrating the 2D hand coordinates with the precomputed scene geometry, *i.e.*, camera intrinsics, camera extrinsics, and depth maps. Specifically, the 2D pixel coordinates of the wrist and fingertips are projected into the robot coordinate frame $\mathcal{F}_R$ through 2D-to-3D back-projection using the estimated depth maps, as illustrated in the bottom right of Figure 2.

Using the generic and automatic pipeline described above, we obtain metrically consistent wrist poses—3D positions and orientations in the robot coordinate frame—and 3D fingertip

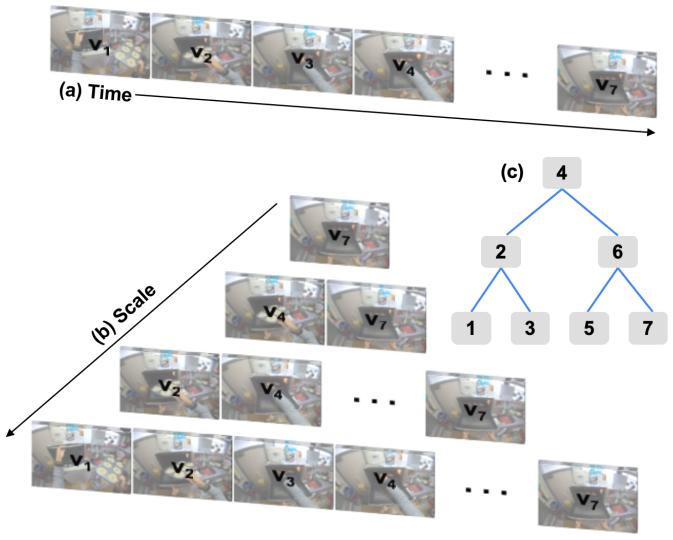

Fig. 5. (a) Conventional autoregression along the temporal dimension. (b) Autoregression along temporal scales. (c) Temporal indices at each scale.

coordinates relative to the wrist from arbitrary in-the-wild egocentric videos *without any labels or annotations*. These data serve as supervision for training the bimanual dexterous hand policy described in the main paper.

Algorithm 1 Autoregressive

$T \leftarrow \frac{H}{r}$
$K \leftarrow V \leftarrow v_t$ ▷ *Omit linear projection for simplicity.*
for $i = 1, \ldots, T$ **do**
  $v_{t+i} \leftarrow \texttt{Attn}(Q_{t+i}, K, V)$ ▷ *Q is learnable query.*
  $K \leftarrow V \leftarrow [V : v_{t+i}]$ ▷ *':' is concatenation.*
end for

III. Details of World Model

A. Temporal-scale Autoregressive Transformer

To predict multiple future observations, a straightforward approach is to employ an autoregressive scheme, as illustrated in panel (a) of Figure 5 and outlined in Algorithm 1. However, a major limitation of autoregressive temporal forecasting lies in the accumulation of prediction errors, which leads to increasingly inaccurate estimates for observations at later time steps. To address this limitation, we draw inspiration from Visual Autoregressive Modeling [55], which generates images by autoregressing across spatial scales. Analogously, we propose to autoregress future observations along temporal scales—that is, to first generate a coarse, sparse observation sequence and then progressively refine it into finer, denser sequences conditioned on the previously generated coarser levels, as illustrated in panel (b) of Figure 5. The time-step indices for the future observations to be generated are determined in two steps: (1) constructing a midpoint binary tree, *i.e.*, a balanced binary tree formed by recursively splitting the interval $[1, T]$ at its midpoints, as shown in panel (c) of Figure 5, where T denotes the prediction horizon; and (2) performing an accumulated breadth-first traversal of the tree across depths, or temporal scales. We refer to this scheme as a *Temporal-scale Autoregressive* scheme, which is outlined in Algorithm 2. This approach yields more stable temporal sequences, particularly at later time steps.

Algorithm 2 Temporal-scale Autoregressive

$T \leftarrow \frac{H}{r}$
$L = \lceil \log_2(T+1) \rceil$ ▷ *Depth of a midpoint binary tree.*
$K \leftarrow V \leftarrow v_t$
$v_{t+T} \leftarrow \texttt{Attn}(Q_{t+T}, K, V)$ ▷ *Jump to last time step.*
$K \leftarrow V \leftarrow [V : V_{t+T}]$
for $l = 1, \ldots, L$ **do**
  $I_l = \texttt{BFS}(\text{tree}) \cup \{T\}$ ▷ *Temporal indices generated by accumulated breadth-first search.*
  $\{v_t\}_{t \in I_l} \leftarrow \texttt{Attn}(\{Q_t\}_{t \in I_l}, K, V)$
  $K \leftarrow V \leftarrow [V : \{v_t\}_{t \in I_l}]$
end for

B. Sparse Attention Transformer

We employ a set of learnable queries that attend to the current and predicted observations to generate actions. Since only specific regions are relevant to action generation, the model attends selectively to relevant parts of the observations rather than to the entire scene. To this end, we adopt the compressed attention and selected attention mechanisms introduced in Native Sparse Attention [56]. The core idea is to group spatially adjacent feature patches into larger blocks, to which *compressed* multi-head attention is applied. *Selected* multi-head attention is then performed only on the patches within the top-k blocks with the highest attention scores during the compressed-attention stage. However, to encourage each attention head to form compact yet diverse attention blocks, we further introduce three regularizers that encourage (1) the top-k attention locations within each head to be spatially concentrated and (2) the top-k locations across different heads to be spatially diverse, yielding less redundant multi-head attention patterns. Specifically, for a given query, each key position is associated with a spatial coordinate $x_k \in \mathbb{R}^d$ (*i.e.*, $d = 1$ for sequences or $d = 2$ for images).

(1) Concentration Loss L_{conc}: After selecting the top-k attention blocks for head h, the soft weights (*i.e.*, attention scores) $w_{h,k}$ define a weighted centroid $\mu_h = \sum_k w_{h,k} x_k$. The concentration loss measures the weighted variance of these selected positions around their centroid:

$$L_{\text{conc}} = \sum_{h=1}^{|H|} \sum_{k=1}^{|K|} w_{h,k} \left\| x_k - \mu_h \right\|^2 . \tag{2}$$

Minimizing L_{conc} encourages each attention head to focus sharply on a spatially compact region.

(2) Overlap Loss $L_{overlap}$: To prevent redundant attention across heads, this term discourages overlap between their selected top-k blocks. Let m_h denote a binary mask marking

TABLE II

EXECUTION SUCCESS RATE ACROSS 10 TRIALS PER TASK WITH DIFFERENT RATIOS OF EGOCENTRIC VIDEOS FOR PRETRAINING. 'V. DATA' MEANS USING EGOCENTRIC VIDEO PRETRAINING; 'H. DATA' MEANS USING HUMAN DEMONSTRATIONS FOR CO-TRAINING WITH ROBOT TELEOPERATED DEMONSTRATIONS.

Method	V. Data	H. Data	Box Opening		Object Passing		Pick and Place		Object Pushing		Overall Success	
			ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
0%	✗	✗	0.90	0.80	0.50	0.20	0.30	0.20	0.90	0.50	0.65	0.43
10%	✓	✗	0.90	0.70	0.50	0.20	0.30	0.10	0.80	0.40	0.63	0.35
50%	✓	✗	1.00	0.80	0.60	0.10	0.50	0.20	0.80	0.60	0.73	0.43
100%	✓	✗	0.90	0.90	0.60	0.30	0.60	0.40	1.00	0.60	0.78	0.55

the top-k indices for head h. For each query, pairwise overlap is computed via dot products between masks:

$$L_{\text{overlap}} = \sum_{\substack{h,h'=1 \\ h \neq h'}}^{|H|} \langle m_h, m_{h'} \rangle \ . \tag{3}$$

Minimizing this loss encourages each head to explore different key subsets, improving representational diversity.

(3) Repulsion Loss $L_{repulse}$: Even if heads attend to different subsets, their centroids may still cluster closely. The repulsion term introduces centroid diversity by penalizing proximity between head-wise centroids:

$$L_{\text{repulse}} = \sum_{\substack{h,h'=1 \\ h \neq h'}}^{|H|} \exp(-\|\mu_h - \mu_{h'}\|^2) \ . \tag{4}$$

This encourages the mean attended block locations of different heads to spread apart.

C. Training objective

Conditioned on $(v_t, z_t, \hat{v}_{t+1:t+H})$, we predict an action chunk $\{a_{t+1:t+H}\} = \pi(v_t, z_t, \hat{v}_{t+1:t+H})$, similar to [4], [8]. Training is multi-task: (i) an action loss L_{act} supervises predicted actions against ground-truth actions using mean absolute error, and (ii) a dynamics loss L_{dyn} aligns predicted features with ground-truth future-frame features using smoothed mean absolute error. The total objective is:

$$\mathcal{L}_{\text{total}} = L_{\text{act}} + \omega L_{\text{dyn}} + \alpha L_{\text{conc}} + \beta L_{\text{overlap}} + \gamma L_{\text{repulse}} \ . \tag{5}$$

The total objective couples kinematics and vision, encouraging the policy to select actions that are consistent with the forecasted visual future.

IV. DATA AND PLATFORM FOR EXPERIMENTS

Video Data. We used the large-scale Ego4D dataset [12], focusing on the 'Hand + Object Interaction' subset, as shown in panel (a) of Figure 1, which captures natural human–object manipulations from a first-person perspective. This subset encompasses diverse scenes, objects, and task variations, making it particularly suitable for training policies that generalize beyond controlled teleoperation data. The videos inherently capture fine-grained hand–object contact, coordinated bimanual movements, and adaptive finger articulations—behaviors that are challenging to reproduce consistently in laboratory-collected demonstrations.

Hardware Platforms. All experiments were conducted on a Unitree H1-2 bimanual humanoid robot with its lower body fixed. The robot is equipped with a pair of 6-DoF Inspire dexterous hands. This configuration provides a total of 26 DoF for the upper body: 7 DoF for each arm and 6 DoF for each hand. For visual perception, we use a single head-mounted ZED stereo camera without third-person or wrist-mounted cameras. The complete hardware setup is illustrated in panels (b) and (d) of Figure 1.

Data Collection. For human demonstrations, we used an Apple Vision Pro (AVP) headset customized with a mount holding a ZED Mini stereo camera. The ZED Mini provided egocentric stereo RGB video, while the AVP enabled tracking of human wrist poses and finger joint angles. To obtain this motion data, we followed [4] and employed the open-source implementation from VisionProTeleop[2]. This setup allowed us to capture synchronized egocentric visual input together with wrist and finger poses during human manipulation demonstrations.

For robot demonstrations, we teleoperated the Unitree H1-2 humanoid robot using the Apple Vision Pro headset. The operator's arm and hand movements were mapped in real time to the robot through OpenTeleVision [9] and AVPTeleop[3]. While the robot executed tasks, a head-mounted ZED stereo camera provided egocentric visual input. To ensure consistency with human demonstrations, we computed the robot's wrist poses and finger joint positions using forward kinematics from its arm and hand joint encoders. The final recorded dataset consisted of synchronized stereo egocentric images together with wrist poses and finger joint positions.

V. ABLATION STUDY

We further investigate the effect of the proportion of egocentric videos used for pretraining, with results reported in Table II. The results show a clear trend: more video data improves policy performance, particularly in out-of-distribution (OOD) evaluations, where generalization is more challenging. Increasing the proportion of video supervision leads to higher

[2]https://github.com/Improbable-AI/VisionProTeleop

[3]https://github.com/unitreerobotics/xr_teleoperate

success rates across all four tasks, with the 100% video-data setting achieving the best overall performance, indicating that large-scale egocentric video pretraining provides strong transferable priors for robust bimanual manipulation.

VI. Detailed Scheme (Algorithm 5) for Temporal-scale Autoregressive Transformer

The following three algorithms describe the detailed steps of the proposed 'Temporal-scale Autoregressive' scheme outlined in Algorithm 2.

Algorithm 3 BUILD_TREE(L, R): Construct Balanced Midpoint Binary Tree

Require: Integers L, R with $L \leq R$
Ensure: Root node of a midpoint binary tree covering interval $[L, R]$

if $L > R$ **then**
  return NULL
end if
$m \leftarrow \left\lfloor \frac{L+R}{2} \right\rfloor$
Create a new node u with u.index $\leftarrow m$
u.left $\leftarrow$ BUILD_TREE$(L, m-1)$
u.right $\leftarrow$ BUILD_TREE$(m+1, R)$
return u

Algorithm 4 BFS_LEVELS(r): Breadth-First Search on a Tree

Require: Root node r (may be NULL)
Ensure: List $TreeLevels$, each entry in $TreeLevels$ is the set of node indices at that BFS depth.

if $r =$ NULL **then**
  return empty list
end if
Initialize queue $Q \leftarrow \{r\}$
$TreeLevels \leftarrow [\,]$
while Q not empty **do**
  $\ell \leftarrow$ length(Q)
  $C \leftarrow [\,]$
  for $i = 1$ to ℓ **do**
    $u \leftarrow$ POP_FRONT(Q)
    append u.index to C
    if u.left $\neq$ NULL **then**
      push u.left to back of Q
    end if
    if u.right $\neq$ NULL **then**
      push u.right to back of Q
    end if
  end for
  append C to $TreeLevels$
end while
return $TreeLevels$

Algorithm 5 Temporal-scale Autoregressive

$T \leftarrow \frac{H}{r}$
$L = \lceil \log_2(T+1) \rceil$ ▷ *Depth of a midpoint binary tree.*

$r \leftarrow$ BUILD_TREE$(1, T)$ ▷ *Build a balanced midpoint binary tree.*
$TreeLevels \leftarrow$ BFS_LEVELS(r) ▷ *Visited nodes of breadth-first search on a tree at each level.*
for $l = 1, \ldots, L$ **do**
  $TreeLevels[l] \leftarrow TreeLevels[l]$ with T removed ▷ $TreeLevels[l]$ is a list or an empty list.
end for

$K \leftarrow V \leftarrow v_t$
$v_{t+T} \leftarrow \texttt{Attn}(Q_{t+T}, K, V)$ ▷ *Jump to last time step.*
$K \leftarrow V \leftarrow [V : V_{t+T}]$

$I_l \leftarrow \emptyset$
for $l = 1, \ldots, L$ **do**
  for each index x in $TreeLevels[l]$ **do**
    $I_l = I_l \cup \{x\}$
  end for
  $I_l = I_l \cup \{T\}$ ▷ *Temporal indices generated by accumulated breadth-first search.*

  $\{v_t\}_{t \in I_l} \leftarrow \texttt{Attn}(\{Q_t\}_{t \in I_l}, K, V)$
  $K \leftarrow V \leftarrow [V : \{v_t\}_{t \in I_l}]$
end for