

DexCast: Learning Real-World Dexterous Control from Human Videos

Zihan Gao^{1,*} Yuhong Zhang^{1,2,*} Mengzhe Wang^{3,*} Kaisheng Liu⁴ Shengpeng Li⁵
Siyao Xiao¹ Runqing Cheng⁶ Xiao Lin⁶ Zhifang Liu¹ Haoqian Wang^{1,†}

¹Tsinghua University ²TsingTec ³Fudan University

⁴Shenzhen University ⁵Nanyang Technological University ⁶Georgia Institute of Technology

{zihangao, dsyuhong, wangmengzhe076}@gmail.com

*Equal contribution. [†]Corresponding author.

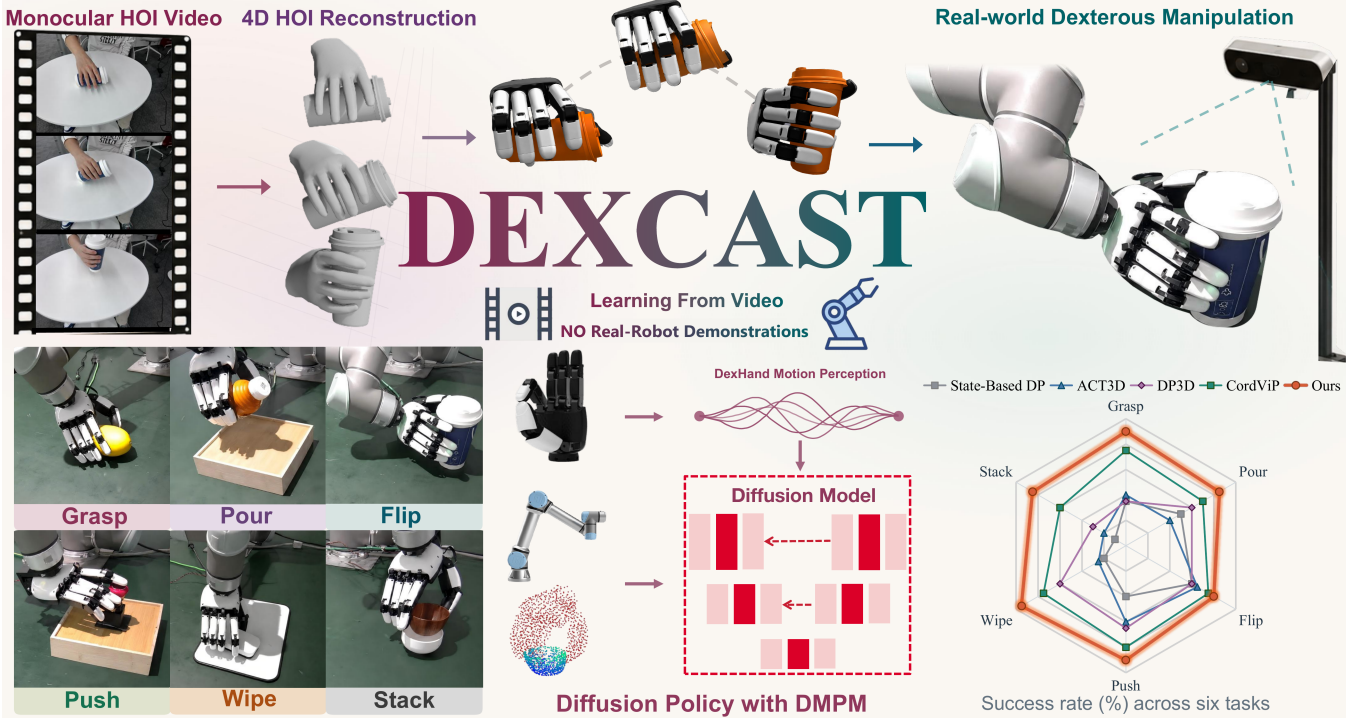


Fig. 1: DEXCAST establishes a new pathway for learning dexterous skills directly from human video demonstrations. DEXCAST leverages 4D hand–object interaction reconstruction to eliminate the need for labor-intensive demonstration collection and introduces a DexHand Motion Perception Module to enhance dexterous manipulation.

Abstract—Recent advances in 3D hand–object interaction (HOI) perception have opened new opportunities for learning dexterous robotic skills from human videos. However, existing dexterous imitation learning methods often rely on costly and manually teleoperated robot demonstrations, requiring substantial human effort. Therefore, we build DEXCAST, a framework that learns dexterous robotic hand control directly from monocular RGB-D videos by leveraging joint HOI reconstructions, without requiring any teleoperation. Furthermore, how to model high-dimensional robotic hand motion and extract temporal context to inform future action generation remains a key challenge. To this end, we propose a DexHand Motion Perception Module (DMPM) to capture inter-joint coordination and temporal dependencies in compact latent representations that condition our diffusion policy. Extensive real-world experiments demonstrate that DEXCAST achieves exceptional performance across diverse dexterous manipulation tasks.

I. INTRODUCTION

Dexterous manipulation with multi-fingered robotic hands has recently achieved rapid progress across diverse tasks

such as grasping, flipping, and in-hand object manipulation [1], [2], [3], [4], [5]. Despite these advances, acquiring high-quality supervision remains a central bottleneck. Most existing methods rely on carefully teleoperated robot demonstrations, which are costly to obtain, time-consuming, and often coupled to specific robot embodiments, camera viewpoints, and scene configurations [6], [7], [8], [9], [10]. This bottleneck is further amplified for dexterous robotic hands, where the high-dimensional action space and complex multi-contact dynamics make demonstration collection substantially more difficult. Moreover, lack of haptic feedback and limited sampling rates can disrupt the fluidity of teleoperated demonstrations; policies trained on these demonstrations struggle to capture the fluid human motion patterns.

The aforementioned limitations motivate a shift from robot data collection toward HOI videos, which offer a more scalable and embodiment-agnostic source of supervision for dexterous skill acquisition [4], [11], [12], [13], [14], [15].

However, directly learning robotic manipulation from such videos remains highly challenging. Existing video-based dexterous learning methods typically model hand motion from human videos and imitate corresponding skills via reinforcement learning, but their validation largely remains confined to simulated environments or their real-world deployment lacks closed-loop perception of interaction dynamics [12], [16]. We therefore propose DEXCAST, a framework that directly transfers manipulation skills from human video demonstrations to closed-loop execution on real robots. Moreover, prior works including [11], [12], [17], [18], [19] either treat human videos as visual priors or encode hand motion through factorized joint representations, limiting the emphasis on modeling the inter-joint coupling and temporal context that are essential for dexterous manipulation. To address this limitation, we introduce the DexHand Motion Perception Module (DMPM), which represents the hand as a directed skeletal graph, enabling the module to model kinematic dependencies and coordinated motion patterns for action generation. Leveraging large-scale HOI data, we pretrain DMPM on over 5M frames to learn transferable motion representations from a broad spectrum of behaviors.

We evaluate DEXCAST on six real-world dexterous manipulation tasks using a five-fingered robotic hand. Without using any real-robot demonstrations, DEXCAST achieves strong real-world performance and supports zero-shot cross-scene deployment under variations in object pose, camera viewpoint, and robot initial state. Extensive experiments and ablations demonstrate the effectiveness of video-derived expert demonstrations and proposed visuomotor policy. In summary, our contributions are as follows:

- 1) We present DEXCAST, a new framework for learning from human videos, which transfers dexterous manipulation skills from monocular human videos to real-world control without any on-robot demonstrations.
- 2) We introduce the DexHand Motion Perception Module (DMPM) to capture finer-grained robotic hand motion context, enhancing dexterous manipulation.
- 3) Through extensive real-world experiments, we demonstrate that DEXCAST substantially advances dexterous imitation learning, achieving an average success rate of up to 87.5% and exhibiting strong generalization.

II. RELATED WORK

A. Dexterous Hand Motion Representation

Learning compact representations of high-dimensional hand motion has emerged as an effective approach to dexterous control. Liconti et al. [20] pretrain latent representations of hand-motion subtrajectories for few-shot imitation learning, while VQ-ACE [21] compresses human hand motions into quantized action chunks to facilitate planning and reinforcement learning. More recently, XL-VLA [22] learns a shared latent action space across dexterous hand embodiments. Being-H0 [11] investigates alternative hand-motion tokenization strategies to identify effective representations for downstream manipulation. These studies motivate

transferable motion representations for dexterous manipulation but do not explicitly account for the intrinsic coupling and strong dependencies among finger joints. Our approach emphasizes topology-aware hand-state encoding by explicitly incorporating the robotic hand’s skeletal structure.

B. Geometry-Aware Visuomotor Policies

Recent diffusion-based visuomotor policies integrate 3D representations into conditional action generation to improve spatial generalization. DP3 [23] extracts compact geometric features from sparse point clouds, while 3D Diffuser Actor [24] integrates 3D scene features, language, and proprioception within a denoising transformer. Subsequent approaches enrich these representations with interaction and temporal structure. CordViP [1] incorporates hand–object correspondences and object-centric contact maps for dexterous manipulation, whereas DP4 [25] captures scene dynamics through a dynamic Gaussian world model. Geometric equivariance provides a complementary inductive bias. Equivariant Diffusion Policy [8] enforces $SO(2)$ symmetry in action denoising, and EquiBot [26] incorporates $SIM(3)$ equivariance to support generalization across translations, rotations, and scales. Inspired by this approach, we employ a $SIM(3)$ -equivariant point-cloud encoder to extract geometric representations for policy conditioning.

C. Learning from Human Videos

Human videos have become an increasingly important source of supervision for embodied learning, particularly for imitation learning. However, how to convert such videos into effective supervision for dexterous robotic manipulation remains an open problem. Representative works such as [11], [4], [27], [28] exploit estimated human hand trajectories from large-scale video corpora to learn generalizable manipulation priors. While these methods benefit from broad behavioral coverage, object states are usually absent or only weakly modeled. Moreover, they still require robot demonstrations during post-training, and therefore do not fully eliminate the dependence on robot data. Methods including [12], [16], [29] recover more structured hand-object interaction trajectories from videos, and use the recovered motion to reproduce corresponding dexterous skills in simulation. These methods provide a more explicit geometric bridge between human videos and robotic manipulation. Nevertheless, these methods are either evaluated only in simulation or deployed on robots by simply replaying simulation rollouts, rather than following a closed-loop mechanism that conditions action generation on real-world observations.

III. METHOD

A. Overview

Our goal is to learn deployable real-world dexterous control directly from RGB-D HOI videos, without any real-robot demonstrations. For this purpose, we transform raw HOI videos into structured expert demonstrations for downstream policy training. Second, to provide temporal context and topology-aware motion representations for action generation,

we propose a DexHand Motion Perception Module (DMPM) and incorporate its frozen encoder into our visuomotor policy. Fig. 2 illustrates the overall pipeline.

B. Demonstration Generation

Here, we detail how to derive expert demonstrations for imitation learning from HOI videos through human-to-robot transfer and data augmentation.

1) *Human-to-Robot Demonstration Generation*: We convert a monocular HOI video into a temporally consistent HOI representation using a video-based reconstruction pipeline following [30] and refine the trajectories of both the hand and object via a residual RL policy [2] to obtain physically plausible poses. The resulting trajectories are transformed into expert demonstrations $\mathcal{D}^i = \{\mathbf{P}_t^h, \mathbf{P}_t^o, \mathbf{s}_t, \mathbf{a}_t\}_{t=1}^{T_i}$. $\mathbf{P}_t^h, \mathbf{P}_t^o \in \mathbb{R}^{1024 \times 3}$ are the complete robotic-hand and object point clouds obtained by sampling their respective meshes. The proprioceptive state comprises the hand and arm joint positions and velocities: $\mathbf{s}_t = [\mathbf{q}_t^h; \mathbf{q}_t^r; \dot{\mathbf{q}}_t^h; \dot{\mathbf{q}}_t^r] \in \mathbb{R}^{36}$. Correspondingly, the imitation target $\mathbf{a}_t$ comprises both absolute joint-position targets and joint-velocity targets.

2) *Data Augmentation*: As original RL rollouts provide limited motion diversity, we further augment the data to achieve a broader spatial distribution while preserving the underlying interaction geometry. We randomly perturb the object pose by translating it within a $20 \text{ cm} \times 20 \text{ cm}$ workspace region and rotating it within a 60° angular range. Such structured augmentation introduces controlled variations in object placement, robot initialization, and workspace configuration, thereby improving the downstream policy’s generalization.

C. DexHand Motion Perception Module

1) *DMPM Architecture*: We formulate robotic hand motion perception within a latent-variable framework and instantiate it as a variational autoencoder (VAE) [31], as shown in Fig. 2(b). DMPM encodes robotic hand motion sequences into a compact latent representation $\mathbf{z}$ and decodes it to reconstruct the original motion sequences.

a) *Encoder*: We introduce a *skeleton* into the residual blocks to explicitly encode the topology of the robotic hand, as shown in Fig. 4. Concretely, the skeleton represents the robotic hand as a structured articulated graph, with nodes representing joints and directed edges tracing the kinematic hierarchy that governs local inter-joint dependencies. This topology restricts feature interactions to local neighborhoods on the skeleton graph, enabling the encoder to model inter-joint couplings and coordinated motion patterns beyond factorized joint representations.

b) *Decoder*: We adopt the non-autoregressive Neural Motion Field (NeMF) decoder [32] as our VAE decoder for efficient sampling. Our decoder $\mathcal{M}$ represents robotic hand motion as a continuous function of time conditioned on a latent code: $\mathcal{M} : (t, z) \mapsto J_t$, where J_t denotes the kinematic pose at time t . Given a latent code z , the entire motion sequence can be sampled in parallel by querying the decoder at different timestamps. The effectiveness of the learned latent representation is empirically evaluated

through the kinematic fidelity and temporal coherence of the reconstructed motion.

c) *Training Objective*: The objective function we aim to minimize is defined as:

$$\begin{aligned} \min_{\theta, \phi} \mathbb{E}_{\{\mathbf{q}_t, \dot{\mathbf{q}}_t\}_{t=1}^T \sim \mathcal{D}} & \left[\sum_{t=1}^T \|\mathbf{W}_q(\hat{\mathbf{q}}_t - \mathbf{q}_t)\|_2^2 \right. \\ & + \lambda_v \sum_{t=1}^T \|\mathbf{W}_v(\hat{\mathbf{q}}_t - \dot{\mathbf{q}}_t)\|_2^2 \\ & \left. + \lambda_{\text{KL}} \text{KL}(q_\phi(\mathbf{z} \mid \{\mathbf{q}_t, \dot{\mathbf{q}}_t\}_{t=1}^T) \parallel p(\mathbf{z})) \right]. \end{aligned} \quad (1)$$

Here, $\hat{\mathbf{q}}$ and $\dot{\hat{\mathbf{q}}}$ denote robotic hand poses and velocities respectively. The motion latent code $\mathbf{z}$ is inferred by the topology-aware encoder $q_\phi(\mathbf{z} \mid \{\mathbf{q}_t, \dot{\mathbf{q}}_t\}_{t=1}^T)$ parameterized by ϕ . The metrics $\mathbf{W}_q$ and $\mathbf{W}_v$ are empirically set to balance the importance of different joints, and λ_v controls the trade-off between pose precision and dynamical consistency. The KL divergence regularizes the learned latent distribution toward a standard Gaussian prior $p(\mathbf{z})$.

2) *Training Data*: The data used for training DMPM are derived from publicly available HOI datasets, including GRAB [33], TCDHands [34], DexYCB [35], H2O [36], ARCTIC [37], HOT3D [38], and HoCap [39]. Inspired by [40], we retarget MANO [41] poses to the robotic hand by minimizing discrepancies between their corresponding wrist-to-fingertip vectors. The resulting pretraining corpus comprises over 5M frames spanning a broad repertoire of hand behaviors. Its scale and diversity provide a rich foundation for learning transferable motion representations.

D. 3D Diffusion Policy with DMPM

We formulate dexterous action generation as a conditional diffusion process enhanced by the pretrained DexHand Motion Perception Module (DMPM). In our implementation, given a 32-frame robotic-hand motion sequence H comprising joint positions and velocities, the frozen DMPM encoder produces a topology-aware latent embedding $\mathbf{f}_d \in \mathbb{R}^{256}$. This representation captures inter-joint dependencies and temporal context in hand motion to inform future action generation. We use separate point-cloud encoding branches Φ_h and Φ_o for the robotic-hand clouds $\mathbf{P}_t^h$ and object clouds $\mathbf{P}_t^o$, respectively. Each branch is built upon a SIM(3)-equivariant PointNet as its backbone [26] and outputs rotation-equivariant and invariant latent representations, Θ_R and Θ_{inv} , respectively. The two point-cloud embeddings are fused into the geometric condition $\mathbf{f}_g$, while an MLP maps robot proprioception to $\mathbf{f}_r$. These features are injected into the diffusion network through FiLM layers [42].

Starting from Gaussian noise $\mathbf{a}_K$, the denoising network ϵ_θ iteratively refines the noisy action for K steps to recover the clean action $\mathbf{a}_0$:

$$\mathbf{a}_{k-1} = \alpha_k \left(\mathbf{a}_k - \gamma_k \epsilon_\theta(\mathbf{a}_k, k, \mathbf{f}_g, \mathbf{f}_d, \mathbf{f}_r) \right) + \sigma_k \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (2)$$

where $\mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes Gaussian noise, and α_k , γ_k , and σ_k are step-dependent coefficients determined by the noise

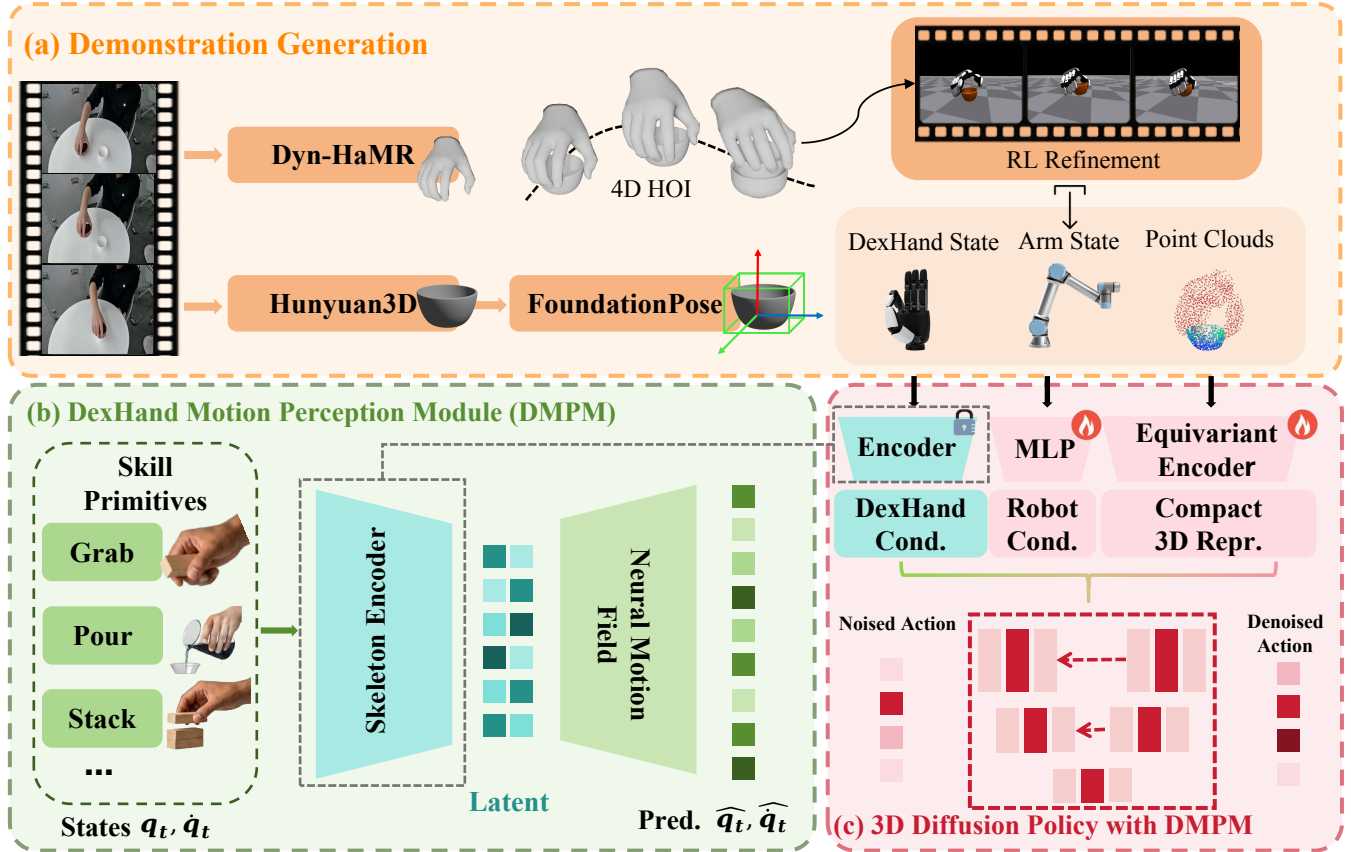


Fig. 2: Overview of DEXCAST. (a) Demonstration acquisition. Given monocular videos, we leverage 4D HOI reconstruction and reinforcement learning to generate training data for dexterous policy learning. (b) We pretrain our VAE-based DexHand Motion Perception Module (DMPM) on large-scale motion capture data. (c) The pretrained DMPM encoder produces a compact latent representation of robotic hand motion, which conditions the diffusion process.

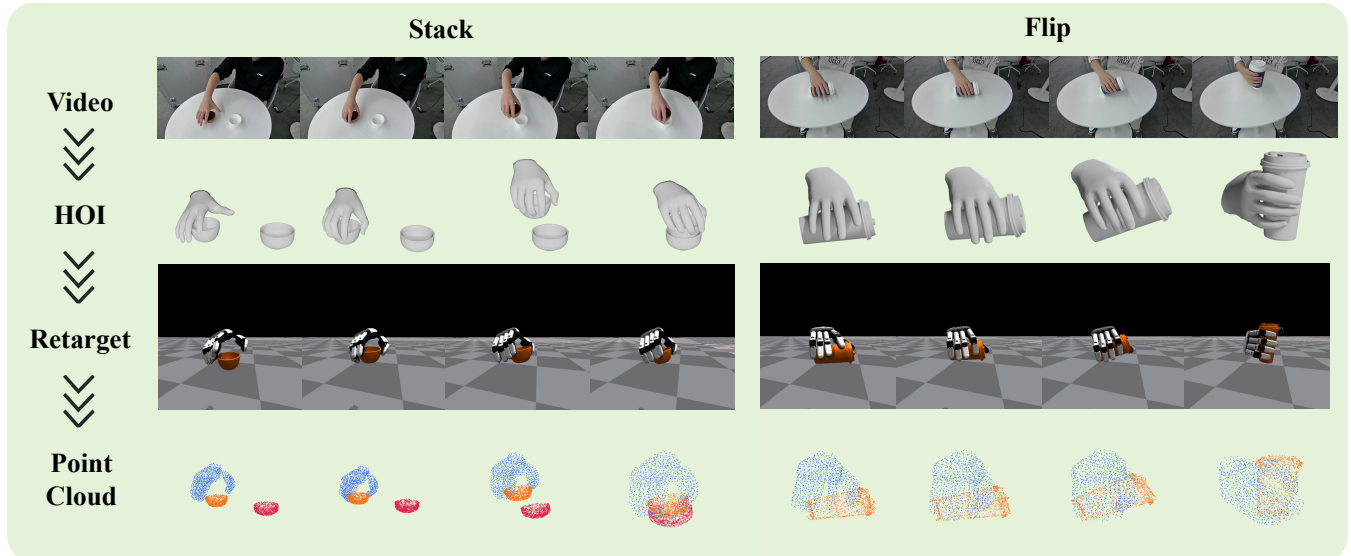


Fig. 3: Visualization of data flow. We recover the full hand–object geometry from raw RGB-D videos, transfer the interaction to a robotic hand, and synthesize complete interaction point clouds for downstream policy training.

scheduler. As the expert demonstrations are derived from human video with a high frame rate, our action chunk size is fixed to be 64. We use a 32-step observation history (≈ 1

s at 30 fps) to predict the next 32 action steps, with a total horizon of 64 steps. We train ϵ_θ using the following objective,

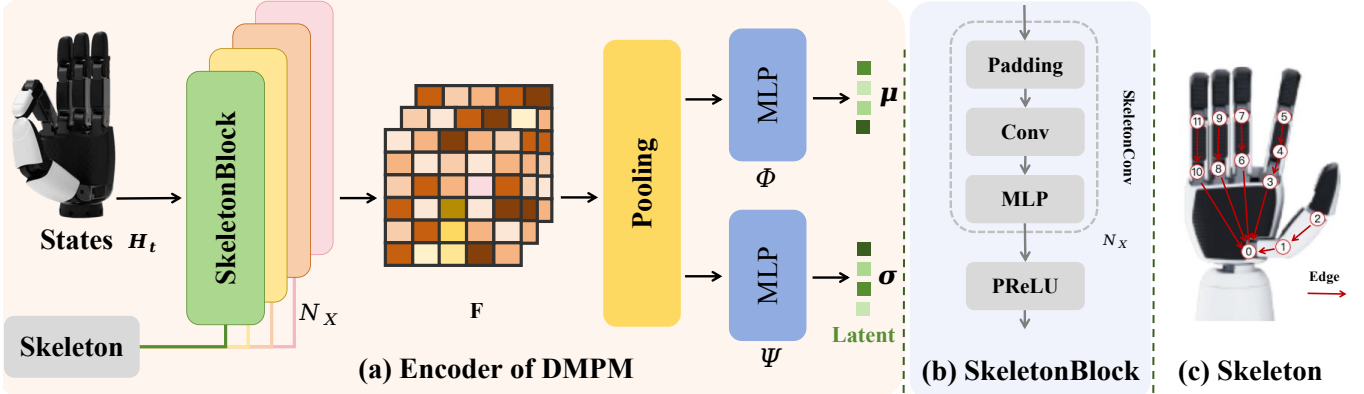


Fig. 4: Details of the DMPM Encoder. (a) Our encoder leverages the articulated kinematic chains of the robotic hand to capture structured inter-joint couplings. (b) SkeletonBlock performs topology-aware residual aggregation along the kinematic graph. (c) Visualization of the skeleton. The skeleton represents the robotic hand as a structured articulated graph, with nodes representing joints and directed edges tracing the kinematic hierarchy.

$$\mathcal{L} = \mathbb{E} \left[\left\| \epsilon - \epsilon_{\theta} \left(\sqrt{\bar{\alpha}_k} \mathbf{a}_0 + \sqrt{1 - \bar{\alpha}_k} \epsilon, k, \mathbf{f}_g, \mathbf{f}_d, \mathbf{f}_r \right) \right\|_2^2 \right]. \quad (3)$$

IV. EXPERIMENTS

A. Experimental Setup

1) *Hardware Setup*: Our system comprises a 6-DoF UR5 robot paired with a 12-DoF five-fingered XHAND robotic hand. A single side-mounted ORBBEC GEMINI 335 RGB-D camera provides object tracking throughout execution and we calibrate the camera extrinsics to the robot base frame.

2) *Task Design*: We design our tasks to cover commonly used human manipulation skills, including pushing, grasping, wiping, flipping, pouring, and stacking. For each task, we generate all expert demonstrations from the same source human video recorded at 30 fps; trajectory horizons range from 150 to 300 steps across tasks. We use a total of 40 demonstrations for training per task, including original RL rollouts and augmented data. We briefly describe each task as follows.

- **Grasp**: Pick up the mango from the tabletop.
- **Pour**: Pour coffee from the bottle.
- **Flip**: Reorient the paper cup to an upright position.
- **Push**: Push the button on the tabletop.
- **Wipe**: Pick up the eraser and clean the whiteboard.
- **Stack**: Stack one cup on top of another.

3) *Baselines*: We benchmark against 3D-based imitation learning algorithms (DP-3D, ACT-3D and CordViP) and a state-based policy. The state-based diffusion policy baseline is conditioned on concatenated robot proprioception and the estimated object 6-DoF pose. DP-3D and ACT-3D are the adaptations of 3D Diffusion Policy [23] and ACT [10], respectively. DP-3D replaces full-scene point clouds with robotic hand-object point clouds, while ACT-3D replaces the image encoder with a point cloud encoder. We emphasize that all point-cloud observations are constructed using our framework, and the number of training point clouds is kept identical across methods. All models are fine-tuned for 10k iterations on each single task, with a batch size of 128, a horizon of 64, 32 observation steps, and 32 action steps.

4) *Deployment Details*: Before deployment, object meshes are reconstructed using Hunyuan3D [43] and scaled to metric units. FoundationPose [44] estimates the object’s 6-DoF pose from RGB-D observations during inference. Points sampled from the object mesh are transformed into the robot base frame using calibrated camera extrinsics to obtain the current object point clouds. The robotic-hand point clouds are generated from joint states through forward kinematics and mesh sampling [45]. We maintain an observation buffer to provide the required input sequence, updating it after each executed action and computing required joint velocities from the buffered joint positions. Consequently, the resulting point clouds are temporally aligned with the proprioceptive state stream. The robot controller executes the complete predicted 32-step action chunk at 8 Hz without considering velocity. After the previous actions are completed, the latest observation is used to predict the next action chunk.

B. Real-world Evaluation

To assess the effectiveness of our framework, we perform real-world evaluations across six tasks. Experiments are conducted under two distinct camera viewpoints: Viewpoint-1 observes the workspace from the front, whereas Viewpoint-2 observes it from the side. For each viewpoint, we execute 20 trials, with the object pose and the arm’s initial state randomized at evaluation time. A trial is considered successful if the robot completes the intended task outcome within the prescribed execution horizon. The results in Tab. I show that our proposed DEXCAST outperforms all baselines across tasks, achieving average success rates of 87.5% and 86.7% under two viewpoints. These remarkable results reflect two complementary successes. First, leveraging precise HOI reconstruction, our framework achieves direct transfer from monocular videos to real-world control, without any real-robot demonstrations. Second, our 3D diffusion policy enhanced with DMPM produces more executable and feasible actions. Fig. 5 visualizes real-world deployment.

Beyond its strong in-distribution (ID) performance, DEXCAST also

Model	Pour		Stack		Avg.
	Seen	Unseen	Seen	Unseen	
DP-3D	60.0	55.0	30.0	30.0	43.8
CordViP	70.0	65.0	60.0	55.0	62.5
Ours	85.0	85.0	85.0	85.0	85.0

TABLE II: Success rates (%) on seen and unseen objects.

TABLE I: Quantitative results on six real-world dexterous manipulation tasks. Each entry reports the success rate (%) under Viewpoint-1 | Viewpoint-2, with the final column reporting the average success rates.

Model	Grasp		Pour		Flip		Push		Wipe		Stack		Avg
State-Based DP	35.0	40.0	50.0	40.0	60.0	50.0	40.0	35.0	20.0	30.0	10.0	25.0	35.8 36.7
ACT-3D	40.0	50.0	40.0	35.0	65.0	70.0	60.0	70.0	25.0	40.0	20.0	30.0	41.7 49.2
DP-3D	35.0	45.0	60.0	50.0	60.0	65.0	65.0	70.0	60.0	55.0	30.0	40.0	51.7 54.2
CordViP	75.0	80.0	70.0	75.0	75.0	70.0	80.0	80.0	75.0	80.0	60.0	65.0	72.5 75.0
Ours	90.0	80.0	85.0	90.0	80.0	85.0	90.0	95.0	95.0	90.0	85.0	80.0	87.5 86.7



Fig. 5: Real-world performance on six dexterous manipulation tasks.

exhibits robust generalization to unseen objects and diverse environments (please refer to the accompanying video for visualizations). As reported in Tab. II, we benchmark DP-3D, CordViP, and our method on the Pour and Stack tasks using both seen and unseen objects. Compared with the baselines, DEXCAST demonstrates markedly stronger adaptability to variations in object shape, maintaining higher success rates under object-level distribution shifts. Moreover, Tab. III highlights another key advantage of our framework. By focusing on interaction-centric geometry, we effectively decouple the policy from background appearance and our policy delivers stronger spatial perception and better performance.

C. Ablation Study

1) *Ablation of DexCast*: We perform fine-grained ablation studies across six tasks to disentangle the individual contributions of the SIM(3)-equivariant architecture and the DMPM. The results displayed in Tab. IV show that these two

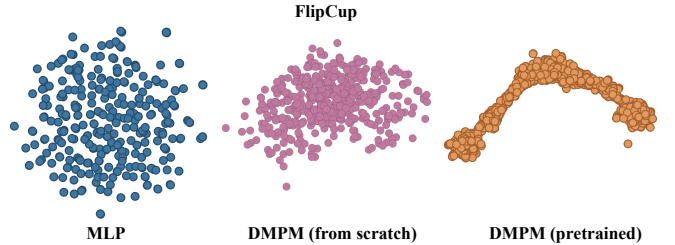


Fig. 6: UMAP visualization of hand-state representations encoded by DMPM and an MLP baseline.

components are effective on their own and complementary when combined, directly supporting higher success rates. We train DEXCAST from scratch to confirm that DMPM’s gains largely stem from large-scale pretraining. Furthermore, our augmentation strategy is critical to achieving high task success rates under our randomized evaluation protocol, as removing it leads to a 61.7% reduction in success rate.

2) *Ablation of the DexHand Motion Perception Module*: We conduct comprehensive ablation studies to systematically analyze the DMPM and inform the final design of DMPM.

To assess the granularity of the learned motion repre-

Model	Pour				Stack				Grasp				Push				Avg.
	Env-1	Env-2	Env-3	Cluttered	Env-1	Env-2	Env-3	Cluttered	Env-1	Env-2	Env-3	Cluttered	Env-1	Env-2	Env-3	Cluttered	
DP-3D	60.0	50.0	45.0	30.0	30.0	25.0	40.0	20.0	35.0	25.0	25.0	20.0	65.0	60.0	55.0	45.0	39.4
CordViP	70.0	65.0	65.0	55.0	60.0	55.0	50.0	40.0	75.0	70.0	75.0	60.0	80.0	70.0	75.0	65.0	64.4
Ours	85.0	85.0	80.0	70.0	85.0	85.0	90.0	70.0	90.0	85.0	85.0	75.0	90.0	90.0	85.0	80.0	83.1

TABLE III: Generalization of DEXCAST to different environments (Env.) (Success Rate %).

Variant	Grasp	Pour	Flip	Push	Wipe	Stack	Avg.
DexCast w/o augmented data	30.0	40.0	20.0	25.0	25.0	15.0	25.8
DexCast w/o SIM(3) & DMPM	60.0	70.0	60.0	75.0	75.0	60.0	66.7
DexCast w/o DMPM	80.0	80.0	70.0	85.0	85.0	70.0	78.3
DexCast w/o SIM(3)	80.0	85.0	75.0	80.0	85.0	80.0	80.8
DexCast (from scratch)	80.0	85.0	60.0	90.0	85.0	80.0	80.0
DexCast (Full)	90.0	85.0	80.0	90.0	95.0	85.0	87.5

TABLE IV: Ablation studies of DexCast. (Success Rate %).

Metric	w/o Skeleton				Skeleton			
	16	32	64	128	16	32	64	128
TMPJPE (mm) ↓	11.497	9.328	8.445	13.141	4.005	3.691	4.540	12.345
MPJPE (mm) ↓	3.355	2.722	2.364	3.835	1.150	1.073	1.316	3.653
Pose Consistency (deg) ↓	2.125	1.811	1.505	2.096	0.776	0.699	0.863	1.941
Hand Jitter (deg/s ²) ↓	0.133	0.215	0.215	0.248	0.320	0.290	0.255	0.288

TABLE V: Ablation studies on skeleton architecture and sequence length. Clip length T varies from 16 to 128 frames.

sentations, we compare DMPM with the MLP-based hand-state encoders commonly adopted in prior work. As shown in Fig. 6, DMPM captures more nuanced patterns in task-specific hand motion, which become more pronounced after pretraining. This comparison suggests that our proprioceptive encoder more effectively models high-dimensional robotic hand motion, providing more informative state representations to condition the policy for action generation.

We further compare encoder variants with and without explicit skeletal modeling. To assess the performance of different variants, we evaluate motion reconstruction using Mean Per-Joint Position Error (MPJPE), Tip Mean Per-Joint Position Error (TMPJPE), Pose Consistency, and Hand Jitter. MPJPE and TMPJPE quantify Cartesian position errors across all hand joints and the fingertip subset, respectively. Pose Consistency assesses the angular reconstruction fidelity, while Hand Jitter measures temporal smoothness using joint angular acceleration. As shown in Table V, incorporating the kinematic structure improves reconstruction accuracy, albeit with a slight reduction in temporal smoothness. Fig. 7 provides a visual comparison of the two variants.

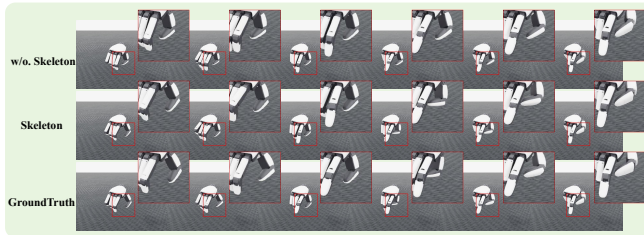


Fig. 7: Visualization of DexHand motion recovery for two DMPM architectures. The DMPM with Skeleton produces higher-fidelity reconstructions.

Metric	32	64	128	256	512	1024	2048
TMPJPE (mm) ↓	6.623	3.370	3.075	3.024	3.402	3.691	4.211
MPJPE (mm) ↓	1.877	0.967	0.893	0.881	0.983	1.073	1.221
Pose Consistency (deg) ↓	1.237	0.639	0.622	0.581	0.656	0.699	0.794
Hand Jitter (deg/s ²) ↓	0.366	0.395	0.388	0.346	0.312	0.290	0.298

TABLE VI: Latent Dimension Ablation. Performance of DMPM under different latent dimensions.

As for the ablations of clip length and latent-code dimensionality, Tab. V shows that the reconstruction quality of the skeleton-based DMPM peaks at a moderate clip length ($T = 32$) and degrades as T increases or decreases. Using the skeleton-based architecture, we further evaluate latent dimensionalities ranging from 32 to 2048 (Tab. VI). The results reveal that increasing the latent size improves reconstruction up to a peak at 256. Larger latent codes provide diminishing returns and can even degrade accuracy while generally producing progressively smoother trajectories.

V. CONCLUSION

Our framework, DEXCAST, establishes a new pathway for learning dexterous skills directly from human video demonstrations. Leveraging a 4D HOI representation and RL refinement, DEXCAST produces training data for policy learning, replacing traditional, labor-intensive demonstration collection. We introduce a DexHand Motion Perception Module (DMPM) to provide a compact, motion-aware conditioning representation for dexterous manipulation. Our 3D diffusion policy enhanced with DMPM demonstrates strong real-world dexterous manipulation performance.

Limitations. Our current 4D HOI reconstruction is limited to rigid objects and remains susceptible to tracking failures under rapid motion and severe occlusion, potentially compromising policy learning and real-world execution. DMPM also remains embodiment-specific, requiring separate skeleton graph design and training for different robotic hands.

REFERENCES

- [1] Y. Fu, Q. Feng, N. Chen, Z. Zhou, M. Liu, M. Wu, T. Chen, S. Rong, J. Liu, H. Dong, *et al.*, “Cordvip: Correspondence-based visuomotor policy for dexterous manipulation in real-world,” *arXiv preprint arXiv:2502.08449*, 2025.
- [2] K. Li, P. Li, T. Liu, Y. Li, and S. Huang, “Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 6991–7003.
- [3] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, Y. Ye, Y. Liang, *et al.*, “Dexgraspvla: A vision-language-action framework towards general dexterous grasping,” *arXiv preprint arXiv:2502.20900*, 2025.

- [4] Q. Li, Y. Deng, Y. Liang, L. Luo, L. Zhou, C. Yao, L. Zeng, Z. Feng, H. Liang, S. Xu, Y. Zhang, X. Chen, H. Chen, L. Sun, D. Chen, J. Yang, and B. Guo, "Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos," *ArXiv*, vol. abs/2510.21571, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:282384634>
- [5] K. Kedia, T. G. W. Lum, J. Bohg, and C. K. Liu, "Simtoolreal: An object-centric policy for zero-shot dexterous tool manipulation," *ArXiv*, vol. abs/2602.16863, 2026. [Online]. Available: <https://api.semanticscholar.org/CorpusID:285787326>
- [6] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki, "Act3d: 3d feature field transformers for multi-task robotic manipulation," *arXiv preprint arXiv:2306.17817*, 2023.
- [7] S. Haldar, J. Pari, A. Rai, and L. Pinto, "Teach a robot to fish: Versatile imitation from one minute of demonstrations," *arXiv preprint arXiv:2303.01497*, 2023.
- [8] D. Wang, S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, and R. Platt, "Equivariant diffusion policy," *arXiv preprint arXiv:2407.01812*, 2024.
- [9] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 5637–5643.
- [10] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.
- [11] 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*, vol. abs/2507.15597, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:280254617>
- [12] 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 *European Conference on Computer Vision*. Springer, 2022, pp. 570–587.
- [13] W. Ye, F. Liu, Z. Ding, Y. Gao, O. Rybkin, and P. Abbeel, "Video2policy: Scaling up manipulation tasks in simulation through internet videos," *ArXiv*, vol. abs/2502.09886, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:276395143>
- [14] H. Chen, T. Dong, T. Wu, L. Wang, Y. Jangir, Y. Niu, Y. Ye, H. Bharadhwaj, Z. Erickson, and J. Ichnowski, "Dexterous manipulation policies from rgb human videos via 3d hand-object trajectory reconstruction," *ArXiv*, vol. abs/2602.09013, 2026. [Online]. Available: <https://api.semanticscholar.org/CorpusID:285454264>
- [15] H. Gupta, G.-Y. Shi, and W. Yuan, "Lucid: Learning embodiment-agnostic intent models from unstructured human videos for scalable dexterous robot skill acquisition," *ArXiv*, vol. abs/2606.11628, 2026. [Online]. Available: <https://api.semanticscholar.org/CorpusID:289166379>
- [16] P. Mandikal and K. Grauman, "Dexvip: Learning dexterous grasping with human hand pose priors from video," in *Conference on Robot Learning*. PMLR, 2022, pp. 651–661.
- [17] K. Shaw, S. Bahl, and D. Pathak, "Videodex: Learning dexterity from internet videos," *ArXiv*, vol. abs/2212.04498, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:254408735>
- [18] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "Dexwild: Dexterous human interactions for in-the-wild robot policies," *ArXiv*, vol. abs/2505.07813, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:278534598>
- [19] Y. Li, X. Wei, J. Luo, Y. Xiao, Y. Bai, G. Zhou, T. Zou, C. Gui, J. Wen, H. Zhang, *et al.*, "Egolive: A large-scale egocentric dataset from real-world human tasks," *arXiv preprint arXiv:2604.23570*, 2026.
- [20] D. Liconti, Y. Toshimitsu, and R. Katzschmann, "Leveraging pre-trained latent representations for few-shot imitation learning on a dexterous robotic hand," *arXiv preprint arXiv:2404.16483*, 2024.
- [21] C. Yang, D. Liconti, and R. K. Katzschmann, "Vq-ace: Efficient policy search for dexterous robotic manipulation via action chunking embedding," *arXiv preprint arXiv:2411.03556*, 2024.
- [22] G. Jiang, Y. Liang, J. Ye, J.-Y. Huang, C. Jing, R. Duan, P. Abbeel, X. Wang, and X. Zou, "Cross-hand latent representation for vision-language-action models," *arXiv preprint arXiv:2603.10158*, 2026.
- [23] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.
- [24] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *arXiv preprint arXiv:2402.10885*, 2024.
- [25] Z. Liu, Y. Wang, K. Wang, L. Liang, X. Xue, and Y. Fu, "Spatial-temporal aware visuomotor diffusion policy learning," *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:280137023>
- [26] J. Yang, Z.-a. Cao, C. Deng, R. Antonova, S. Song, and J. Bohg, "Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning," *arXiv preprint arXiv:2407.01479*, 2024.
- [27] R. Zheng, D. Niu, Y. Xie, J. Wang, M. Xu, Y. Jiang, F. Castañeda, F. Hu, Y. L. Tan, L. Fu, T. Darrell, F. Huang, Y. Zhu, D. Xu, and L. J. Fan, "Egoscale: Scaling dexterous manipulation with diverse egocentric human data," *ArXiv*, vol. abs/2602.16710, 2026. [Online]. Available: <https://api.semanticscholar.org/CorpusID:285726412>
- [28] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen, G. Yang, J. Zhang, S. Yi, G. Shi, and X. Wang, "Humanoid policy human policy," *ArXiv*, vol. abs/2503.13441, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:277103562>
- [29] J. Hsieh, K.-H. Tu, K.-H. Hung, and T.-W. Ke, "Dexman: Learning bimanual dexterous manipulation from human and generated videos," *ArXiv*, vol. abs/2510.08475, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:281951555>
- [30] Y. Zhang, Z. Gao, S. Li, L.-H. Chen, K. Liu, R. Cheng, X. Lin, J. Liu, Z. Li, J. Feng, *et al.*, "Robowheel: A data engine from real-world human demonstrations for cross-embodiment robotic learning," *arXiv preprint arXiv:2512.02729*, 2025.
- [31] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013. [Online]. Available: <https://api.semanticscholar.org/CorpusID:216078090>
- [32] C. He, J. Saito, J. Zachary, H. Rushmeier, and Y. Zhou, "Nemf: Neural motion fields for kinematic animation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4244–4256, 2022.
- [33] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *European Conference on Computer Vision*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:221245720>
- [34] L. Hoyet, K. Ryall, R. McDonnell, and C. O'Sullivan, "Sleight of hand: Perception of finger motion from reduced marker sets," in *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2012, pp. 79–86.
- [35] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox, "DexYCB: A benchmark for capturing hand grasping of objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [36] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2o: Two hands manipulating objects for first person interaction recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 138–10 148.
- [37] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [38] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, R. Newcombe, R. Wang, J. J. Engel, and T. Hodan, "HOT3D: Hand and object tracking in 3D from egocentric multi-view videos," *CVPR*, 2025.
- [39] J. Wang, Q. Zhang, Y.-W. Chao, B. Wen, X. Guo, and Y. Xiang, "Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction," 2024. [Online]. Available: <https://arxiv.org/abs/2406.06843>
- [40] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang, "Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning," *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12 248–12 255, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270923809>
- [41] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.
- [42] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film:

- Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [43] Z. Zhao, Z. Lai, Q. Lin, Y. Zhao, H. Liu, S. Yang, Y. Feng, M. Yang, S. Zhang, X. Yang, *et al.*, “Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation,” *arXiv preprint arXiv:2501.12202*, 2025.
- [44] B. Wen, W. Yang, J. Kautz, and S. Birchfield, “Foundationpose: Unified 6d pose estimation and tracking of novel objects,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 17 868–17 879.
- [45] Z. Wei, Z. Xu, J. Guo, Y. Hou, C. Gao, Z. Cai, J. Luo, and L. Shao, “D(r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping,” *ArXiv*, vol. abs/2410.01702, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:273025803>