FlowHOI: Flow-based Semantics-Grounded Generation of Hand-Object Interactions for Dexterous Robot Manipulation

Huajian Zeng[1], Lingyun Chen[2], Jiaqi Yang[1], Yuantai Zhang[1], Fan Shi[3], Peidong Liu[4], Xingxing Zuo[*,1]

huajian-zeng.github.io/projects/flowhoi

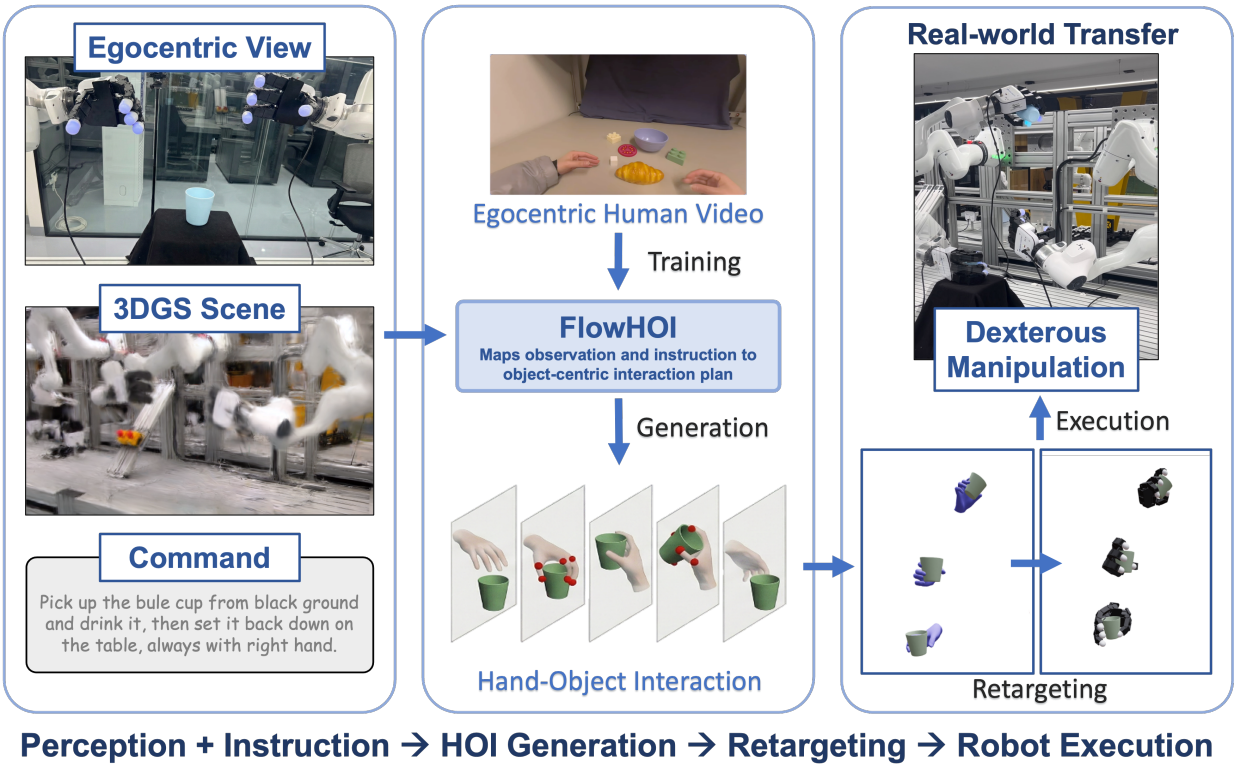

Fig. 1: We generate hand-object interaction (HOI) motions conditioned on an egocentric observation, a text command, and 3D scene context. We first learn a grasping prior from HOI data extracted from large-scale egocentric videos, then generate semantically grounded manipulation motions that respect the language instruction, the surrounding 3D scene, and geometric constraints. The generated motions can be retargeted to robot hands for real-world execution.

***Abstract*—Recent vision-language-action (VLA) models can generate plausible end-effector motions, yet they often fail in long-horizon, contact-rich tasks because the underlying hand-object interaction (HOI) structure is not explicitly represented. An embodiment-agnostic interaction representation that captures this structure would make manipulation behaviors easier to validate and transfer across robots. We propose FlowHOI, a two-stage flow-matching framework that generates semantically grounded, temporally coherent HOI sequences, comprising hand poses, object poses, and hand-object contact states, conditioned on an egocentric observation, a language instruction, and a 3D Gaussian splatting (3DGS) scene reconstruction. We decouple geometry-centric grasping from semantics-centric manipulation, conditioning the latter on compact 3D scene tokens and adding a motion-text alignment loss to ground the generated interactions in both the scene layout and the language instruction. To address the scarcity of high-fidelity HOI supervision, we introduce a pipeline that recovers aligned hand-object trajectories from large-scale egocentric videos, yielding an HOI prior for robust generation. Across the GRAB and HOT3D benchmarks, FlowHOI achieves the highest action recognition accuracy and a 1.7× higher physics-simulation success rate on GRAB than the strongest diffusion-based baseline, while delivering a 40×**

[1]Huajian Zeng, Jiaqi Yang, Yuantai Zhang, and Xingxing Zuo are with Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. {huajian.zeng, jiaqi.yang, yuantai.zhang, xingxing.zuo}@mbzuai.ac.ae

[2]Lingyun Chen is with the Technical University of Munich (TUM), Munich, Germany. lingyun.chen@tum.de

[3]Fan Shi is with the National University of Singapore (NUS), Singapore. fan.shi@nus.edu.sg

[4]Peidong Liu is with Westlake University, Hangzhou, China. liupeidong@westlake.edu.cn

*Corresponding author.

inference speedup. We further demonstrate real-robot execution across four contact-rich, dexterous manipulation tasks, achieving significantly higher success rates than prior methods due to improved physical plausibility and semantic alignment.

I. Introduction

Robotic manipulation in household environments is fundamentally interaction-centric [1, 2, 3]: tasks such as opening a lid or pouring from a container are characterized not by the end-effector trajectory alone, but by where contact occurs given the scene geometry, how stable contact is preserved, and how the object's state evolves consistently with the language instruction and scene affordances.

Recent vision-language-action (VLA) models [4, 5, 6], fine-tuned from large VLMs [7, 8, 9, 10], generate plausible end-effector trajectories but struggle with contact-rich, long-horizon dexterous tasks [11]. This motivates a semantically grounded, embodiment-agnostic intermediate representation [12, 13] that explicitly encodes how contact induces language-specified state changes [14], decoupling manipulation from robot-specific control to facilitate transfer across embodiments.

In this work, we propose FlowHOI, a two-stage flow-matching framework that, given an initial egocentric observation, a language instruction, and a 3D Gaussian splatting (3DGS) scene reconstruction, generates semantically grounded, temporally coherent hand-object interaction (HOI) sequences of hand poses, object poses, and contact states, anchored in the scene and aligned with the instruction. The resulting representation is a natural interface for geometric and physical constraints and can be retargeted to downstream robotic dexterous manipulation [15, 16].

Building such a model raises three central challenges. **(i) Geometric consistency & semantic grounding.** Long-horizon interactions must satisfy both the 3D scene geometry and the instruction's semantics. Naively entangling these objectives causes contact drift or semantically inconsistent motions. Inspired by how humans grasp before manipulating [17], we decompose generation into a geometry-centric *Grasping* stage with a pretrained prior and a semantics-centric *Manipulation* stage conditioned on compact 3D scene tokens with a motion-text alignment loss. **(ii) Inference efficiency.** Diffusion-based HOI generators [18, 19, 20, 21] need tens to hundreds of denoising steps (3–7 s per sequence), prohibitive for real-time planning. We adopt conditional flow matching, cutting inference to 0.16 s (up to 40× speedup) while staying temporally coherent and physically plausible. **(iii) Data scarcity.** High-fidelity HOI supervision is scarce because interactions are high-dimensional, occluded, and tightly coupled with contact [22, 23]. We introduce a reconstruction pipeline that recovers aligned hand-object trajectories from large-scale egocentric videos [24], enabling robust priors that generalize across objects and tasks.

To our knowledge, FlowHOI is the *first* method to formulate HOI generation as a unified, conditional flow-matching process in a semantics-grounded way. Our contributions are a two-stage framework with *stage-specific conditioning* that decouples geometry-centric grasping from semantics-centric manipulation for up to 40× faster generation, semantic grounding via a motion-text alignment loss and a hybrid geometric-semantic 3D scene representation, and a reconstruction pipeline that extracts large-scale HOI data from egocentric videos for a robust prior. On GRAB and HOT3D, FlowHOI attains the highest action-recognition accuracy and a 1.7× higher physics-simulation success rate over the strongest baseline, with validated real-world retargeting.

II. Related Work

Hand-Object Interaction Generation. HOI synthesis under realistic contact is supported by mocap- and vision-based datasets [25, 26, 27, 28, 29]. Early kinematic methods use contact-aware priors, optimization, or latent-variable models in canonical object spaces [30, 31, 32, 33, 14, 34, 35, 36], while physics-based RL [37, 38] ensures feasibility but scales poorly. Recent diffusion methods [39, 40, 41, 42] improve temporal coherence, yet DiffH2O [18] still shows physical artifacts and slow inference and LatentHOI [19] is limited in length and data scale. We instead use flow matching with explicit grasping and manipulation phases for faster, more plausible, semantically aligned generation.

Robot Learning and Motion Priors from Videos. Learning from human videos [43, 44, 45] is scalable but faces an embodiment gap only partly bridged by egocentric datasets [29, 46, 47], while VLA models [4, 5, 6], language-conditioned policies [48, 49], and video generators [50, 51, 52, 53] lack explicit interaction structure. Diffusion-based human-motion and HOI methods [54, 55, 56, 57, 58, 59, 60, 61, 62, 63] target full-body motion on small mocap sets [64, 65, 66, 67], and WHOLE [68] lifts egocentric HOI into a world-grounded representation but yields per-clip reconstructions, not a prior. We instead learn a robot-agnostic, fine-grained HOI prior from egocentric videos [24] retargetable across embodiments.

III. Methodology

As illustrated in Fig. 2, FlowHOI generates HOI motion from a single-frame egocentric observation, a text command, and 3D scene context in two stages: a *Grasping* stage that produces approach-and-grasp motion from a pretrained prior fine-tuned on reconstructed egocentric HOI data, and a *Manipulation* stage that produces the subsequent interaction conditioned on 3D scene context and language. Both stages are conditional flow-matching (CFM) [69] generators with an x-prediction parameterization [70] for temporal stability. We summarize the CFM formulation in Supp. Sec. A.

A. Problem Formulation

From the first observation I we extract an initial state $\mathbf{x}_{\text{init}}$, object geometry $\mathcal{M}$, instruction T, and a 3DGS scene $\mathcal{G}$, and generate an N-frame HOI trajectory $\mathbf{x} = \{(\mathbf{H}_t^l, \mathbf{H}_t^r, \mathbf{O}_t)\}_{t=0}^{N-1} \in \mathbb{R}^{N\times D}$ stacking per-frame left/right hand and object states. We learn a conditional model $p_\theta(\mathbf{x} \mid \mathbf{c})$ with $\mathbf{c}$ encoding $\mathbf{x}_{\text{init}}, \mathcal{M}, T, \mathcal{G}$.

Data representation. Following [18], each frame couples the object pose ($\mathbb{R}^9$, 6D rotation [71]) with bimanual MANO [72]

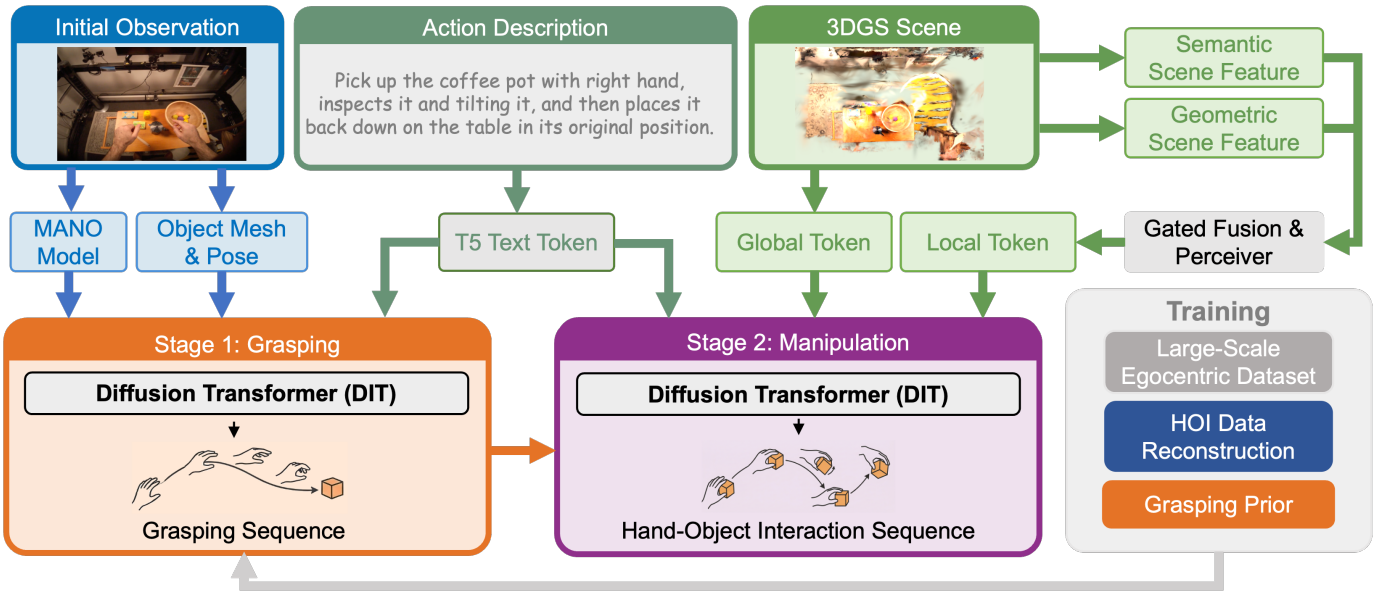

Fig. 2: **Overview of FlowHOI.** A two-stage pipeline conditioned on an egocentric observation, text command, and 3D scene: (1) approach-and-grasp from a prior fine-tuned on reconstructed egocentric HOI data, and (2) manipulation conditioned on scene and language.

hand states ($\mathbb{R}^{54}$: root, 6D orientation, 24D PCA pose, and per-joint object signed distances for contact-aware generation), anchored at the transition-frame object position to reduce inter-scene variance (full notation in Supp. Sec. A).

B. Hand-Object Data Reconstruction from Egocentric Videos

HOI datasets [26, 73] are far smaller than full-body MoCap [67], so we build a three-step pipeline that turns raw egocentric EgoDex [24] videos (no object annotations) into high-fidelity HOI data for pretraining the grasping prior: grasp-to-manipulation transition detection, object reconstruction via SAM3 [74]/DepthAnything3 [75]/SAM3D [76], and contact-based MANO-object alignment under fingerpad-contact and non-penetration constraints (full pipeline and quality evaluation in Supp. Sec. D).

C. Two-Stage Hand-Object Interaction Generation

Motivated by the human grasp-then-manipulate pattern [17], we split HOI into a *Grasping* and a *Manipulation* phase. Unlike DiffH2O [18], which also separates the phases, our key design is *stage-specific conditioning*: grasping uses only object geometry, language, and the initial state, while manipulation additionally consumes 3DGS scene tokens with a motion-text alignment objective, so each stage focuses on its most relevant signals rather than entangling all sources in one denoiser.

Initial state estimation. We estimate the initial hand pose [77] and object mesh/6D pose [76] into a world-frame state $\mathbf{x}_{\text{init}}$, encode geometry $\mathcal{M}$ with a Basis Point Set [78], and encode the instruction T with a frozen T5-Large [79].

Hybrid 3D scene tokens. Manipulation conditions on hybrid tokens built from a single 3DGS reconstruction [80, 81] (moving hands and objects masked out). A geometric stream (Concerto [82]) and a language-aligned semantic stream (SceneSplat [83]) over sampled Gaussian centroids are fused by a learnable channel-wise gate $\mathbf{F} = \tilde{\mathbf{E}} + \sigma(\boldsymbol{\alpha}) \odot \tilde{\mathbf{U}}$ [84] and compressed by a Perceiver bottleneck [85] into compact *local* tokens $\mathbf{S}_{\text{local}}$ tied to specific interaction regions, complemented by a *global* occupancy token $\mathbf{S}_{\text{global}}$ from a 3D ViT [86] that discourages catastrophic collisions (full architecture in Supp. Sec. H).

Grasping stage. Since the object is static, grasping generates only the hand state $\mathbf{x}^g = (\mathbf{H}^l, \mathbf{H}^r)$, conditioned on $\mathbf{c}^g = \{\text{BPS}(\mathcal{M}), \text{T5}(T_g), \mathbf{x}_{\text{init}}\}$, where T_g is a grasp-focused sub-instruction from T via an MLLM [87] (isolating grasping from manipulation semantics) and $\mathbf{x}_{\text{init}}$ anchors generation. We train f_θ^g with x-prediction [70] along the CFM optimal-transport path using the flow loss $\mathcal{L}_{\text{flow}}^g = \mathbb{E}\|f_\theta^g(\mathbf{x}_\tau^g, \tau, \mathbf{c}^g) - \mathbf{x}_1^g\|_2^2$ and a TMR-inspired symmetric InfoNCE alignment loss $\mathcal{L}_{\text{align}} = \frac{1}{2}(\mathcal{L}_{\text{t2m}} + \mathcal{L}_{\text{m2t}})$ [88, 89] that grounds motion in language, giving $\mathcal{L}_{\text{grasp}} = \mathcal{L}_{\text{flow}}^g + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$ ($\lambda_{\text{align}} = 0.1$).

Manipulation stage. Manipulation generates the full sequence $\mathbf{x}^m \in \mathbb{R}^{N \times D}$ conditioned on $\mathbf{c}^m = \{\text{BPS}(\mathcal{M}), \text{T5}(T), \mathbf{S}_{\text{local}}, \mathbf{S}_{\text{global}}, \mathbf{x}_{t_g}^g\}$, treating the previously generated grasping trajectory as a fixed prefix. At each Euler step we softly inpaint this prefix into the noisy trajectory and clamp the transition state, so unlike DiffH2O's diffusion-step inpainting [18] our continuous-time formulation enforces transition consistency at the ODE level (details in Supp. Sec. J). Training uses a masked x_1-prediction loss on the manipulation portion plus the alignment loss, $\mathcal{L}_{\text{manip}} = \mathcal{L}_{\text{flow}}^m + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$.

IV. Experiments

A. Setup

We evaluate on GRAB [26] (high-fidelity mocap) and HOT3D [29] (real-world egocentric video for 3DGS scenes and unseen objects), pretraining the grasping prior on curated EgoDex [24] (Sec. III-B). We retrain DiffH2O [18] and LatentHOI [19] under identical splits with CLIP [90]

Dataset	Method	Physical Interaction Quality				Motion Quality			Realizable Physical Feasibility			
		IV [cm^3] (↓)	ID [cm] (↓)	CR [%] (↑)	IVU (↓)	AR (↑)	SD [m] (↑)	OD [m] (↑)	Phy [%] (↑)	SR [%] (↑)	HT [s] (↑)	Time [s] (↓)
GRAB	Real Mocap	6.49	0.43	6.62	0.09	1.00	-	0.18	95.56	77.06	2.40	-
	DiffH2O	13.11	1.32	8.14	0.16	0.87	0.12	0.13	89.44	33.03	0.35	6.34
	LatentHOI	13.76	**0.99**	**9.18**	0.14	0.78	0.11	0.14	**96.40**	28.44	0.29	3.57
	Ours	**10.93**	1.28	6.85	**0.13**	**0.95**	**0.13**	**0.16**	90.90	**55.96**	**1.50**	**0.16**
HOT3D	Real Mocap	2.88	0.63	1.69	0.15	1.00	-	0.23	16.80	6.67	0.08	-
	DiffH2O	3.25	0.66	2.07	**0.18**	0.71	**0.15**	**0.23**	16.47	4.00	0.03	6.54
	LatentHOI	**1.54**	**0.44**	1.59	0.21	0.65	0.12	0.18	14.36	0.00	0.00	3.21
	Ours	3.36	0.72	**2.14**	0.19	**0.78**	0.10	0.20	**17.15**	**5.33**	**0.03**	**0.16**

TABLE I: Quantitative evaluation on GRAB and HOT3D. Best generation method in **bold**, Real Mocap is the upper bound. IV/ID bolded only with non-trivial CR (cf. LatentHOI's 0% physics SR, Supp. Sec. E).

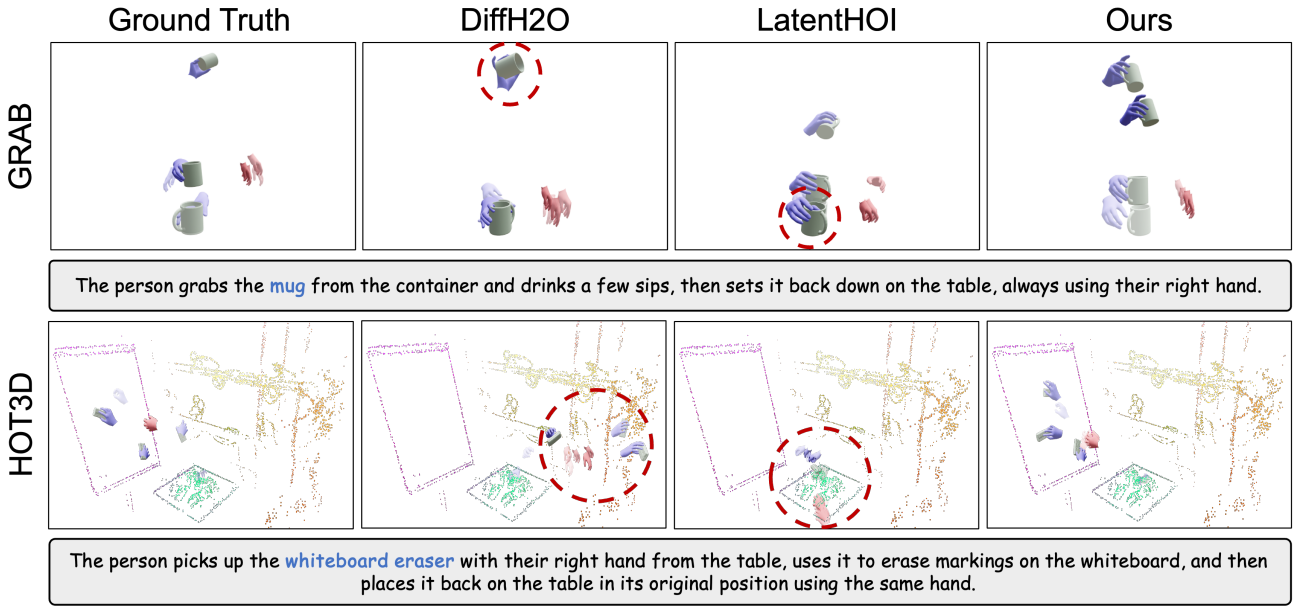

Fig. 3: Qualitative comparison vs. DiffH2O [18] and LatentHOI [19] (top: GRAB, bottom: HOT3D with 3D scene). Ours yields more natural grasps and physically plausible manipulations that better follow the instruction. Best viewed in the supplementary video.

replaced by T5 [79]. Metrics (definitions in Supp. Sec. I) cover physical interaction quality (interpenetration volume IV, depth ID, contact ratio CR, IV-per-contact IVU), motion quality (action-recognition accuracy AR, diversity SD/OD), and realizable feasibility: the heuristic plausibility Phy [19] and a physics-grounded success rate (SR), holding time (HT), and generation time in Isaac Gym [91, 92, 93]. HOT3D is noisier (Real Mocap only 6.67% SR), so we weight CR and AR there (Supp. Sec. E), and treat SR as the primary indicator since low IV/IVU is trivial without contact.

B. Results

Generation (Table I, Fig. 3). FlowHOI attains the lowest IV and IVU and the highest AR on GRAB, and the highest CR on HOT3D, whereas LatentHOI's lower IV/ID comes with much weaker contact and a 0% physics SR. It reaches the best GRAB SR (55.96% vs. 28.44%/33.03%) and a 40× planning speedup (0.16 s vs. 6.34/3.57 s), with diversity comparable to baselines and qualitatively more realistic grasps and smoother manipulations. **Real world.** On a dual Franka+Allegro platform [94, 95], FlowHOI acts as a *planner*, producing one open-loop trajectory per task (drinking, pouring, tilting, squeezing) retargeted via pyroki [91] and DexTrack [93] (Fig. 1). It reaches 57.5% average SR, 5.75× DiffH2O's 10% (Supp. Table V, Sec. G), with failures mainly from embodiment mismatch and open-loop slippage. **Ablations (Supp. Sec. B).** Pretraining the grasping prior cuts grasp error from 0.10 to 0.06 m (40%), T5 and the motion-text alignment loss raise AR, and our fused geometric-semantic scene tokens give the best AR and final displacement error.

V. Conclusion

We presented FlowHOI, a two-stage flow-matching framework that decouples geometry-centric grasping from semantics-centric manipulation via 3D scene tokens and a motion-text alignment loss. It attains the highest action-recognition accuracy, a 1.7× higher physics SR, and a 40× speedup over DiffH2O [18], with validated real-robot retargeting. **Limitations and future work.** FlowHOI relies on single-observation initial-state estimation and degrades under heavy occlusion or unreliable reconstruction. Its trajectories are kinematic and run open-loop, so they cannot react to online perturbations or post-grasp contact slippage. Promising directions include closed-loop re-planning, mobile manipulation that jointly models base motion and reachability, and learning interaction priors from exocentric videos.

REFERENCES

[1] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Houghton Mifflin, 1979.

[2] Matthew T Mason. *Mechanics of robotic manipulation*. MIT press, 2001.

[3] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065)*, volume 1, pages 348–353. IEEE, 2000.

[4] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

[6] Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. *arXiv preprint arXiv:2506.19850*, 2025.

[7] 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.

[8] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[9] 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.

[10] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025.

[11] Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Zhang Chen, Tianrui Guan, Fanlian Zeng, Ka Num Lui, Yuyao Ye, Yitao Liang, et al. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. *arXiv preprint arXiv:2502.20900*, 2025.

[12] Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17359–17369, 2025.

[13] Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, et al. A0: An affordance-aware hierarchical model for general robotic manipulation. *arXiv preprint arXiv:2504.12636*, 2025.

[14] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021.

[15] Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. Spot: Se (3) pose trajectory diffusion for object-centric manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4853–4860. IEEE, 2025.

[16] Tyler Ga Wei Lum, Olivia Y Lee, C Karen Liu, and Jeannette Bohg. Crossing the human-robot embodiment gap with sim-to-real rl using one human demonstration. *arXiv preprint arXiv:2504.12609*, 2025.

[17] Roland S Johansson and J Randall Flanagan. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience*, 10 (5):345–359, 2009.

[18] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[19] Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17416–17425, 2025.

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

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[22] Hongming Fu, Wenjia Wang, Xiaozhen Qiao, Shuo Yang, Zheng Liu, and Bo Zhao. Egograsp: World-space hand-object interaction estimation from egocentric videos. *arXiv preprint arXiv:2601.01050*, 2026.

[23] Yue Xu, Yong-Lu Li, Zhemin Huang, Michael Xu Liu,

Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. Egopca: A new framework for egocentric hand-object interaction understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5273–5284, 2023.

[24] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[25] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021.

[26] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

[27] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.

[28] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023.

[29] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7061–7071, 2025.

[30] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023.

[31] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021.

[32] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3035–3044, 2024.

[33] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020.

[34] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13263–13273, 2022.

[35] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021.

[36] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2023.

[37] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 464–473. IEEE, 2024.

[38] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

[39] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, volume 42, pages 1–12. Wiley Online Library, 2023.

[40] Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. *arXiv preprint arXiv:2402.14810*, 2024.

[41] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024.

[42] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Tridi: Trilateral diffusion of 3d humans, objects, and interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5523–5535, 2025.

[43] Ajay Mandlekar, Jonathan Booher, Max Spero, Albert Tung, Anchit Gupta, Yuke Zhu, Animesh Garg, Silvio Savarese, and Li Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1048–1055. IEEE, 2019.

[44] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karam-

cheti, 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.

[45] 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.

[46] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 4125–4141, 2020.

[47] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913, 2025.

[48] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[49] Ce Hao, Kelvin Lin, Siyuan Luo, and Harold Soh. Language-guided manipulation with diffusion policies and constrained inpainting. *arXiv preprint arXiv:2406.09767*, 2024.

[50] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

[51] Boyang Wang, Nikhil Sridhar, Chao Feng, Mark Van der Merwe, Adam Fishman, Nima Fazeli, and Jeong Joon Park. This&that: Language-gesture controlled video generation for robot planning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12842–12849. IEEE, 2025.

[52] Jiageng Mao, Sicheng He, Hao-Ning Wu, Yang You, Shuyang Sun, Zhicheng Wang, Yanan Bao, Huizhong Chen, Leonidas Guibas, Vitor Guizilini, et al. Robot learning from a physical world model. *arXiv preprint arXiv:2511.07416*, 2025.

[53] Raktim Gautam Goswami, Amir Bar, David Fan, Tsung-Yen Yang, Gaoyue Zhou, Prashanth Krishnamurthy, Michael Rabbat, Farshad Khorrami, and Yann LeCun. World models can leverage human videos for dexterous manipulation. *arXiv preprint arXiv:2512.13644*, 2025.

[54] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[55] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023.

[56] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18000–18010, 2023.

[57] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022.

[58] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022.

[59] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.

[60] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023.

[61] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024.

[62] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2878–2888, 2025.

[63] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

[64] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

[65] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking

human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.

[66] Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. *arXiv preprint arXiv:2504.10414*, 2025.

[67] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[68] Yufei Ye, Jiaman Li, Ryan Rong, and C Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *arXiv preprint arXiv:2602.22209*, 2026.

[69] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[70] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[71] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

[72] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[73] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021.

[74] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[75] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[76] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

[77] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025.

[78] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019.

[79] 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.

[80] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. In *European Conference on Computer Vision*, pages 382–400. Springer, 2024.

[81] Zhaoyang Lv, Maurizio Monge, Ka Chen, Yufeng Zhu, Michael Goesele, Jakob Engel, Zhao Dong, and Richard Newcombe. Photoreal scene reconstruction from an egocentric device. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[82] Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. *arXiv preprint arXiv:2510.23607*, 2025.

[83] Yue Li, Qi Ma, Runyi Yang, Huapeng Li, Mengjiao Ma, Bin Ren, Nikola Popovic, Nicu Sebe, Ender Konukoglu, Theo Gevers, et al. Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. *arXiv preprint arXiv:2503.18052*, 2025.

[84] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[85] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[86] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024.

[87] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[88] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages

9488–9497, 2023.

[89] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[90] 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.

[91] Chung Min Kim, Brent Yi, Hongsuk Choi, Yi Ma, Ken Goldberg, and Angjoo Kanazawa. Pyroki: A modular toolkit for robot kinematic optimization. *arXiv preprint arXiv:2505.03728*, 2025.

[92] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[93] Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, and Li Yi. Dextrack: Towards generalizable neural tracking control for dexterous manipulation from human references. *arXiv preprint arXiv:2502.09614*, 2025.

[94] Sami Haddadin. The franka emika robot: A standard platform in robotics research. *IEEE Robotics & Automation Magazine*, 2024.

[95] Allegro hand v5. https://www.allegrohand.com/, accessed 2026. Wonik Robotics.

[96] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[97] 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.

[98] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315, 1997.

[99] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems*, 34:15816–15829, 2021.

[100] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[101] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[102] Moo K Chung. Gaussian kernel smoothing. *arXiv preprint arXiv:2007.09539*, 2020.

[103] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[104] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[105] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[106] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multimodal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

[107] SkylandX. MetaCam: 3D scanner for spatial intelligence. https://skylandx.com/, accessed 2026.

[108] Xiaolei Lang, Laijian Li, Chenming Wu, Chen Zhao, Lina Liu, Yong Liu, Jiajun Lv, and Xingxing Zuo. Gaussian-lic: Real-time photo-realistic slam with gaussian splatting and lidar-inertial-camera fusion. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8500–8507. IEEE, 2025.

[109] Tamar Flash and Neville Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience*, 5(7):1688–1703, 1985.

FlowHOI: Flow-based Semantics-Grounded Generation of Hand-Object Interactions for Dexterous Robot Manipulation

Supplementary Material

This supplementary document provides additional details and results that complement the main paper. We begin with implementation details of the model architecture and training configurations in Sec. A, followed by an ablation study on the flow matching prediction target in Sec. C. We then describe the hand-object data reconstruction pipeline and dataset preprocessing in Sec. D and Sec. E, respectively. Next, we elaborate on the physics simulation evaluation protocol in Sec. F and provide details on the real-world experiment setup in Sec. G. We then analyze the scene representation in Sec. H. Full evaluation metric definitions are provided in Sec. I. We additionally provide a stage-transition soft-inpainting ablation in Sec. J, additional qualitative results in Sec. K, and failure cases in Sec. L. We also provide a supplementary video with animated results.

A. Implementation details

This section provides additional implementation details beyond Sec. IV-A of the main paper.

A. Flow matching formulation

We collect here the flow matching formulation used by FlowHOI (Sec. III of the main paper). Conditional flow matching (CFM) [69] transports a Gaussian prior $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ over $\mathbf{x} \in \mathbb{R}^d$ to the data distribution $q(\mathbf{x})$ along a continuous probability path. Since matching the marginal vector field of the intermediate distribution p_τ is intractable, CFM constructs a tractable conditional path per sample $\mathbf{x}_1 \sim q$. With the optimal-transport linear interpolation,

$$\mathbf{x}_\tau = \big(1-(1-\sigma_{\min})\tau\big)\mathbf{x}_0 + \tau\,\mathbf{x}_1, \quad \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \tau \sim \mathcal{U}[0,1], \tag{S1}$$

the conditional vector field admits a closed form:

$$\mathbf{u}_\tau(\mathbf{x}_\tau \mid \mathbf{x}_1) = \frac{\mathbf{x}_1 - (1-\sigma_{\min})\mathbf{x}_\tau}{1-(1-\sigma_{\min})\tau}. \tag{S2}$$

A network $\mathbf{v}_\theta(\mathbf{x}, \tau, \mathbf{c})$ conditioned on side information $\mathbf{c}$ is regressed to this target with a mean-squared loss. In FlowHOI we adopt the equivalent x-prediction parameterization that directly regresses the clean target $\mathbf{x}_1$ for improved temporal stability [70] (analysis in Sec. C).

Neural ODE formulation. The continuous-time transport from p_0 to p_1 underlying conditional flow matching is defined by the neural ODE [96]:

$$\frac{d}{d\tau}\,\boldsymbol{\phi}_\tau(\mathbf{x}) = \mathbf{v}(\boldsymbol{\phi}_\tau(\mathbf{x}), \tau)\,, \qquad \boldsymbol{\phi}_0(\mathbf{x}) = \mathbf{x}, \tag{S3}$$

where $\boldsymbol{\phi}_\tau$ is the flow map. The optimal-transport path of Eq. S1 is one specific instantiation of this ODE.

Role of $\sigma_{\min}$. The constant $\sigma_{\min} > 0$ in Eq. S1 controls the residual stochasticity at the end of the flow: at $\tau = 1$ the conditional distribution becomes $p_1(\mathbf{x} \mid \mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1, \sigma_{\min}^2\mathbf{I})$ rather than a Dirac delta, preventing the path from degenerating and keeping the closed-form vector field of Eq. S2 numerically stable as $\tau \to 1$. We set $\sigma_{\min} = 10^{-5}$, small enough that its effect on sample quality is negligible.

Euler discretization at inference. At inference, sampling starts from $\mathbf{x}(0) = \mathbf{x}_0 \sim p_0$ and integrates Eq. S3 to $\tau = 1$ with a numerical ODE solver. We use Euler discretization with $K = 50$ steps:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\tau\ \mathbf{v}_\theta(\mathbf{x}_k, \tau_k, \mathbf{c}), \qquad \Delta\tau = \tfrac{1}{K}, \quad \tau_k = \tfrac{k}{K}. \tag{S4}$$

B. Architecture and training

Both models use a DiT [97] backbone with pre-norm. The manipulation model uses d=512, d_{ff}=1024, while the grasping model uses d=256, d_{ff}=512. Both use L=8 layers and dropout 0.1. Text is encoded by a frozen T5-Large [79], projected to d and injected via cross-attention. Object geometry uses BPS [78] encoding, projected to 256-dim and concatenated per-frame.

The scene-tokenization pipeline below is consumed only by the manipulation stage; the grasping stage is geometry-centric (cf. Sec. III-C) and does not include scene tokens in its conditioning. The scene is represented by up to 25,000 points sampled via Farthest Point Sampling (FPS) [98]. Geometric features from Concerto [82] d_e=1536 and semantic features from SceneSplat [83] d_u=768 are fused via bidirectional cross-attention with d_h=512. 3D coordinates use Fourier positional encoding [99], projected to 64-dim. Fused features are compressed by a Perceiver [85] bottleneck into K=256 scene tokens of dimension d. For global scene encoding, we voxelize the scene at 48×48×24 resolution and process it with a ViT, injected via AdaLN [97].

We use AdamW [100] with lr 10^{-4}, weight decay 0.01, and linear annealing. The grasping model trains for 500K steps, and the manipulation model for 200K steps. We adopt logit-normal time sampling and classifier-free guidance with drop probability 0.1 during training and scale 2.5 at inference with 50-step Euler integration.

Each frame is a 117-dim vector encoding bimanual wrist poses, hand articulations in 24D PCA, per-joint signed distances to the object, and object pose. The full sequence length is N=200 frames with N_g=50 for grasping and 150 for manipulation.

Contrastive Alignment Loss. We employ a TMR-inspired [88] symmetric InfoNCE loss to align motion and text embeddings in a shared 512-dim latent space. The motion encoder is a 4-layer Transformer encoder with a learnable `[CLS]` token. The text encoder applies mean pooling over frozen T5-Large [79] embeddings and projects them to the same 512-dim space via a two-layer MLP. The

TABLE II: Ablation studies. (a) Scene encoder choices; FDE: final displacement error of object pose. (b) Effect of pretraining with large-scale egocentric HOI data on GRAB; GE: grasp error at the end of the grasping stage. (c) Text encoder choices.

(a) Scene encoder.

Method	AR (↑)	FDE (↓)
w/o scene	0.73	0.33
PointNet++ [101]	0.59	0.44
Concerto [82]	0.75	0.26
SceneSplat [83]	0.73	0.30
Ours	**0.78**	**0.24**

(b) Pretraining grasping prior on GRAB.

Model	IV $[cm^3]$ (↓)	ID $[cm]$ (↓)	CR [%] (↑)	IVU (↓)	GE $[m]$ (↓)
w/o prior	**10.40**	1.31	6.78	0.13	0.10
w/ prior	10.93	**1.28**	**6.85**	0.13	**0.06**

(c) Text encoder.

Method	AR (↑)
CLIP [90]	0.89
CLIP + Align. Loss	0.88
T5 [79]	0.92
T5 + Align. Loss	**0.94**

temperature τ is initialized to 0.7, learnable, and clamped to $[0.01, 100]$. The alignment loss is weighted by λ_{align}=0.1 and directly added to the flow matching loss. These alignment components are only used during training and discarded at inference.

Baseline reproduction. For DiffH2O [18] and LatentHOI [19] we use the official codebases and follow each paper's published training schedule (optimizer, learning rate, batch size, total iterations) on GRAB and HOT3D, without re-tuning. The only deviation is the CLIP [90] → T5-Large [79] text encoder swap applied uniformly for fairness; we re-initialize the linear projection that maps text features into each model's hidden dimension. All other hyperparameters (*e.g.*, diffusion steps and CFG scale for DiffH2O, latent dimensionality and decoder depth for LatentHOI) follow the original papers.

B. Ablation Tables

We provide the full ablation tables discussed in Sec. IV of the main paper: the effect of pretraining the grasping prior, and the text/scene conditioning study.

C. Flow matching prediction target analysis

Different from standard flow matching that predicts the velocity field $\mathbf{v}$, we propose to directly predict the clean data $\mathbf{x}_1$ as the model output. Recent studies [70] have conducted extensive analyses on the choice of prediction targets and consistently advocate for directly predicting the clean data (x-prediction), rather than noise (ϵ-prediction) or flow velocity (v-prediction). These works are motivated by the *manifold assumption*, which posits that natural data, such as images, concentrate on a low-dimensional manifold embedded in a high-dimensional ambient space, while quantities such as noise ϵ or flow velocity $v = x - \epsilon$ are inherently off-manifold and distributed across the full ambient space.

In our problem, although the interaction sequence is represented in a high-dimensional space, physically and semantically plausible hand-object motions are highly constrained (e.g., kinematics, joint limits, contact consistency, and temporal coherence), suggesting a similarly thin *interaction manifold*. Under velocity-based training, small high-frequency errors in $\mathbf{v}_\theta$ are directly injected into the ODE integration and can accumulate across steps, manifesting as temporal jitter [70]. In contrast, x-prediction repeatedly estimates an on-manifold state and then derives the vector field from it, which empirically improves temporal stability. To validate this hypothesis, we conduct an ablation study comparing the two prediction targets (x-prediction and v-prediction) on the

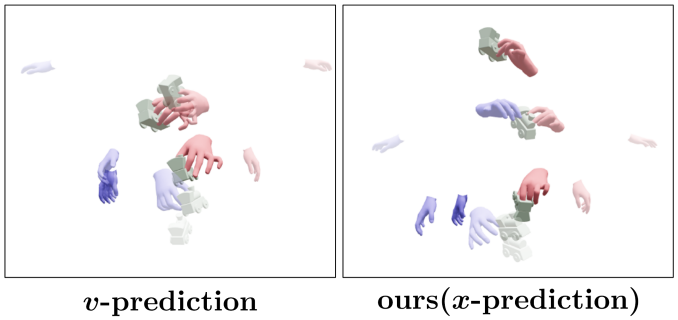

Fig. 4: Qualitative comparison of prediction targets. Our x-prediction produces temporally smooth and stable hand trajectories, while v-prediction exhibits noticeable spatial jitter and inconsistent grasping poses across frames.

GRAB [26] dataset. We evaluate temporal smoothness using two complementary metrics (see Sec. I for formal definitions): (1) *jerk*, the third derivative of object position/rotation with respect to time, and (2) *hand acceleration*, which measures the second-order temporal stability of the generated hand motion, decomposed into global wrist acceleration (Acc_g) and local finger acceleration (Acc_l) for both positional and rotational components. As shown in Tab. III, x-prediction achieves lower angular jerk and consistently lower local hand acceleration (both positional and rotational), indicating smoother finger articulation. The global wrist metrics are comparable between the two targets, suggesting that the primary benefit of x-prediction lies in stabilizing fine-grained hand dynamics rather than coarse wrist trajectories. Fig. 4 further visualizes the generated trajectories, where v-prediction exhibits noticeable high-frequency jitter whereas x-prediction produces temporally smooth results.

D. Hand-object data reconstruction details

The goal of this pipeline is to reconstruct high-fidelity hand-object interaction data from large-scale egocentric videos [24], which provides large-scale training data for pretraining the grasping model and improves generalization across objects and tasks. This section provides the full algorithmic details, loss formulations, and hyperparameters of the reconstruction pipeline summarized in Sec. III-B of the main paper. An overview of the pipeline is illustrated in Fig. 5.

A. Step 1: transition frame detection

To separate the grasp phase from the manipulation phase, we detect a transition frame using wrist motion cues. We first smooth the wrist trajectory with a Gaussian kernel [102]

Prediction Target	Jerk (Pos.) $[m/s^3]$ (↓)	Jerk (Ang.) $[rad/s^3]$ (↓)	Acc_g (Pos.) $[m/s^2]$ (↓)	Acc_l (Pos.) $[m/s^2]$ (↓)	Acc_g (Rot.) $[rad/s^2]$ (↓)	Acc_l (Rot.) $[rad/s^2]$ (↓)
v-prediction	4.9721	137.9271	0.2523	0.1232	0.9567	3.2414
x-prediction (Ours)	**0.0438**	**0.4071**	**0.0204**	**0.0055**	**0.0835**	**0.0339**

TABLE III: Ablation study on flow matching prediction targets on GRAB [26]. x-prediction yields lower angular jerk and hand acceleration while maintaining competitive interaction quality. Acc_g/Acc_l: global/local acceleration (see Sec. I).

(σ=2.0) and compute wrist speed as the displacement between consecutive frames. Candidate transitions are identified at time steps where the speed reaches a local minimum within a 1-second window and the orientation changes by more than 30° in the window.

B. Step 2: 3D object reconstruction

We use an LLM (GPT-4o [87]) to extract the label of the object of interest from the given text prompt. The object is stable during the grasping phase, allowing us to reconstruct its 3D geometry from the video start frame to the transition frame. We uniformly sample eight frames from this interval and apply SAM3 [74] to segment the target object using the specified label. For each sampled frame, the highest-confidence mask is selected. In cases where SAM3 fails to produce a valid segmentation, we examine nearby frames within the same interval to recover a suitable mask.

Since SAM3D [76] reconstructs geometry from a single RGB image but does not recover metric scale, we first estimate metric depth for the initial frame using DepthAnything3 [75] and back-project it into a 3D point map using the known camera parameters. We then apply the segmentation mask to extract the object point map, which provides the metric reference. Given this metrically scaled point map and the initial RGB frame, we reconstruct the object mesh with SAM3D. Given the pose of the RGB image, we further transform the reconstructed metric mesh from SAM3D into the transition frame's world coordinates using the known camera extrinsics.

C. Step 3: hand-object alignment

Despite sequential video observations from multiple cameras in EgoDex [24], misalignment persists between the reconstructed object and the tracked hand keypoints in the world frame. We align the hand and object based on two hypotheses: (i) the fingerpads should be in contact (at least three fingers including the thumb) with the object surface at the moment of grasp completion, and (ii) no deep penetration should occur.

At the transition frame between the grasping and manipulation stages, we first fit the MANO hand mesh from the tracked hand keypoints using optimization-based inverse kinematics (IK) [91], which allows us to extract the fingerpad vertices from the mesh. We then estimate an optimal object translation offset $\Delta \boldsymbol{p}^o \in \mathbb{R}^3$ that minimizes a weighted fingerpad-to-object surface discrepancy at the transition frame by gradient descent. If the resulting grasp does not satisfy the contact hypotheses, we further fine-tune the MANO parameters correction $\Delta\boldsymbol{\theta}$ with:

$$\min_{\Delta \boldsymbol{p}^o, \Delta\boldsymbol{\theta}} \mathcal{L}_{\text{dist}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{S5}$$

where $\Delta \boldsymbol{p}^o \in \mathbb{R}^3$ is the object translation offset, $\Delta\boldsymbol{\theta} \in \mathbb{R}^{24}$ is the MANO parameters correction in PCA space, $\mathcal{L}_{\text{dist}}$ penalizes MANO fingerpad to object surface distances, $\mathcal{L}_{\text{pen}}$ penalizes negative signed distances of fingerpad vertices, and $\mathcal{L}_{\text{reg}}$ regularizes deviation from the IK-fitted pose.

Finally, we propagate both the estimated object translation offset $\Delta \boldsymbol{p}^o$ and the optimized MANO pose correction $\Delta\boldsymbol{\theta}$ at the transition frame to all frames prior to the transition, resulting in aligned hand-object data for the entire grasping phase.

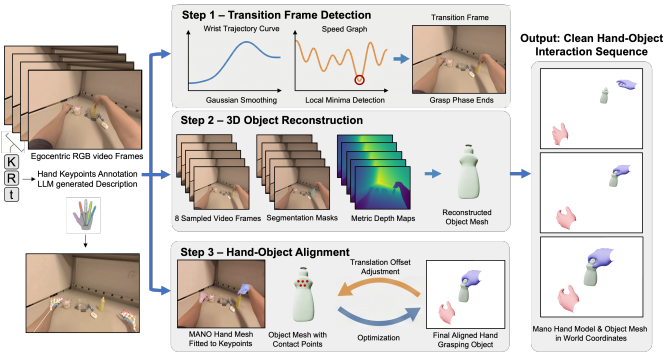

Fig. 5: **Detailed hand-object data reconstruction pipeline.** Given an egocentric RGB video with tracked hand keypoints and camera parameters, we reconstruct HOI data in three steps. (1) *Transition frame detection*: we smooth wrist trajectories and identify the grasp-to-manipulation transition via local speed minima and direction change. (2) *3D object reconstruction*: we segment the target object with SAM3 [74], estimate metric depth with DepthAnything3 [75], and reconstruct a mesh with SAM3D [76] using pre-transition frames where the object is static. (3) *Hand-object alignment*: we fit MANO hand meshes via inverse kinematics and optimize an object translation offset to satisfy fingerpad contact and non-penetration constraints at the transition frame, then propagate the alignment to all frames.

TABLE IV: Geometric and pose accuracy of our EgoDex reconstruction pipeline measured on HOT3D's ground-truth meshes and 6-DoF poses.

Metric	Mean	Median	Std
Chamfer distance	9.8 mm	6.5 mm	21.8 mm
F-score @ 5 cm	96.5%	100.0%	10.3%
Position error	5.1 cm	3.6 cm	6.4 cm
Mask IoU	0.70	0.78	0.24

D. Reconstruction quality validation on HOT3D

We validate the EgoDex reconstruction pipeline on HOT3D, which provides ground-truth object meshes and 6-DoF poses, by running the same pipeline on HOT3D's egocentric RGB streams and comparing against the released annotations. Table IV summarises the result over the public training split.

A median Chamfer distance of 6.5 mm and a mean F-score@5cm of 96.5% indicate that nearly all reconstructed

surface points lie within 5 cm of the ground truth, with sub-centimeter median geometric deviation. The median 6-DoF position error of 3.6 cm is sufficient for initializing the grasping stage, and the upstream SAM3 segmentation (mask IoU median 0.78) confirms reliable object isolation in the egocentric stream.

E. Dataset preprocessing details

This section details our preprocessing and annotation procedures for each dataset used in Sec. IV-A of the main paper.

A. GRAB

GRAB [26] is a comprehensive full-body grasping and manipulation dataset containing 1,335 sequences of human interactions with 51 everyday objects. Each sequence provides high-quality 3D motion capture data for both the human body and hands, together with accurate object meshes and trajectories. Although GRAB does not include scene information, we include it in our evaluation due to the high fidelity of its hand-object interaction data and the complexity of the manipulation actions it contains.

Following [18], we focus exclusively on the hand-object interaction components and discard full-body motion information. We adopt the dataset split (1,125 train / 210 test) and textual action descriptions introduced in [18]. The raw motion capture data is downsampled from $\sim$120 Hz to 20 Hz.

B. HOT3D

For training our semantically grounded and scene-aware manipulation model, there is currently no dataset that simultaneously satisfies all of the following requirements: (i) high-quality 3D annotations for hand-object interactions [73, 26], (ii) high-fidelity real-world scene capture comparable to large-scale indoor datasets [103, 104], and (iii) explicit action-level semantic descriptions.

HOT3D [29] is close to meeting these requirements. The dataset provides accurate 3D hand and object pose annotations synchronized with egocentric videos, enabling detailed analysis of fine-grained manipulation activities.

In our experiment, we use the publicly available training split of HOT3D, as the ground-truth annotations for the official test set have not been released and are primarily intended for benchmark evaluation. The training split consists of 136 recordings collected from nine subjects, with each recording lasting approximately two minutes and containing multiple hand-object interactions.

To obtain scene representations, we reconstruct each recording using 3DGS [105, 81] from the recorded egocentric videos captured by the Project Aria glasses [106]. We then automatically extract hand-object interaction clips by detecting object motion and temporally segmenting the sequences accordingly. Specifically, each interaction clip is divided into a grasping stage and a manipulation stage. The grasping stage is defined as the interval from the end of the previous interaction to the moment of grasp completion. The manipulation stage covers the remaining period where the hand actively manipulates the object.

Following this procedure, we obtain 1,802 interaction clips in total, split into 1,441 for training and 361 for testing, where the test set contains interactions involving unseen objects. Finally, we use OpenAI GPT-4o [87] to generate natural-language descriptions for each extracted interaction clip, including the objects involved and the actions performed.

HOT3D evaluation noise and LatentHOI reproduction. Compared to GRAB's mocap supervision, HOT3D's evaluation regime carries three additional sources of noise: roughly 20% of object poses are missing in the public training split, the initial hand pose is estimated by an off-the-shelf tracker rather than mocap-tracked, and the 3DGS scenes contain reconstruction artifacts (mesh holes, texture seams, transient floaters). These factors compound at the simulator level: the Real Mocap upper bound on HOT3D physics-simulation SR is only 6.67% (Table I), so absolute SR is dominated by simulator and supervision noise rather than by generator quality.

LatentHOI [19] collapses to 0% physics-simulation SR on HOT3D (Table I), which we attribute to its autoregressive latent decoder. Small early-frame errors are amplified through the rollout, and under HOT3D's noisier initialization the latent decoder converges on near-static, contact-avoiding sequences that satisfy the heuristic Phy criterion (the object stays above the ground) without any meaningful contact. To rule out a broken retraining as the cause, we also evaluated LatentHOI under its original CLIP text encoder and observed the same HOT3D collapse. On GRAB, LatentHOI's SR (28.44%) is consistent with the trends reported by the original authors, indicating that the CLIP$\rightarrow$T5 swap applied uniformly to all baselines for fairness is not the source of the HOT3D failure.

F. Physics simulation analysis

This section provides additional details on the physics simulation experiments described in Sec. IV-A of the main paper.

We evaluate the realizable physical feasibility of generated HOI sequences in Isaac Gym [92]. The evaluation pipeline consists of three steps: (1) retargeting the generated MANO [72] hand trajectories to the Allegro Hand [95] via inverse kinematics [91], (2) executing the retargeted motions using a physics-based tracking controller [93] that produces robot joint torques, and (3) evaluating whether the object remains stably grasped and the task is completed. We describe each step in detail below.

Retargeting. Given the generated MANO hand trajectories, we first convert the MANO parameters to 21 keypoint positions per hand at each frame via forward kinematics. Since the Allegro Hand is approximately 1.6$\times$ larger than the human hand, we uniformly scale the MANO keypoints accordingly to match the robot hand size. We then solve for the 16-DOF Allegro joint angles and the 6-DOF wrist pose via inverse kinematics using pyroki [91], a JAX-based least-squares IK solver. (The Allegro Hand has 4 fingers with 4 DOF each = 16 DOF, while MANO models 5 fingers with 3 PCA-coupled DOF each plus a 6-DOF wrist root, hence the structural mismatch we resolve via the offset described below.)

The optimization minimizes a weighted combination of cost terms: (i) a *local alignment* cost that preserves relative inter-joint distances and angles to maintain hand structure, (ii) a *global alignment* cost that matches the robot link positions to the target keypoints, (iii) *wrist position and rotation* costs that anchor the robot wrist to the MANO wrist pose using translation error and SO(3) geodesic distance respectively, and (iv) *temporal smoothness* costs on both the joint angles and the root trajectory to suppress jitter. Joint angle limits from the Allegro URDF are enforced as hard constraints. Since the Allegro Hand has four fingers while MANO models five, we approximate the pinky targets by offsetting the ring finger keypoints laterally.

Tracking. We use DexTrack [93], a reinforcement learning-based tracking controller, to execute the retargeted trajectories under physics simulation. The controller is trained with Proximal Policy Optimization (PPO) in Isaac Gym [92] across thousands of parallel environments. The policy observes the current hand joint positions and velocities, fingertip states, object pose, and the residual between the current state and the reference trajectory at the current and future timesteps. It outputs cumulative residual position targets for the 16 actuated hand joints, which are tracked by low-level PD controllers. The reward function encourages (i) minimizing the distance between the fingertips and the object surface, (ii) tracking the reference hand pose with separate coefficients for global translation, wrist rotation, and finger joint angles, (iii) matching the target object position and orientation, and (iv) smooth joint velocity profiles. To improve robustness, we apply domain randomization over hand joint stiffness and damping, hand and object masses, friction coefficients, and observation and action noise during training.

Evaluation. As shown in Fig. 6, although the kinematic trajectories of both the human hand (a) and the retargeted Allegro Hand (b) appear visually plausible, executing them under physics simulation (c) can reveal failures: the object may detach or slip from the grasp due to insufficient contact forces, incorrect friction modeling, or dynamically unstable grasp configurations. This demonstrates that kinematic feasibility alone does not guarantee physically stable interactions, motivating the use of physics simulation as a more rigorous evaluation protocol.

For each method, we run 10 simulation trials per test sequence and report the average success rate and holding time. An episode is considered successful if the object remains within 5 cm of its target position while maintaining finger contact throughout the trajectory. All simulation parameters follow the default settings from DexTrack [93].

G. Real-world experiment details

This section provides additional details on the real-world dexterous manipulation experiments showcased in the main paper.

Hardware Setup. Our platform consists of two Franka Emika Panda robotic arms [94], each equipped with an Allegro Hand v5 [95]. The 3D scene is reconstructed offline prior to each experiment: we capture a multi-view video sequence using a MetaCam [107] scanner and reconstruct the scene via Gaussian-LIC [108], yielding a 3DGS representation consistent with our training pipeline.

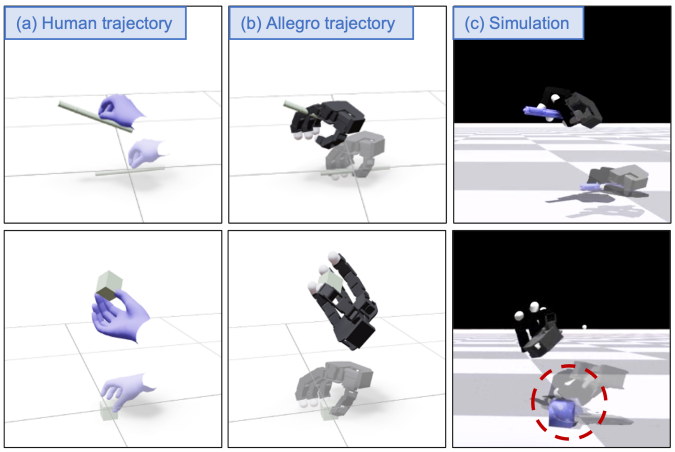

Fig. 6: Comparison of HOI trajectories: (a) kinematic human hand trajectory, (b) retargeted Allegro Hand trajectory, and (c) physics simulation in Isaac Gym [92]. Although the kinematic trajectories in (a) and (b) appear plausible, executing them under physics simulation reveals failures: the object detaches and slips from the grasp (bottom row, red circle), indicating that kinematically feasible motions do not guarantee physically stable interactions.

Retargeting Pipeline. Given the generated MANO hand trajectories, we decompose the retargeting into two stages: (1) *Arm retargeting*: the generated MANO wrist 6D pose at each frame is used as the end-effector target for the Franka arm. We solve the 7-DOF joint angles using the Franka built-in inverse kinematics solver. (2) *Hand retargeting*: the MANO finger articulations are retargeted to the 16-DOF Allegro Hand joint angles using pyroki [91], following the same procedure described in Sec. F. The retargeted trajectories are further refined offline by the DexTrack [93] sim-to-real policy, which was trained entirely in Isaac Gym [92] and deployed without additional fine-tuning on the real robot.

Execution. The refined joint trajectories are executed in an open-loop fashion via a standard joint impedance controller on both the Franka arms and the Allegro Hands at 1 kHz. Since the execution is open-loop, no real-time object state feedback is required during task execution.

Evaluation. We evaluate the system on four contact-rich household manipulation tasks: drinking from a cup, pouring liquid between containers of different sizes, tilting a container, and squeezing dressing. Across all tasks, the generated HOI trajectories are consistently retargeted and successfully executed, producing stable contact-rich interactions that match the intended behaviors (Fig. 7, per-task success rates in Table V). We refer the reader to the supplementary video for animated results.

Task diversity (joint-angle analysis). The four real-world tasks are not kinematically identical. Pouring and tilting demand distinct wrist rotation ranges and grasp-stability con-

TABLE V: Real-world execution success rate (SR, out of 10 trials) on four contact-rich tasks. Aggregated over 40 trials, FlowHOI achieves 57.5% $[42.2\%, 71.5\%]$ vs. DiffH2O 10.0% $[4.0\%, 23.0\%]$ (95% Wilson CI), and the intervals do not overlap.

Method	Pouring	Drinking	Tilting	Squeezing	Avg.
DiffH2O [18]	2/10	0/10	2/10	0/10	10.0%
FlowHOI (Ours)	**5/10**	**4/10**	**7/10**	**7/10**	**57.5%**

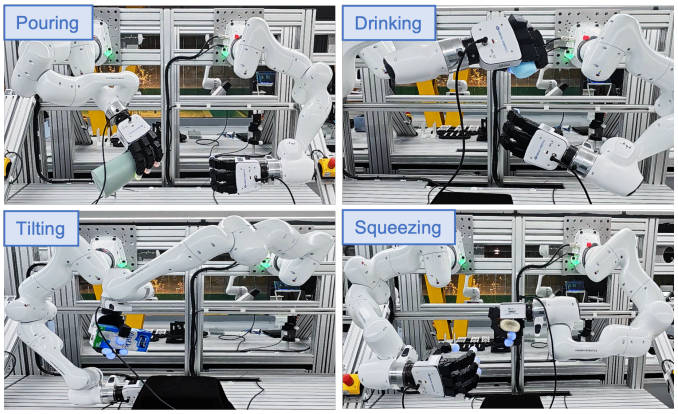

Fig. 7: Real-world execution on four contact-rich tasks (pouring, drinking, tilting, squeezing): the generated HOI is retargeted to a Franka Panda + Allegro Hand and run open-loop.

straints. Drinking is a compound lift-tilt-return trajectory. Squeezing requires coordinated finger flexion with distinct joint targets and force profiles.

H. Scene Representation Details and Analysis

This section provides additional details on the scene representation described in Sec. III-C of the main paper.

Fig. 8 visualizes the geometric features (Concerto [82]) and semantic features (SceneSplat [83]) of two example scenes in top-down view, colored by PCA of the respective feature spaces. The geometric features (left column) primarily encode spatial properties such as surface normals, curvature, and local point distributions. As highlighted by the black dashed circles, geometric features can distinguish objects on a table from the table surface itself based on their different spatial structures. The semantic features (right column) encode language-aligned semantics, grouping points by object category or functional role. Points on the same object share coherent representations regardless of spatial location, while semantically different objects are clearly separated (black dashed circles). This complementarity motivates our bidirectional fusion: geometric features provide precise spatial grounding for localization and collision avoidance, while semantic features enable the model to identify task-relevant objects for text-conditioned generation.

I. Evaluation Metric Details

This section provides detailed definitions and implementation details of the evaluation metrics reported in Sec. IV-A of the main paper.

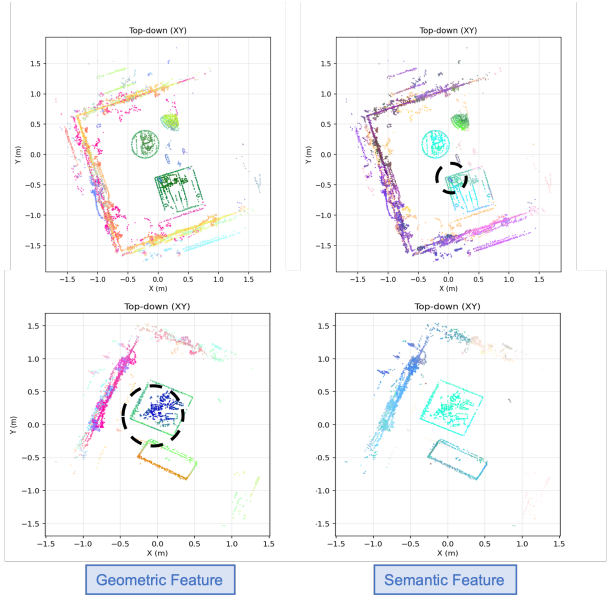

Fig. 8: **Scene feature analysis.** Top-down views of two scenes comparing geometric features (Concerto [82], left) and semantic features (SceneSplat [83], right), colored by PCA of the respective feature spaces. Black dashed circles highlight where geometric features successfully distinguish objects from the supporting surface (bottom-left) and where semantic features successfully discriminate semantically distinct regions (top-right). This complementarity motivates our bidirectional fusion.

A. Physical interaction quality

Let $\mathcal{M}_h^t$ and $\mathcal{M}_o^t$ denote the hand and object meshes at frame t, respectively. For a hand mesh vertex $\mathbf{v}$, we denote by $d(\mathbf{v}, \mathcal{M}_o^t)$ the signed distance to the object surface, where negative values indicate penetration. All physical interaction metrics (IV, ID, CR, IVU) are computed independently for each hand and then averaged across hands.

Interpenetration Volume (IV). Interpenetration volume measures the volumetric intersection between the hand and object meshes:

$$\mathrm{IV}_t = \mathrm{Vol}\big(\mathcal{M}_h^t \cap \mathcal{M}_o^t\big)\,, \tag{S6}$$

and is averaged over time:

$$\mathrm{IV} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{IV}_t. \tag{S7}$$

Interpenetration Depth (ID). Interpenetration depth is defined as the mean penetration distance across all penetrating hand vertices:

$$\mathrm{ID}_t = \frac{1}{|\mathcal{V}_{\mathrm{pen}}^t|} \sum_{\mathbf{v}\in\mathcal{V}_{\mathrm{pen}}^t} \max\big(0, -d(\mathbf{v}, \mathcal{M}_o^t)\big)\,, \tag{S8}$$

where $\mathcal{V}_{\mathrm{pen}}^t = \{\mathbf{v} \in \mathcal{V}_h \mid d(\mathbf{v}, \mathcal{M}_o^t) < 0\}$ is the set of penetrating vertices at frame t. ID is temporally averaged over all frames.

Contact Ratio (CR). The contact ratio quantifies the proportion of hand vertices that are close to the object surface:

$$\mathrm{CR}_t = \frac{1}{|\mathcal{V}_h|} \sum_{\mathbf{v} \in \mathcal{V}_h} \mathbb{I}\big(d(\mathbf{v}, \mathcal{M}_o^t) \le \delta\big)\,, \tag{S9}$$

where δ is a fixed distance threshold (set to $5\,\mathrm{mm}$ in all experiments). The final CR score is obtained by averaging over interaction frames (frames where hand-object contact occurs).

Interpenetration Volume per Contact Unit (IVU). Since low penetration can be trivially achieved by avoiding contact, we additionally report interpenetration volume normalized by the contact area at each frame. For each contact frame t (where $\mathrm{CR}_t > 0$), we compute the per-frame IVU as:

$$\mathrm{IVU}_t = \frac{\mathrm{IV}_t}{A^t_{\mathrm{contact}}}, \quad A^t_{\mathrm{contact}} = \mathrm{CR}_t \cdot \mathrm{Area}(\mathcal{M}_h), \tag{S10}$$

and the final IVU is the mean over all contact frames:

$$\mathrm{IVU} = \frac{1}{|\mathcal{T}_c|} \sum_{t \in \mathcal{T}_c} \mathrm{IVU}_t, \tag{S11}$$

where $\mathcal{T}_c = \{t \mid \mathrm{CR}_t > 0\}$ is the set of frames with non-zero contact.

B. Motion quality

Action Recognition Accuracy (AR). Following IMoS [39], we pretrain a 4-layer Transformer encoder with a `[CLS]` classification head on each dataset's training split (29 action classes on GRAB, 14 on HOT3D), holding it fixed thereafter and using it to score every method's generated trajectory under identical conditions. The classifier consumes the bimanual MANO joint-position sequence, augmented with the object-pose sequence on HOT3D where scene context is available. AR is the top-1 classification accuracy on the test split:

$$\mathrm{AR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\arg\max f_{\mathrm{act}}(\hat{\mathbf{x}}_i) = y_i)\,. \tag{S12}$$

Sample Diversity (SD). For each conditioning input, we generate K motion samples and compute the mean pairwise ℓ_2 distance between their hand joint trajectories, normalized by the number of frames:

$$\mathrm{SD} = \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} \frac{1}{T'} \left\| \mathbf{h}^{(i)} - \mathbf{h}^{(j)} \right\|_2, \tag{S13}$$

where $\mathbf{h}^{(i)} \in \mathbb{R}^{2J \cdot 3 \cdot T'}$ is the flattened joint position trajectory of the i-th sample. We use $K = 4$ in our experiments.

Overall Diversity (OD). Overall diversity is computed as the mean pairwise ℓ_2 distance between hand joint trajectories across all generated test samples, normalized by the number of frames:

$$\mathrm{OD} = \frac{2}{S(S-1)} \sum_{1 \le i < j \le S} \frac{1}{T'} \left\| \mathbf{h}_i - \mathbf{h}_j \right\|_2, \tag{S14}$$

where S is the total number of generated motions.

C. Realizable physical feasibility

Physical Plausibility (Phy). Following LatentHOI [19], physical plausibility is a heuristic, per-sequence binary score. A sequence is deemed physically plausible if (i) at least one hand joint remains within a signed distance of $5\,\mathrm{mm}$ to the object surface for every frame during the interaction phase, and (ii) the object stays above the ground plane throughout the sequence. The final Phy score is the percentage of test sequences satisfying both criteria. Note that this metric only checks kinematic proximity and does not account for dynamic stability. As discussed in Sec. F, a high Phy score does not necessarily imply physically realizable interactions.

Success Rate (SR). For simulation evaluation, a trial is considered successful if the object is lifted above the ground plane by at least $5\,\mathrm{cm}$ while maintaining contact with the hand. The success rate is the fraction of such sequences:

$$\mathrm{SR} = \frac{1}{N} \sum_{n=1}^{N} s_n, \tag{S15}$$

where $s_n \in \{0, 1\}$ indicates whether the object in the n-th trial is lifted off the ground ($> 5\,\mathrm{cm}$) with sustained hand-object contact.

Holding Time (HT). Holding time measures the duration during which objects are stably held in the hand and is averaged over successful trials:

$$\mathrm{HT} = \frac{\sum_{n=1}^{N} s_n\, \tau_n}{\sum_{n=1}^{N} s_n}, \tag{S16}$$

where τ_n denotes the duration during which the object is stably held in the n-th trial.

D. Ablation-specific metrics

Grasp Error (GE). Grasp error measures the mean per-joint Euclidean distance between the generated and ground-truth hand poses at the transition frame t_g:

$$\mathrm{GE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \frac{1}{|\mathcal{J}^{(i)}|} \sum_{j \in \mathcal{J}^{(i)}} \left\| \hat{\mathbf{j}}^{(i)}_{t_g,j} - \mathbf{j}^{(i)}_{t_g,j} \right\|_2, \tag{S17}$$

where $\hat{\mathbf{j}}^{(i)}_{t_g,j}, \mathbf{j}^{(i)}_{t_g,j} \in \mathbb{R}^3$ are the generated and ground-truth 3D positions of the j-th MANO joint at the transition frame, and $\mathcal{J}^{(i)}$ is the set of joints belonging to the grasping hand(s), determined by proximity to the object at the ground-truth transition frame.

Final Displacement Error (FDE). Final displacement error measures the Euclidean distance between the generated and ground-truth object poses at the end of the manipulation stage:

$$\mathrm{FDE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left\| \hat{\mathbf{O}}^{(i)}_{N-1} - \mathbf{O}^{(i)}_{N-1} \right\|_2, \tag{S18}$$

where $\hat{\mathbf{O}}^{(i)}_{N-1}$ and $\mathbf{O}^{(i)}_{N-1}$ denote the generated and ground-truth object poses at the final frame.

Jerk. Jerk measures the smoothness of generated object motion as the third derivative of position with respect to time, i.e., the rate of change of acceleration. Human motion naturally

TABLE VI: Effect of stage-transition soft inpainting on GRAB. Inpainting suppresses high-frequency artifacts at the grasp/manipulation boundary by an order of magnitude.

Setting	Jerk_{pos} (↓)	Jerk_{ang} (↓)	$\text{Acc}_g^{\text{pos}}$ (↓)	GE $[m]$ (↓)
w/o inpainting	0.341	3.218	0.157	0.14
w/ inpainting	**0.044**	**0.407**	**0.020**	**0.06**

follows a *minimum jerk* principle [109], and lower jerk values indicate smoother, more natural trajectories. Given the object trajectory $\mathbf{p}^t \in \mathbb{R}^3$ over T uniformly sampled frames, we compute the positional jerk via finite differences:

$$\text{Jerk}_{\text{pos}} = \frac{1}{T}\sum_{t=1}^{T}\left\|\frac{d^3\mathbf{p}^t}{dt^3}\right\|_2, \tag{S19}$$

where derivatives are approximated using central differences (`np.gradient`). We analogously define the angular jerk Jerk_{ang} over the object rotation trajectory $\boldsymbol{\theta}^t \in \mathbb{R}^3$ (axis-angle representation).

Hand Acceleration. To assess the stability of generated hand motion, we compute acceleration metrics that separately evaluate global wrist movement and local finger articulation. Given the MANO joint positions $\mathbf{J}^t \in \mathbb{R}^{J\times 3}$ at frame t, we decompose them into wrist positions $\mathbf{p}_w^t \in \mathbb{R}^3$ and finger positions $\mathbf{p}_f^t \in \mathbb{R}^{(J-1)\times 3}$. Local finger positions are computed relative to the wrist: $\tilde{\mathbf{p}}_f^t = \mathbf{p}_f^t - \mathbf{p}_w^t$, isolating finger articulation from global wrist movement. We then compute acceleration via second-order finite differences and report the mean absolute acceleration:

$$\text{Acc}_g^{\text{pos}} = \frac{1}{T}\sum_t\left|\frac{d^2\mathbf{p}_w^t}{dt^2}\right|, \quad \text{Acc}_l^{\text{pos}} = \frac{1}{T}\sum_t\left|\frac{d^2\tilde{\mathbf{p}}_f^t}{dt^2}\right|, \tag{S20}$$

where $|\cdot|$ denotes element-wise absolute value and the result is averaged over all spatial components and both hands. The rotational counterparts $\text{Acc}_g^{\text{rot}}$ and $\text{Acc}_l^{\text{rot}}$ are defined analogously over the wrist and finger axis-angle rotations, respectively. Separating global and local components ensures that wrist trajectory smoothness and finger articulation stability are evaluated independently.

J. Stage-Transition Inpainting Ablation

We ablate the soft inpainting mechanism (Sec. III-C, the $\mathbf{M}\odot\mathbf{x}_\tau^m + (1-\mathbf{M})\odot\mathbf{x}_{[0,t_g]}^g$ replacement during inference) on GRAB. Without inpainting, the manipulation network receives no grasping context, causing velocity discontinuities at the transition frame. Table VI reports motion-smoothness metrics: positional jerk (Jerk_{pos}), angular jerk (Jerk_{ang}), global wrist acceleration ($\text{Acc}_g^{\text{pos}}$), and grasp error (GE).

K. Additional Qualitative Results

We provide additional qualitative results that complement the main-paper comparison (Fig. 3). Fig. 9 provides further qualitative comparisons with DiffH2O [18] and LatentHOI [19] on GRAB [26]. Fig. 10 presents scene-conditioned results on HOT3D [29], demonstrating that the generated trajectories adapt to the surrounding scene layout. **We strongly encourage the reader to view the supplementary video for animated results.**

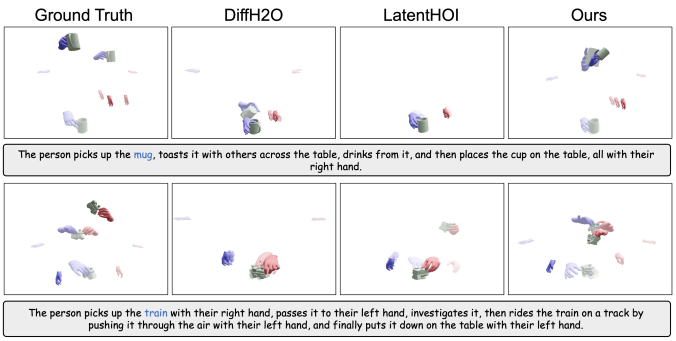

Fig. 9: **Qualitative comparison on GRAB.** Comparison with DiffH2O [18] and LatentHOI [19] on diverse actions. Our method produces more physically plausible hand-object interactions with accurate grasping poses and smooth bimanual coordination.

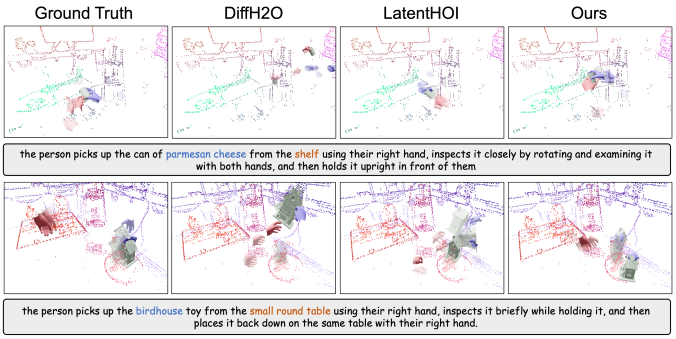

Fig. 10: **Additional qualitative results on HOT3D with scene context.** The generated hand trajectories adapt to the surrounding scene layout, correctly interacting with objects on shelves and tables.

L. Failure Cases

We identify two representative failure modes of our framework (see Fig. 11):

(a) Object Size Mismatch. When the object is too large or too small relative to the hand, the generated grasp may not fully enclose the object, resulting in the object appearing to float in the air without stable contact. A potential improvement is to incorporate explicit object size conditioning or adaptive grasp aperture prediction based on object geometry.

(b) Incorrect Contact Establishment. The hand and object may fail to establish proper contact during the grasping phase, leading to physically implausible configurations where fingers do not properly wrap around the object surface. Future work could incorporate contact-aware losses or physics-based refinement during inference to ensure proper finger-object contact. Additionally, improving the quality of training data with more accurate contact annotations could also help mitigate this issue.

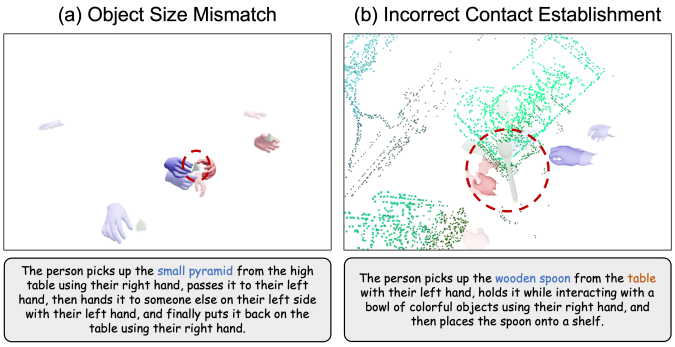

Fig. 11: **Representative failure cases.** Red dashed circles highlight the erroneous regions. (a) *Object size mismatch*: the small pyramid floats away from the hand due to a mismatch between object size and generated grasp aperture. (b) *Incorrect contact establishment*: the hand fails to properly wrap around the wooden spoon, resulting in physically implausible finger-object contact. Left: GRAB [26], right: HOT3D [29] with scene context.