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

Huajian Zeng[1] **Lingyun Chen**[1,2] **Jiaqi Yang**[1,3] **Yuantai Zhang**[1]
Fan Shi[4] **Peidong Liu**[5] **Xingxing Zuo**[1,*]

[1] Mohamed bin Zayed University of Artificial Intelligence
[2] Technical University of Munich [3] ShanghaiTech University
[4] National University of Singapore [5] Westlake University

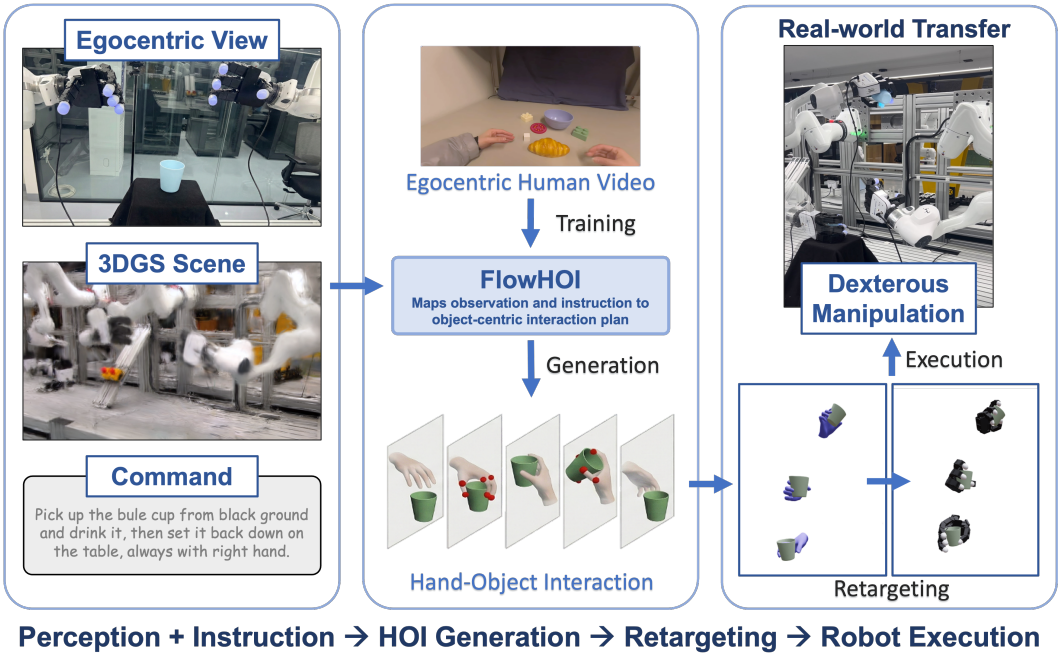

Figure 1: We present a method for generating hand-object interaction (HOI) motions conditioned on egocentric observation, text command, and 3D scene context. We first learn a grasping prior with HOI data extracted from large-scale egocentric videos, and then generate semantically grounded manipulation motions that respect language instructions as well as the surrounding 3D scene context 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

Project page: `https://huajian-zeng.github.io/projects/flowhoi`

*Corresponding author. (Email: `xingxing.zuo@mbzuai.ac.ae`)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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 employing a motion-text alignment loss to semantically ground the generated interactions in both the physical scene layout and the language instruction. To address the scarcity of high-fidelity HOI supervision, we introduce a reconstruction pipeline that recovers aligned hand-object trajectories and meshes 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$\times$ higher physics-simulation success rate on GRAB than the strongest diffusion-based baseline, while delivering a 40$\times$ inference speedup. We further demonstrate real-robot execution across four contact-rich, dexterous manipulation tasks, highlighting the effectiveness of retargeting our generated HOI representations to robotic dexterous hands, achieving significantly higher success rates than prior methods due to improved physical plausibility and semantic alignment.

1 Introduction

Robotic manipulation in everyday household environments is fundamentally interaction-centric [1, 2, 3]: tasks such as opening a lid, pouring from a container, or placing an object are not characterized by the robot's end-effector trajectory alone, but by the underlying context of interaction, namely where contact occurs given the scene geometry, how stable contact is preserved, and how the object's state evolves in a way that is semantically consistent with the language instruction and the 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 is established and transitioned to induce language-specified state changes [14], decoupling manipulation from robot-specific control and facilitating transfer across embodiments.

In this work, we propose FlowHOI, a two-stage flow-matching framework for producing semantically grounded hand-object interaction (HOI) sequences. Given an initial egocentric observation, a language instruction, and a 3D Gaussian splatting (3DGS) scene reconstruction, FlowHOI generates temporally coherent and physically plausible HOI sequences comprising hand poses, object poses, and hand-object contact states, all anchored in the observed scene context and aligned with the language instruction. The resulting HOI representation provides a natural interface for enforcing geometric and physical constraints and can be retargeted to downstream robotic dexterous manipulation, where it facilitates physically plausible execution [15, 16].

Building such a model raises three central challenges: **(i) Geometric consistency & semantic grounding.** Long-horizon interactions are simultaneously constrained by 3D scene geometry and by the language instruction's semantics. Entangling both objectives in a single generator weakens contact establishment (Supp. Sec. J). Inspired by how humans first establish a stable grasp before manipulating an object [17], we decompose generation into a geometry-centric *Grasping* stage that uses a pretrained grasping prior, followed by 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] require tens to hundreds of denoising steps, taking 3–7 s per sequence, which is prohibitive for real-time downstream planning. We adopt conditional flow matching, reducing inference to 0.16 s per sequence (up to 40$\times$ speedup) while maintaining temporally coherent and physically plausible generation. **(iii) Data scarcity.** High-fidelity HOI supervision is scarce because hand-object interactions are high-dimensional, frequently occluded, and tightly coupled with contact dynamics [22, 23]. We address this by introducing a reconstruction pipeline that recovers aligned hand-object trajectories and meshes from large-scale egocentric videos [24]. We use the

resulting data to pretrain the grasping stage, which improves grasp accuracy and the smoothness of the grasp-to-manipulation transition.

To the best of our knowledge, FlowHOI is the *first* method to formulate HOI generation as a unified, conditional flow-matching process in a semantics-grounded way. In summary, our contributions are:

- We introduce a two-stage HOI generation framework that decouples geometry-centric grasping from semantics-centric manipulation and employs flow matching for efficient generation, achieving up to 40$\times$ speedup over prior diffusion-based methods.
- We semantically ground HOI generation by integrating a motion-text alignment loss to enforce consistency with language instructions, and conditioning on a hybrid 3D scene representation that encodes both geometric and semantic context from the surrounding scene.
- We design a reconstruction pipeline to extract large-scale, high-fidelity HOI data from egocentric videos, which we use to pretrain a grasping prior that improves grasp initialization and the grasp-to-manipulation transition.
- On GRAB and HOT3D, FlowHOI attains the highest action-recognition accuracy and a 1.7$\times$ higher physics-simulation success rate over the strongest baseline. Retargeted to a real dexterous robot and executed open-loop, the generated interactions succeed in 57.5% of trials on four contact-rich tasks, compared with 10% for DiffH2O.

2 Related work

Hand-Object Interaction Generation. Synthesizing articulated hand-object motions under realistic contact has been supported by both mocap- and vision-based datasets [25, 26, 27, 28, 29]. Early kinematic methods leverage contact-aware priors, optimization, or implicit representations to reduce interpenetration [30, 31, 32, 33], often with latent-variable models in canonical object spaces [14, 34, 35, 36]. Physics-based RL approaches ensure dynamic feasibility but scale poorly across objects and suffer from sim-to-real gaps [37, 38]. Recent diffusion methods [39, 40, 41] improve temporal coherence, yet DiffH2O [18] still suffers from physical artifacts, semantic inconsistencies, and slow inference, and LatentHOI [19] remains limited in interaction length and data scale. Concurrent work [42] extends diffusion to trilateral human-object-interaction modelling. Object-centric generators synthesize 6-DoF object trajectories in 3D scenes from goal conditions [43] or from egocentric observations with flow matching [44], but do not model the articulated hand. We instead propose a flow matching framework that explicitly models distinct grasping and manipulation phases, achieving faster generation with improved physical plausibility and semantic alignment.

Robot learning and motion priors from videos. Learning manipulation from human videos [45, 46, 47] is scalable but faces an embodiment gap. Egocentric datasets [29, 48, 49] partially bridge it, while VLA models [4, 5, 6], language-conditioned policies [50, 51], and video generators [52, 53, 54, 55] map perception to actions but lack explicit interaction structure. In parallel, diffusion-based human motion synthesis [56, 57, 58, 59, 60, 61] and HOI-extending variants [62, 63, 64, 65] target full-body motion and rely on small mocap datasets [66, 67, 68, 69]. Closely related to our reconstruction pipeline, WHOLE [70] lifts hand-object interactions from egocentric videos into a world-grounded representation, but produces a per-clip reconstruction rather than a generative prior. World-space hand motion estimators [71, 72] likewise recover hand trajectories from egocentric videos, but estimate hand-only motion for each clip rather than learning a prior over hand-object interactions. We instead acquire a robot-agnostic, fine-grained HOI sequence prior recovered from egocentric videos [24] that can be retargeted to different embodiments.

3 Preliminary: flow matching

We briefly review the flow matching framework for generative modeling [73]. Let $q(\mathbf{x})$ denote the unknown data distribution over $\mathbf{x} \in \mathbb{R}^d$ and $p_0(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ a simple prior. Flow matching learns a time-dependent vector field $\mathbf{v}(\mathbf{x}, \tau)$ that transports samples from p_0 to a target distribution $p_1 \approx q$ along a continuous probability path [74]. Since directly matching the marginal vector field of p_τ is generally intractable, *conditional flow matching* (CFM) [73] constructs a tractable conditional path

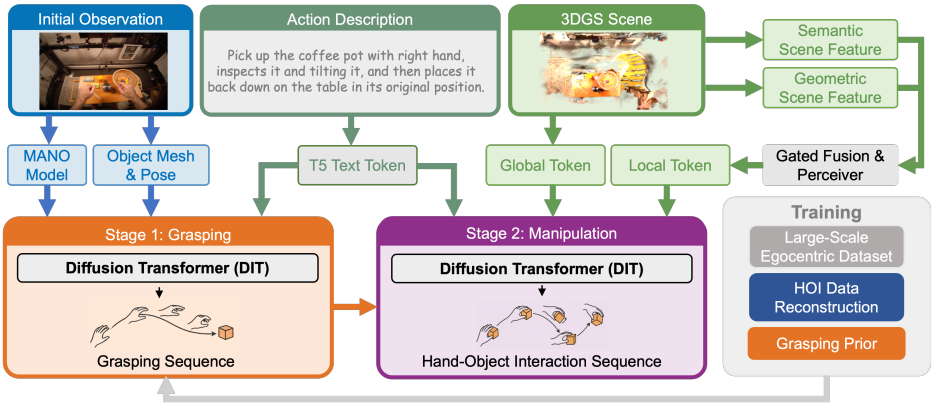

Figure 2: **Overview of FlowHOI.** Two-stage pipeline conditioned on an egocentric observation, text command, and 3D scene: (1) approach-and-grasp from a prior fine-tuned on reconstructed egocentric HOI data, and (2) manipulation conditioned on scene and language.

$p_\tau(\mathbf{x} \mid \mathbf{x}_1)$ for each data sample $\mathbf{x}_1 \sim q$. With the optimal-transport choice of linear interpolation,

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

the conditional vector field admits a closed form:

$$\mathbf{u}_\tau(\mathbf{x}_\tau \mid \mathbf{x}_1) = \frac{\mathbf{x}_1 - (1 - \sigma_{\min})\mathbf{x}_\tau}{1 - (1 - \sigma_{\min})\tau}. \tag{2}$$

A neural 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. Samples are generated at inference by integrating $\mathbf{v}_\theta$ from $\tau = 0$ to $\tau = 1$ with a numerical ODE solver (we use K-step Euler). The role of $\sigma_{\min}$ and discretization details are deferred to Supp. Sec. A.

In FlowHOI, instead of regressing the velocity $\mathbf{u}_\tau$, we adopt the equivalent x-prediction parameterization that directly regresses the clean target $\mathbf{x}_1$ for improved temporal stability [75]. The corresponding analysis is in Supp. Sec. B.

4 Methodology

In this section, we present our flow-based HOI motion generation conditioned on a single-frame initial egocentric observation, text command, and 3D scene context. As illustrated in Fig. 2, our framework consists of two stages: a *Grasping* stage that generates approach-and-grasp motions from a pretrained prior fine-tuned on reconstructed egocentric HOI data, and a *Manipulation* stage that produces subsequent interaction motions conditioned on 3D scene context and language instructions.

4.1 Problem formulation

We address HOI motion generation in practical scenarios. Given the first egocentric observation I, from which we extract the initial hand-object state $\mathbf{x}_{\text{init}}$ and object geometry $\mathcal{M}$, an action description T, and a 3DGS scene representation $\mathcal{G}$, our goal is to generate a temporally coherent hand-object interaction motion over N frames. We denote the object pose at frame t as $\mathbf{O}_t$, and the left and right hand states as $\mathbf{H}_t^l$ and $\mathbf{H}_t^r$, respectively. The full HOI trajectory is defined as

$$\mathbf{x} = \{(\mathbf{H}_t^l, \mathbf{H}_t^r, \mathbf{O}_t)\}_{t=0}^{N-1} \in \mathbb{R}^{N \times D}, \tag{3}$$

where D denotes the per-frame feature dimension of the hand and object representations. Given a dataset of M paired interactions $\{(\mathbf{x}^{(i)}, \mathbf{c}^{(i)})\}_{i=1}^{M}$, our objective is to learn a conditional generative model $p_\theta(\mathbf{x} \mid \mathbf{c})$, where $\mathbf{c}$ is the conditioning information composed of encoded $\mathbf{x}_{\text{init}}$, $\mathcal{M}$, T, and $\mathcal{G}$.

Data representation. Following [18], each frame t contains the object pose $\mathbf{O}_t = (\boldsymbol{p}_t^o, \boldsymbol{r}_t^o) \in \mathbb{R}^9$ with continuous 6D rotation [76], and bimanual MANO [77] hand states $\mathbf{H}_t^h = (\tilde{\boldsymbol{p}}_t^h, \boldsymbol{r}_t^h, \boldsymbol{\theta}_t^h, \mathbf{s}_t^h) \in \mathbb{R}^{54}$ comprising root translation, 6D orientation, 24D MANO PCA pose coefficients, and a per-joint

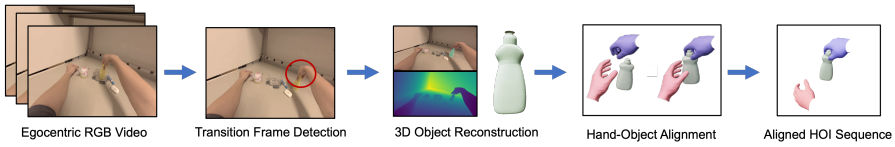

Figure 3: **Hand-object data reconstruction.** Detect the grasp-to-manipulation transition, reconstruct the object mesh via segmentation and metric depth, and align MANO under contact and non-penetration constraints. Detailed pipeline in supp.

signed-distance vector $\mathbf{s}_{t,k}^h = \boldsymbol{j}_k^{h,t} - \Pi_{\mathcal{M}_t}(\boldsymbol{j}_k^{h,t})$ to the posed object surface that encourages contact-aware generation. Hand translations $\tilde{\boldsymbol{p}}_t^h = \boldsymbol{p}_t^h - \boldsymbol{p}_{t_g}^o$ are anchored at the transition-frame object position ($t_g = N_g - 1$) to reduce inter-scene variance. Full notation is in Supp. Sec. A.

4.2 Hand-object data reconstruction from egocentric videos

Existing HOI datasets [26, 78] are orders of magnitude smaller than full-body MoCap collections [69] due to severe self-occlusion and tightly coupled contact dynamics in egocentric views. To bridge this gap, we build a reconstruction pipeline (Fig. 3) that converts raw egocentric videos from EgoDex [24], which provides RGB streams, camera parameters, text descriptions, and tracked hand keypoints but no object annotations, into high-fidelity HOI training data used exclusively to pretrain our grasping prior (Sec. 4.3).

The pipeline consists of three steps. **(1) Transition detection:** we smooth wrist trajectories and identify the grasp-to-manipulation transition via local speed minima and direction changes. **(2) Object reconstruction:** we segment the target object with SAM3 [79], estimate metric depth with DepthAnything3 [80], and reconstruct a mesh with SAM3D [81] from pre-transition frames where the object is static. **(3) Hand-object alignment:** we fit MANO meshes via inverse kinematics, optimize an object translation offset to satisfy fingerpad contact and non-penetration constraints at the transition frame, and propagate the alignment to all frames. Full algorithmic details, loss formulations, hyperparameters, and evaluations of reconstruction results are provided in the supplementary material.

4.3 Two-stage hand-object interaction generation

Motivated by the human grasp-then-manipulate motor pattern [17], we decompose HOI into a *Grasping* and a *Manipulation* phase. While DiffH2O [18] also separates these 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—geometry/reachability vs. semantics/scene—rather than entangling all sources in a single denoiser.

Initial state estimation. From the first egocentric observation, we estimate the initial hand pose with [82], recover the object mesh and 6D pose with SAM3D [81], and lift both into a common world frame defined by the first camera's extrinsics, yielding $\mathbf{x}_{\text{init}} = (\mathbf{O}_0, \mathbf{H}_0^l, \mathbf{H}_0^r)$. We encode object geometry $\mathcal{M}$ with a Basis Point Set [83] and the action description T with a frozen T5-Large [84]; both are reused at every Euler step.

Hybrid 3D scene tokens. Manipulation is constrained by both coarse spatial layout and fine-grained, language-aligned object semantics, so we build complementary *global* and *local* tokens from a single 3DGS reconstruction of the scene [85, 86] (with moving hands and objects masked out). We sample N_s Gaussian centroids via Farthest Point Sampling [87], $\mathbf{X} \in \mathbb{R}^{N_s \times 3}$, and extract a geometric embedding via Concerto [88] and a language-aligned semantic embedding via SceneSplat [89], stacked into $\mathbf{E} \in \mathbb{R}^{N_s \times d_e}$ and $\mathbf{U} \in \mathbb{R}^{N_s \times d_u}$.

We fuse semantic and geometric cues with a linear-complexity gated mechanism adapted from squeeze-and-excitation [90]: after projecting both modalities into a shared d_h-dim space, $\tilde{\mathbf{E}} = \Phi_e(\mathbf{E})$ and $\tilde{\mathbf{U}} = \Phi_u(\mathbf{U})$, we combine them with a learnable channel-wise gate $\boldsymbol{\alpha} \in \mathbb{R}^{d_h}$:

$$\mathbf{F} = \tilde{\mathbf{E}} + \sigma(\boldsymbol{\alpha}) \odot \tilde{\mathbf{U}}, \tag{4}$$

allowing adaptive modulation of semantics and geometry while preserving linear complexity.

To explicitly encode spatial layout, we concatenate a Fourier positional encoding $\gamma(\mathbf{X}) \in \mathbb{R}^{N_s \times d_h}$ of the 3D point coordinates and apply a final projection to get the hybrid scene tokens:

$$\mathbf{P} = \Psi\big([\mathbf{F}, \gamma(\mathbf{X})]\big) \in \mathbb{R}^{N_s \times d_s}, \tag{5}$$

where d_s denotes the dimension of the scene token. Cross-attending each of the N per-frame motion tokens to all N_s scene tokens would cost $\mathcal{O}(N \cdot N_s)$, so we compress $\mathbf{P}$ into $L \ll N_s$ latents via a Perceiver bottleneck [91],

$$\mathbf{S}_{\text{local}} = \text{Perceiver}(\mathbf{P}; \mathbf{Q}_0) \in \mathbb{R}^{L \times d_s}, \tag{6}$$

with learnable queries $\mathbf{Q}_0 \in \mathbb{R}^{L \times d_s}$, reducing the per-layer cost to $\mathcal{O}(L \cdot N_s + N \cdot L)$.

For coarse layout, we voxelize the scene into an occupancy grid $\mathbf{V} \in \{0,1\}^{H_z \times H_x \times H_y}$, split it into non-overlapping $p \times p \times p$ cubic patches, prepend a learnable [CLS] token, and feed the sequence into a 3D vision transformer [92]; its [CLS] output is the global token $\mathbf{S}_{\text{global}} = \text{ViT}(\mathbf{V}) \in \mathbb{R}^{1 \times d_s}$. Hyperparameters ($H_x, H_y, H_z$, p, N_s, L) are in the supplementary. The two representations play complementary roles: $\mathbf{S}_{\text{global}}$ mainly supports scene-aware interaction placement and the selection of the interacting hand, while $\mathbf{S}_{\text{local}}$ refines fine-grained hand-object contact geometry tied to specific interaction regions (Supp. Sec. G.1).

Grasping stage. Since the object is static during grasping, we generate only the hand state $\mathbf{x}^g = (\mathbf{H}^l, \mathbf{H}^r) \in \mathbb{R}^{N_g \times D_h}$ (D_h: per-frame hand feature dim), conditioned on

$$\mathbf{c}^g = \{\, \text{BPS}(\mathcal{M});\ \text{T5}(T_g);\ \mathbf{x}_{\text{init}} \,\}, \tag{7}$$

where T_g is a grasp-focused sub-instruction extracted from T via an MLLM [93], isolating grasping conditioning from manipulation semantics, while $\mathbf{x}_{\text{init}}$ is concatenated to the noisy motion token at every sampling step to anchor generation. We train a network f_θ^g with x-prediction [75] to recover the clean target $\mathbf{x}_1^g$ from $\mathbf{x}_\tau^g$ along the linear path of Eq. (1),

$$\mathcal{L}_{\text{flow}}^g = \mathbb{E}_{\tau, \mathbf{x}_0, \mathbf{x}_1^g} \big\| f_\theta^g(\mathbf{x}_\tau^g, \tau, \mathbf{c}^g) - \mathbf{x}_1^g \big\|_2^2, \tag{8}$$

with the corresponding velocity field recovered through Eq. (2). To ground generation in language semantics, we add a TMR-inspired symmetric InfoNCE alignment loss [94, 95]

$$\mathcal{L}_{\text{align}} = \tfrac{1}{2}(\mathcal{L}_{\text{t2m}} + \mathcal{L}_{\text{m2t}}), \tag{9}$$

that aligns text and motion embeddings in a shared latent space. The total grasping loss is $\mathcal{L}_{\text{grasp}} = \mathcal{L}_{\text{flow}}^g + \lambda_{\text{align}} \mathcal{L}_{\text{align}}$ with $\lambda_{\text{align}} = 0.1$. At inference we Euler-integrate the recovered vector field from $\tau = 0$ to $\tau = 1$ to obtain the grasping trajectory terminating at t_g.

Manipulation stage. Unlike grasping, manipulation requires reasoning over the full sequence $\mathbf{x}^m \in \mathbb{R}^{N \times D}$ while preserving the established grasp and respecting scene constraints. We treat the previously generated grasping trajectory $\mathbf{x}_{[0,t_g]}^g$ as a known prefix and condition the manipulation generator on

$$\mathbf{c}^m = \{\, \text{BPS}(\mathcal{M});\ \text{T5}(T);\ \mathbf{S}_{\text{local}};\ \mathbf{S}_{\text{global}};\ \mathbf{x}_{t_g}^g \,\}. \tag{10}$$

With temporal mask $\mathbf{M} \in \{0,1\}^N$ (zero on $t \le t_g$, one elsewhere), at each Euler step with flow time $\tau < 0.9$ we softly inpaint the grasping prefix into the noisy trajectory,

$$\bar{\mathbf{x}}_\tau^m = \mathbf{M} \odot \mathbf{x}_\tau^m + (1 - \mathbf{M}) \odot \mathbf{x}_{[0,t_g]}^g, \tag{11}$$

where $\mathbf{x}_{[0,t_g]}^g$ is padded with the static initial object pose $\mathbf{O}_0$ to match the dimension of $\mathbf{x}^m$. Disabling the replacement for $\tau \ge 0.9$ lets the model refine the full sequence, and we clamp the transition via $\bar{\mathbf{x}}_\tau^m[t_g] \leftarrow \mathbf{x}_{t_g}^g$. Compared to DiffH2O's diffusion-step inpainting [18], our continuous-time formulation enforces transition consistency at the ODE level. Training uses a masked x_1-prediction loss applied only to the manipulation portion,

$$\mathcal{L}_{\text{flow}}^m = \mathbb{E}_{\tau, \mathbf{x}_0, \mathbf{x}_1^m} \big\| \mathbf{M} \odot \big(f_\theta^m(\bar{\mathbf{x}}_\tau^m, \tau, \mathbf{c}^m) - \mathbf{x}_1^m \big) \big\|_2^2, \tag{12}$$

and the total manipulation loss is $\mathcal{L}_{\text{manip}} = \mathcal{L}_{\text{flow}}^m + \lambda_{\text{align}} \mathcal{L}_{\text{align}}$ (Eq. 9).

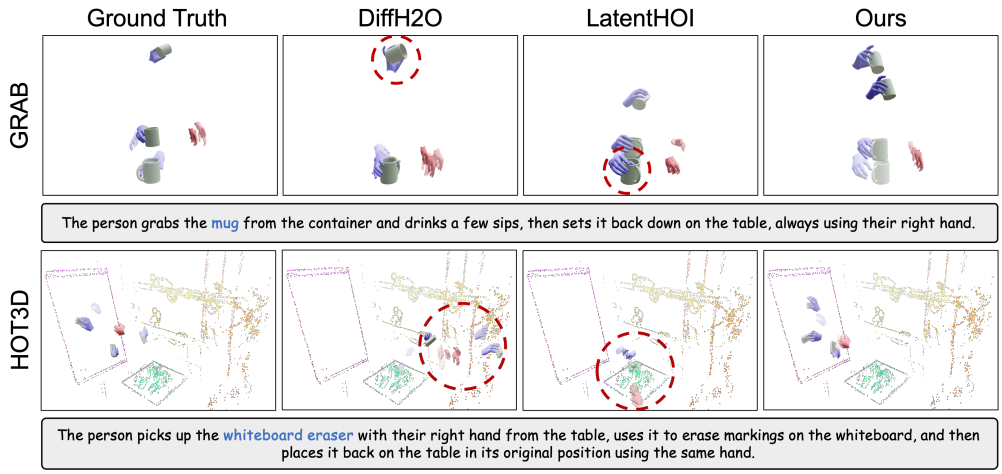

Figure 4: Qualitative comparison vs. DiffH2O [18] and LatentHOI [19] (top: GRAB, bottom: HOT3D with 3D scene). Ours yields more natural grasps and physically plausible manipulations that better follow the instruction. Best viewed in the supplementary video.

Table 1: Quantitative evaluation on GRAB and HOT3D. Best among generation methods in **bold**; Real Mocap is the upper bound. IV/ID are bolded only when paired with non-trivial CR (cf. LatentHOI on HOT3D: low IV/ID with 0% physics SR; Supp. Sec. D). Binomial SE on SR $\leq 1.1\%$ on GRAB and $\leq 0.4\%$ on HOT3D.

		Physical Interaction Quality				Motion Quality			Realizable Physical Feasibility			
Dataset	Method	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**

5 Experiments

5.1 Datasets

We evaluate our method on two datasets, namely HOT3D [29] and GRAB [26]. We pretrain our grasping model with our curated EgoDex [24] dataset as described in Sec. 4.2. EgoDex is used solely for training the grasping prior model and excluded from our evaluations. GRAB [26] provides high-fidelity hand-object motion capture and is used for quantitative evaluation without scene context. HOT3D [29] offers real-world egocentric recordings with accurate hand and object pose annotations together with the source video required by our 3DGS scene reconstruction, enabling scene-conditioned evaluation and generalization to unseen objects. HOT3D's annotations are noisier than GRAB's mocap supervision: even Real Mocap reaches only 6.67% physics-simulation SR (Table 1), so we treat CR and AR as the more informative signals on HOT3D (analysis in Supp. Sec. D). Note that neither dataset provides natural language action descriptions or pre-built 3D scene reconstructions. Details of our reconstruction and annotation procedures are provided in the supplementary material.

5.2 Implementation details

We implement FlowHOI in PyTorch [96] and train on an RTX 6000 Ada GPU (48GB) with batch size 64, using conditional flow matching [73] with optimal-transport paths, x-prediction, and 50-step Euler integration at inference. For DiffH2O [18] and LatentHOI [19], we run the official implementations and retrain on GRAB and HOT3D under the same data splits and evaluation protocol. For fair comparison we replace the CLIP [97] text encoder with the more capable T5 [84] encoder for all baselines. Additional details are in the supplementary material.

5.3 Evaluation metrics

Aligned with prior work [18, 19, 39, 98], we report three metric groups (formal definitions in Supp. Sec. H). *Physical interaction quality* comprises interpenetration volume (IV), depth (ID), contact ratio (CR), and IV per contact unit (IVU). *Motion quality* comprises action-recognition accuracy (AR) and sample/overall diversity (SD/OD). *Realizable physical feasibility* comprises two complementary measures. The heuristic physical plausibility score (Phy) of [19] is a kinematic-only check (sustained contact and object above ground) and does not enforce dynamics. For physics-grounded evaluation, we additionally retarget each generated HOI to the Allegro Hand via IK [99] and execute it in Isaac Gym [100] with a physics-based tracker [101], reporting the simulation success rate (SR), holding time (HT), and the per-sequence HOI generation time. We treat SR as the primary holistic indicator, since low IV/IVU can be trivially achieved by avoiding contact.

5.4 HOI generation comparison

We compare FlowHOI with DiffH2O [18] and LatentHOI [19] on GRAB and HOT3D (Table 1), focusing on long-horizon contact consistency, semantic alignment, and inference efficiency.

Contact consistency over long horizons. On GRAB, FlowHOI achieves the lowest IV and IVU, indicating that contact geometry remains consistent over extended interactions with less error accumulation. On HOT3D, FlowHOI attains the highest contact ratio while remaining competitive with DiffH2O on penetration metrics. LatentHOI achieves lower IV/ID at the cost of substantially weaker contact and a 0% physics-simulation SR (Table 1); a detailed analysis of its reproduction is in Supp. Sec. D.

Semantically grounded generation without sacrificing diversity. FlowHOI achieves the highest AR, reflecting strong language and scene grounding from the manipulation stage's 3D scene tokens and motion-text alignment loss. Sample-level and overall diversity remain comparable to prior methods, with the highest OD on GRAB.

Realizable physical feasibility and efficiency. In physics simulation, FlowHOI achieves the best SR on GRAB (55.96% vs. 28.44% for LatentHOI and 33.03% for DiffH2O) and the longest execution duration of 1.5 s, confirming that the generated trajectories remain stable and executable after retargeting. Although LatentHOI achieves the highest heuristic Phy score on GRAB, this metric does not capture physical stability under dynamics, as reflected by its substantially lower SR. By adopting flow matching, the planning step requires only 0.16 s per sequence, a $40\times$ reduction relative to the diffusion-step ODE of DiffH2O (6.34 s) and LatentHOI (3.57 s). Qualitative results in Fig. 4 further corroborate these findings, showing more realistic grasp configurations and smoother manipulation trajectories, particularly for long-horizon interactions.

Input-matched comparison. No method uses scene context on GRAB, so the comparison there is input-matched. On HOT3D, FlowHOI additionally receives 3D scene tokens that the baselines do not use. Without scene conditioning, FlowHOI reaches an AR of 0.73 on HOT3D (Table 3a), compared with 0.71 for DiffH2O and 0.65 for LatentHOI. The full HOT3D results in Table 1 thus combine the generator with scene conditioning, whose contribution Table 3a and Supp. Sec. G.1 isolate.

Post-grasp motion. After the transition frame, the manipulation stage continues to predict finger articulation and contact states, so the object can rotate relative to the hand instead of being carried rigidly. On 174 GRAB hand instances with an asymmetric object and more than 30° of ground-truth hand-relative reorientation, FlowHOI reduces the hand-relative rotation error from 83.0° for DiffH2O to 46.4° (Supp. Sec. K).

5.5 Showcase of real-world applications

We further evaluate the physical feasibility of executing generated HOI sequences on a real-world dexterous manipulation platform, consisting of two Franka Emika Panda robotic arms [102], each equipped with an Allegro Hand v5 [103]. We consider four contact-rich household manipulation tasks: drinking from a cup, pouring liquid between containers of different sizes, tilting a container, and squeezing dressing. In this setup FlowHOI plays the role of an interaction *planner* rather than a closed-loop controller: it produces a single reference trajectory from one observation, which is then refined offline and executed open-loop. The pipeline takes egocentric RGB and a Gaussian-LIC [104] 3DGS scene as input (Fig. 1). The pyroki solver [99] converts the robot proprioceptive

Table 2: 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); 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%**

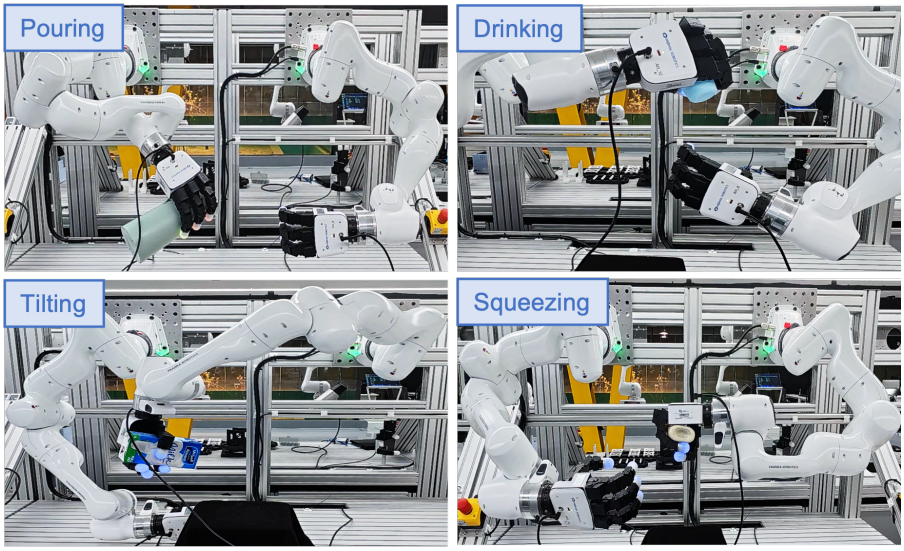

Figure 5: Real-world execution on four contact-rich tasks (pouring, drinking, tilting, squeezing): generated HOI retargeted to a Franka Panda + Allegro Hand and run open-loop.

state into the initial MANO pose required by FlowHOI; after generation, the same solver retargets the predicted MANO trajectory to 16-DOF Allegro joint actions, which are refined by DexTrack [101] and executed open-loop via a joint impedance controller, producing stable contact-rich interactions (Fig. 5). Quantitatively (Table 2), FlowHOI achieves a 57.5% average SR, $5.75\times$ DiffH2O's 10%. Residual failures stem primarily from embodiment mismatch (e.g., Allegro fingertips not fitting the mug handle for pouring) and post-grasp contact slippage under open-loop tracking. Table 2 therefore measures executability under nominal conditions, not robustness to state-estimation errors or disturbances.

5.6 Ablation study

Two-stage decomposition. We compare with a single-stage FlowHOI that generates the full sequence in one pass and is matched in representation, training data, objective, and parameter budget (Supp. Sec. J). The two-stage design raises the contact ratio and Phy and lowers interpenetration on both GRAB and HOT3D. AR and simulation SR are comparable, and the single-stage model is 40 ms faster. The benefit of the decomposition thus lies in contact establishment rather than overall accuracy, and it also enables stage-specific conditioning and grasp-only pretraining.

Effect of Pretraining Grasping Model. Table 3b compares grasping models with and without pretraining on large-scale egocentric data (Sec. 4.2). We report grasp error (GE), defined as the distance between the generated hand pose and the ground-truth grasp at the end of the grasping stage. Pretraining substantially reduces GE ($0.10\,\text{m} \rightarrow 0.06\,\text{m}$, 40% relative reduction), indicating more accurate grasp initialization. Penetration-related metrics remain comparable, indicating the prior improves grasp accuracy without regressing the contact behavior already enforced by our contact-aware data representation. With the manipulation generator held fixed and only the grasp prefix varied (Supp. Sec. C.5), pretraining also lowers the wrist velocity across the stage boundary from 0.583 to 0.340 m/s and raises handedness accuracy from 96.34% to 99.02%, while Phy and simulation SR show no clear difference at our evaluation scale.

Table 3: Ablation studies. (a) Scene encoder choices on HOT3D; 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++ [105]	0.59	0.44
Concerto [88]	0.75	0.26
SceneSplat [89]	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 [97]	0.89
CLIP + Align. Loss	0.88
T5 [84]	0.92
T5 + Align. Loss	**0.94**

Semantic grounding via text and scene conditioning. Results are summarized in Table 3. For text encoding, the T5 encoder consistently outperforms CLIP-based variants in action recognition accuracy. Incorporating the motion-text alignment loss (Eq. (9)) further improves the performance, indicating stronger correspondence between generated motions and language instructions. For scene encoding on HOT3D, removing scene information leads to degraded action recognition accuracy and large final displacement error (FDE). Geometry-only or semantics-only scene representations provide partial improvements, while our fused representation achieves the best performance on both metrics. This demonstrates that jointly modeling spatial geometric constraints and scene semantics is critical for accurate object motion and semantically consistent manipulation.

6 Conclusion

We presented FlowHOI, a two-stage flow-matching framework that generates semantically grounded HOI sequences conditioned on an egocentric observation, a language instruction, and a 3DGS scene reconstruction. By decoupling geometry-centric grasping from semantics-centric manipulation with 3D scene tokens and a motion-text alignment loss, FlowHOI grounds interactions in both the physical scene and the language instruction. A reconstruction pipeline recovering HOI trajectories from egocentric videos further provides a grasping prior that improves grasp initialization. On GRAB and HOT3D, FlowHOI achieves the highest action-recognition accuracy, a $1.7\times$ higher physics-simulation success rate and a $40\times$ inference speedup over DiffH2O [18] (the strongest diffusion-based baseline in our evaluation), and up to $\sim 21\%$ lower interpenetration on GRAB compared to LatentHOI [19]. Real-robot experiments on four contact-rich tasks show that the generated interactions can be retargeted to a dexterous robot hand and executed open-loop.

Limitations and future work. FlowHOI targets contact-rich articulated-hand HOI generation and retargeting under nominal conditions, and within this scope it has three limitations. (i) It relies on initial-state estimation from a single observation. (ii) It generates purely kinematic trajectories, which on the robot depend on offline refinement by a physics-based tracker [101] and are executed open-loop, so FlowHOI cannot react online to perturbations or contact slippage. (iii) Each sequence contains a single grasp-to-manipulation transition. The generated post-grasp motion includes hand-relative reorientation, but we do not evaluate repeated release and regrasp, force-level contact modes, deformable objects, or general in-hand manipulation, and multi-phase tasks such as opening a lid are not covered. Future work includes closed-loop state estimation, replanning, and contact recovery, a phase planner that chains the grasping and manipulation generators across multi-segment tasks, scaling to mobile manipulation by jointly modeling base motion and reachability with HOI generation, and learning priors from large-scale exocentric videos.

Acknowledgments and Disclosure of Funding

This work is partially supported by "The MBZUAI-WIS Joint Program for Artificial Intelligence Research". We thank the Agentic Robotics Lab at MBZUAI for their hardware support.

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, Rolandos Alexandros Potamias, Taku Komura, 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] Huajian Zeng, Abhishek Saroha, Daniel Cremers, and Xi Wang. Gmt: Goal-conditioned multimodal transformer for 6-dof object trajectory synthesis in 3d scenes. In *2026 International Conference on 3D Vision (3DV)*, pages 1691–1701. IEEE, 2026.

[44] Abhishek Saroha, Huajian Zeng, Xingxing Zuo, Daniel Cremers, and Xi Wang. Egoflow: Gradient-guided flow matching for egocentric 6dof object motion generation. *arXiv preprint arXiv:2604.01421*, 2026.

[45] 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.

[46] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

[55] Raktim Gautam Goswami, Amir Bar, David Fan, Tsung-Yen Yang, Gaoyue Zhou, Prashanth Krishnamurthy, Michael Rabbat, Farshad Khorrami, and Yann LeCun. World models for learning dexterous hand-object interactions from human videos. *arXiv preprint arXiv:2512.13644*, 2025.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1805–1815, 2025.

[72] Huajian Zeng, Chaohua Yao, Yuantai Zhang, Jiaqi Yang, Rolandos Alexandros Potamias, and Xingxing Zuo. Stablehand: Quality-aware flow matching for world-space dual-hand motion estimation from egocentric video. *arXiv preprint arXiv:2605.18553*, 2026.

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

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

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

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

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

[96] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[97] 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.

[98] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023.

[99] 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.

[100] 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.

[101] 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.

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

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

[104] 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.

[105] 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.

[106] 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.

[107] 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.

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

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

[110] 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.

[111] 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.

[112] 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.

[113] 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 multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

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

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

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. B. We then describe the hand-object data reconstruction pipeline, together with the effect of EgoDex pretraining, and dataset preprocessing in Sec. C and Sec. D, respectively. Next, we elaborate on the physics simulation evaluation protocol in Sec. E and provide details on the real-world experiment setup in Sec. F. We then analyze the scene representation, the granularity of scene conditioning, and hand-scene penetration under clutter in Sec. G. Full evaluation metric definitions are provided in Sec. H. We additionally provide a stage-transition soft-inpainting ablation in Sec. I, a comparison of single-stage and two-stage generation in Sec. J, a post-grasp dexterity analysis in Sec. K, additional qualitative results in Sec. L, and failure cases in Sec. M. We also provide a supplementary video with animated results.

A Implementation details

This section provides additional implementation details beyond Sec. 5.2 of the main paper.

A.1 Flow matching: deferred definitions

We collect here the flow matching definitions deferred from Sec. 3 of the main paper.

Neural ODE formulation. The continuous-time transport from p_0 to p_1 underlying conditional flow matching is defined by the neural ODE [74]:

$$\frac{d}{d\tau}\,\boldsymbol{\phi}_\tau(\mathbf{x}) = \mathbf{v}(\boldsymbol{\phi}_\tau(\mathbf{x}), \tau)\,, \qquad \boldsymbol{\phi}_0(\mathbf{x}) = \mathbf{x}, \tag{13}$$

where $\boldsymbol{\phi}_\tau$ is the flow map. The optimal-transport path of Eq. 1 is one specific instantiation of this ODE.

Role of $\sigma_{\min}$. The constant $\sigma_{\min} > 0$ in Eq. 1 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. 2 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. 13 to $\tau = 1$ with a numerical ODE solver. We use Euler discretization with $K = 50$ steps:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\tau\; \mathbf{v}_\theta(\mathbf{x}_k, \tau_k, \mathbf{c}), \qquad \Delta\tau = \tfrac{1}{K},\;\; \tau_k = \tfrac{k}{K}. \tag{14}$$

A.2 Architecture and training

Both models use a DiT [106] 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 [84], projected to d and injected via cross-attention. Object geometry uses BPS [83] 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. 4.3) 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) [87]. Geometric features from Concerto [88] d_e=1536 and semantic features from SceneSplat [89] d_u=768 are fused via bidirectional cross-attention with d_h=512. 3D coordinates use Fourier positional encoding [107], projected to 64-dim. Fused features are compressed by a Perceiver [91] bottleneck into K=256 scene tokens of dimension d. For global scene encoding, we voxelize the scene at $48\times48\times24$ resolution and process it with a ViT, injected via AdaLN [106].

We use AdamW [108] 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.

Table 4: 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. H).

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**

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 [94] 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 [84] embeddings and projects them to the same 512-dim space via a two-layer MLP. The temperature τ is initialized to 0.7, learnable, and clamped to $[0.01, 100]$. The alignment loss is weighted by $\lambda_{\text{align}}{=}0.1$ 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 [97] $\rightarrow$ T5-Large [84] 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 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 [75] 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 [75]. 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 GRAB [26] dataset. We evaluate temporal smoothness using two complementary metrics (see Sec. H 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. 4, 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. 6 further visualizes the generated trajectories, where v-prediction exhibits noticeable high-frequency jitter whereas x-prediction produces temporally smooth results.

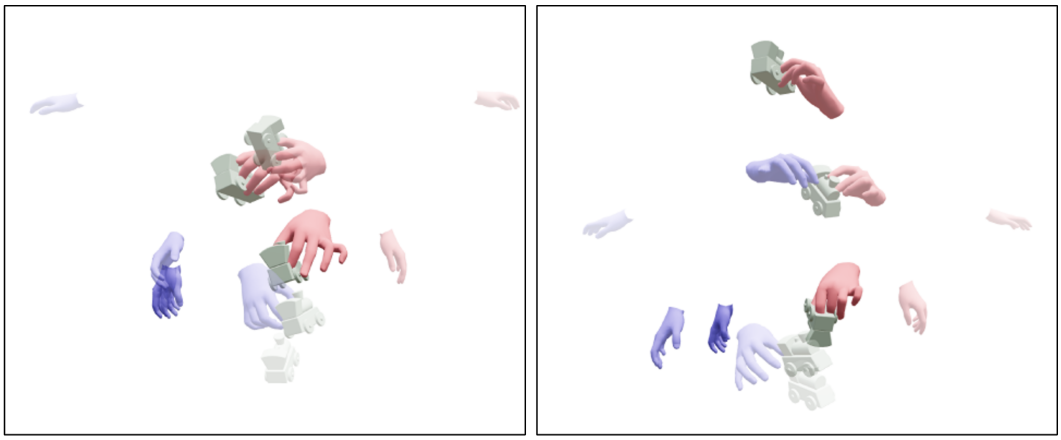

Figure 6: Qualitative comparison of prediction targets. Our x-prediction produces temporally smooth and stable hand trajectories, while v-prediction exhibits noticeable spatial jitter and inconsistent grasping poses across frames.

C 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. This section provides the full algorithmic details, loss formulations, and hyperparameters of the reconstruction pipeline summarized in Sec. 4.2 of the main paper. An overview of the pipeline is illustrated in Fig. 7.

C.1 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 [109] (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.

C.2 Step 2: 3D object reconstruction

We use an LLM (GPT-4o [93]) 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 [79] 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 [81] reconstructs geometry from a single RGB image but does not recover metric scale, we first estimate metric depth for the initial frame using DepthAnything3 [80] 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.3 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

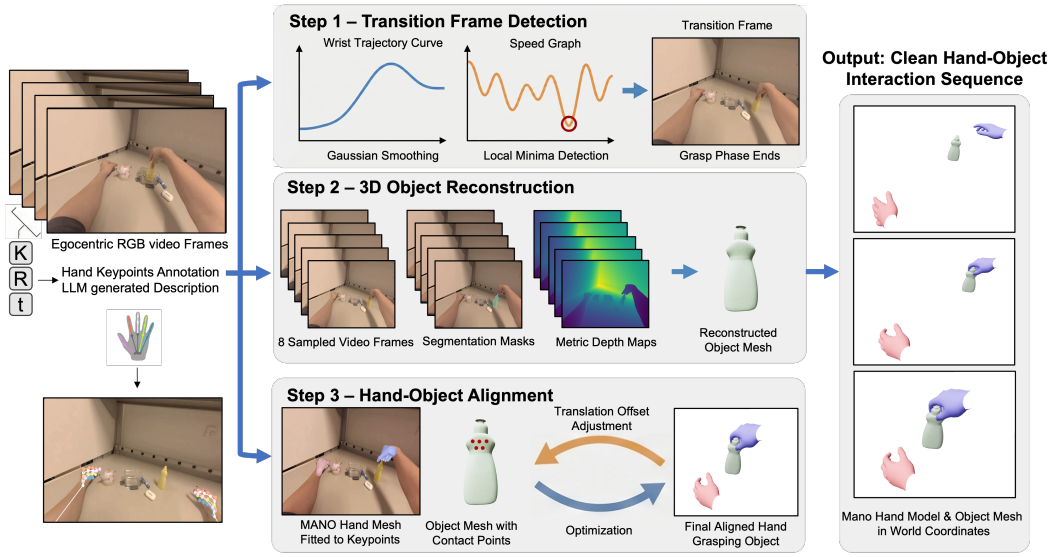

Figure 7: **Detailed hand-object data reconstruction pipeline.** Given an egocentric RGB video with tracked hand keypoints and camera parameters, we reconstruct HOI data in three steps. (1) *Transition frame detection*: we smooth wrist trajectories and identify the grasp-to-manipulation transition via local speed minima and direction change. (2) *3D object reconstruction*: we segment the target object with SAM3 [79], estimate metric depth with DepthAnything3 [80], and reconstruct a mesh with SAM3D [81] 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.

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) [99], 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:

$$\min_{\Delta \boldsymbol{p}^o, \Delta\boldsymbol{\theta}} \; \mathcal{L}_{\text{dist}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{15}$$

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.

C.4 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 5 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.

Table 5: 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

Table 6: Effect of EgoDex pretraining. Only the grasp prefix differs. RG is measured on the real robot in a separate set of 10 squeezing trials, and all other metrics on GRAB.

Grasping prior	GE $[m]$ (↓)	TV $[m/s]$ (↓)	HA [%] (↑)	Phy [%] (↑)	SR [%] (↑)	RG (of 10) (↑)
No EgoDex pretraining	0.10	0.583	96.34	**91.06**	51.38	6
EgoDex-pretrained (ours)	**0.06**	**0.340**	**99.02**	90.90	**55.96**	6

C.5 Effect of EgoDex pretraining

We fix the manipulation generator and vary only whether the grasp prefix comes from an EgoDex-pretrained model. Besides GE, Phy, and SR on GRAB, we report the wrist velocity across the stage boundary (TV), the handedness accuracy (HA), which is the fraction of clips whose first manipulation frame engages the ground-truth hand, and the real-robot grasp success (RG) over 10 separate squeezing trials. Pretraining lowers GE and TV and raises HA (Table 6). Phy, SR, and RG show no clear difference, and all models are trained with a single seed.

D Dataset preprocessing details

This section details our preprocessing and annotation procedures for each dataset used in Sec. 5.1 of the main paper.

D.1 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.

D.2 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 [78, 26], (ii) high-fidelity real-world scene capture comparable to large-scale indoor datasets [110, 111], 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 [112, 86] from the recorded egocentric videos captured by the Project Aria glasses [113]. 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 [93] 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 1), 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 1), 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→T5 swap applied uniformly to all baselines for fairness is not the source of the HOT3D failure.

E Physics simulation analysis

This section provides additional details on the physics simulation experiments described in Sec. 5.3 of the main paper.

We evaluate the realizable physical feasibility of generated HOI sequences in Isaac Gym [100]. The evaluation pipeline consists of three steps: (1) retargeting the generated MANO [77] hand trajectories to the Allegro Hand [103] via inverse kinematics [99], (2) executing the retargeted motions using a physics-based tracking controller [101] 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× 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 [99], 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 [101], 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 [100] 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

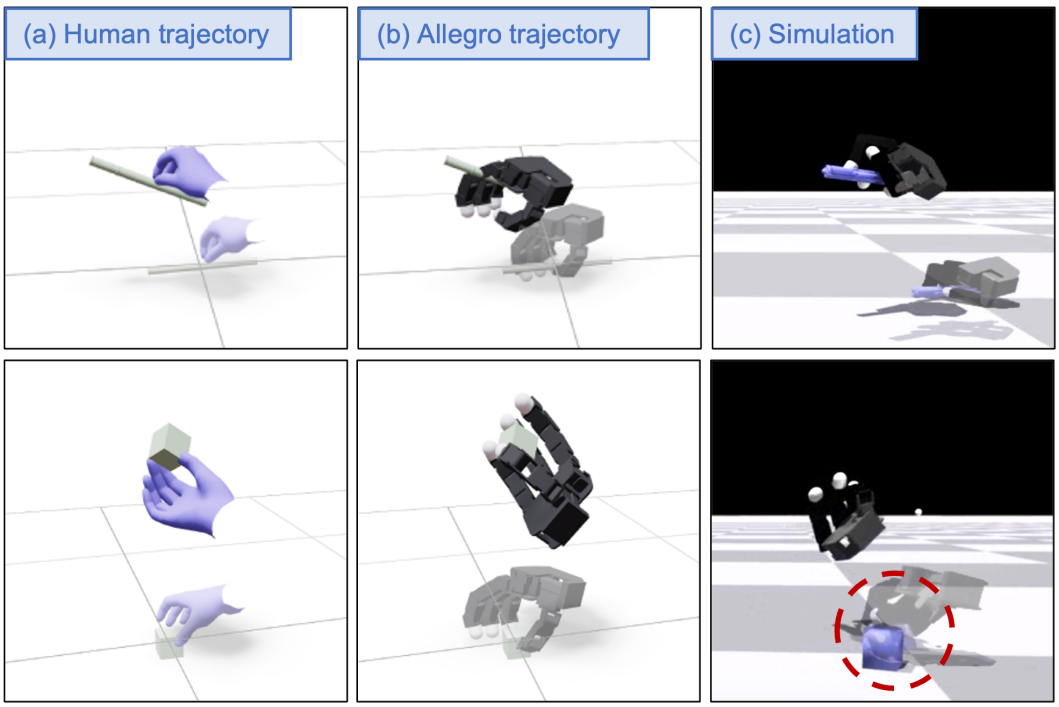

Figure 8: Comparison of HOI trajectories: (a) kinematic human hand trajectory, (b) retargeted Allegro Hand trajectory, and (c) physics simulation in Isaac Gym [100]. Although the kinematic trajectories in (a) and (b) appear plausible, executing them under physics simulation reveals failures: the object detaches and slips from the grasp (bottom row, red circle), indicating that kinematically feasible motions do not guarantee physically stable interactions.

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. 8, 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 [101].

F 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 [102], each equipped with an Allegro Hand v5 [103]. The 3D scene is reconstructed offline prior to each experiment: we capture a multi-view video sequence using a MetaCam [114] scanner and reconstruct the scene via Gaussian-LIC [104], yielding a 3DGS representation consistent with our training pipeline.

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 [99], following the same procedure described in Sec. E. The retargeted trajectories are further refined offline by the DexTrack [101] sim-to-real policy, which was trained entirely in Isaac Gym [100] 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, with 10 trials per task (Table 2). We refer the reader to Fig. 5 in the main paper and the supplementary video for visual 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 constraints. Drinking is a compound lift-tilt-return trajectory. Squeezing requires coordinated finger flexion with distinct joint targets and force profiles.

Phase decomposition of the tasks. Each task uses a single grasp-to-manipulation transition. In squeezing, the finger closure that establishes contact belongs to the grasping stage, and the compression belongs to the manipulation stage. Multi-phase tasks such as opening a bottle lid and deformable objects such as a toothpaste tube are outside the scope of this work.

G Scene representation details and analysis

This section provides additional details on the scene representation described in Sec. 4.3 of the main paper.

Fig. 9 visualizes the geometric features (Concerto [88]) and semantic features (SceneSplat [89]) 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.

G.1 Scene-conditioning granularity

On HOT3D, we compare no scene conditioning, global-only, local-only, and global-plus-local conditioning, with everything else fixed. In addition to GE and HA (Sec. C.5), we report the hand-scene penetration depth (PD, Sec. G.2). The global token mainly improves interaction placement and hand selection (GE, HA), and the local tokens mainly improve hand-object contact (IV, CR). Combining both gives the best GE, HA, Phy, and PD, although PD differs only slightly across variants (Table 7).

G.2 Hand-scene penetration under scene clutter

We split the HOT3D test sequences into three equal groups by a clutter score, the number of occupied scene voxels (8.3 cm) within 0.5 m of the ground-truth object trajectory, excluding the target object. Penetration is measured from MANO vertices to the MPS semidense point cloud released with HOT3D, which contains only the static scene. We report the mean penetration depth over penetrating vertex-frames and a GT-tolerance success rate, the fraction of clips whose maximum per-frame

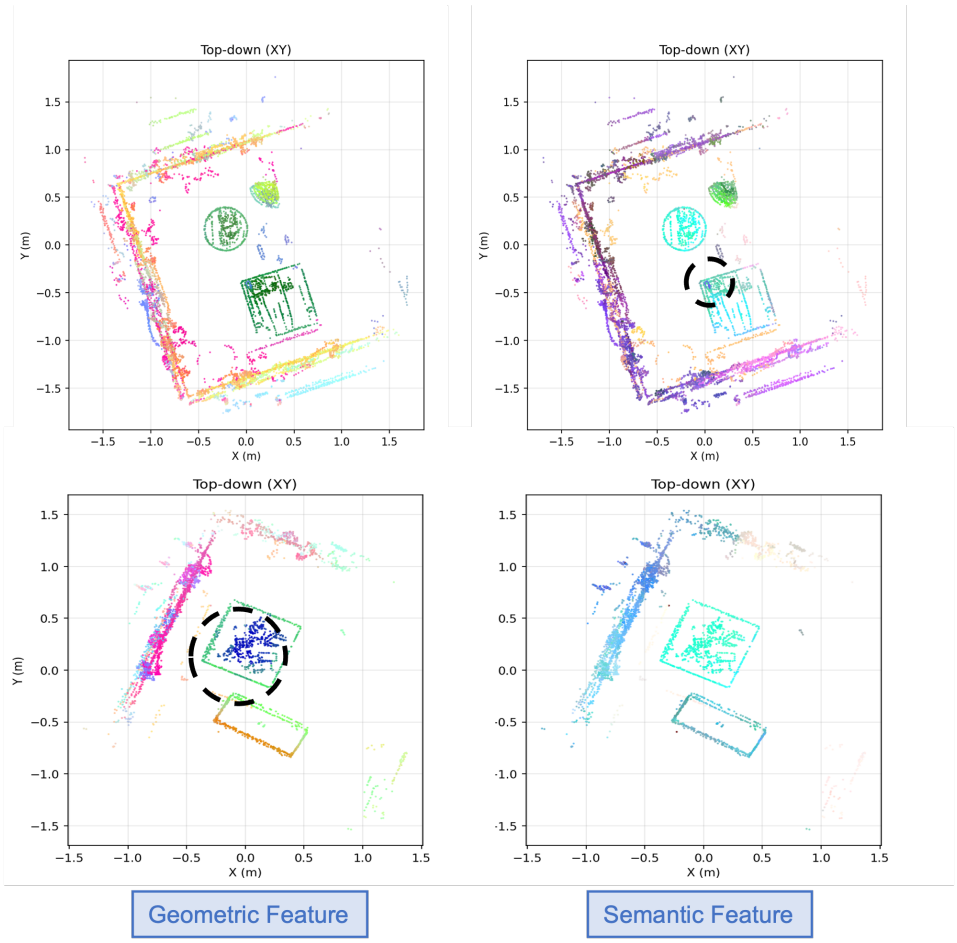

Figure 9: **Scene feature analysis.** Top-down views of two scenes comparing geometric features (Concerto [88], left) and semantic features (SceneSplat [89], right), colored by PCA of the respective feature spaces. Black dashed circles highlight where geometric features successfully distinguish objects from the supporting surface (bottom-left) and where semantic features successfully discriminate semantically distinct regions (top-right). This complementarity motivates our bidirectional fusion.

Table 7: Scene-conditioning granularity on HOT3D. GE: grasp error at the transition frame. HA: handedness accuracy. PD: hand-scene penetration depth.

Scene conditioning	IV $[cm^3]$ (↓)	CR [%] (↑)	Phy [%] (↑)	GE $[m]$ (↓)	HA (↑)	PD $[mm]$ (↓)
None	3.26	2.27	16.47	0.20	0.83	2.08
Global only	3.30	2.09	16.63	0.14	0.88	1.97
Local only	**3.22**	**2.33**	16.10	0.19	0.85	1.98
Global + local (ours)	3.36	2.14	**17.15**	**0.11**	**0.92**	**1.95**

penetration stays below 6.66 mm (the 95th percentile of the ground truth). The ground truth is evaluated with the same pipeline as a reference. FlowHOI stays within 0.03 mm of the ground-truth penetration depth in every group, and its success-rate gap to the ground truth shrinks from 6.7 to 0.9 points as clutter increases (Table 8).

H Evaluation metric details

This section provides detailed definitions and implementation details of the evaluation metrics reported in Sec. 5.3 of the main paper.

Table 8: Hand-scene penetration against the HOT3D MPS static point cloud, stratified by scene clutter.

Metric	Source	Low (n=120)	Medium (n=120)	High (n=120)
Penetration depth [mm] (↓)	FlowHOI	1.97	1.94	2.04
	Ground truth	1.94	1.94	2.02
GT-tolerance success rate [%] (↑)	FlowHOI	70.0	80.8	85.8
	Ground truth	76.7	85.0	86.7

H.1 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:

$$\mathrm{IV}_t = \mathrm{Vol}\big(\mathcal{M}_h^t \cap \mathcal{M}_o^t\big)\,, \tag{16}$$

and is averaged over time:

$$\mathrm{IV} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{IV}_t. \tag{17}$$

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

$$\mathrm{ID}_t = \frac{1}{|\mathcal{V}_{\mathrm{pen}}^t|} \sum_{\mathbf{v}\in\mathcal{V}_{\mathrm{pen}}^t} \max\big(0, -d(\mathbf{v}, \mathcal{M}_o^t)\big)\,, \tag{18}$$

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:

$$\mathrm{CR}_t = \frac{1}{|\mathcal{V}_h|} \sum_{\mathbf{v}\in\mathcal{V}_h} \mathbb{I}\big(d(\mathbf{v}, \mathcal{M}_o^t) \le \delta\big)\,, \tag{19}$$

where δ is a fixed distance threshold (set to 5 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:

$$\mathrm{IVU}_t = \frac{\mathrm{IV}_t}{A_{\mathrm{contact}}^t}, \quad A_{\mathrm{contact}}^t = \mathrm{CR}_t \cdot \mathrm{Area}(\mathcal{M}_h), \tag{20}$$

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

$$\mathrm{IVU} = \frac{1}{|\mathcal{T}_c|} \sum_{t\in\mathcal{T}_c} \mathrm{IVU}_t, \tag{21}$$

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

H.2 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:

$$\mathrm{AR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\arg\max f_{\mathrm{act}}(\hat{\mathbf{x}}_i) = y_i)\,. \tag{22}$$

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:

$$\mathrm{SD} = \frac{2}{K(K-1)} \sum_{1\le i<j\le K} \frac{1}{T'} \left\| \mathbf{h}^{(i)} - \mathbf{h}^{(j)} \right\|_2, \tag{23}$$

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:

$$\mathrm{OD} = \frac{2}{S(S-1)} \sum_{1\le i<j\le S} \frac{1}{T'} \left\| \mathbf{h}_i - \mathbf{h}_j \right\|_2, \tag{24}$$

where S is the total number of generated motions.

H.3 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 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. E, 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 cm while maintaining contact with the hand. The success rate is the fraction of such sequences:

$$\mathrm{SR} = \frac{1}{N}\sum_{n=1}^{N} s_n, \tag{25}$$

where $s_n \in \{0, 1\}$ indicates whether the object in the n-th trial is lifted off the ground (> 5 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:

$$\mathrm{HT} = \frac{\sum_{n=1}^{N} s_n\, \tau_n}{\sum_{n=1}^{N} s_n}, \tag{26}$$

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

H.4 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:

$$\mathrm{GE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \frac{1}{|\mathcal{J}^{(i)}|} \sum_{j\in\mathcal{J}^{(i)}} \left\| \hat{\mathbf{j}}^{(i)}_{t_g,j} - \mathbf{j}^{(i)}_{t_g,j} \right\|_2, \tag{27}$$

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.

Table 9: Effect of stage-transition soft inpainting on GRAB. Inpainting suppresses high-frequency artifacts at the grasp/manipulation boundary by an order of magnitude.

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

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:

$$\mathrm{FDE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left\| \hat{\mathbf{O}}_{N-1}^{(i)} - \mathbf{O}_{N-1}^{(i)} \right\|_2, \tag{28}$$

where $\hat{\mathbf{O}}_{N-1}^{(i)}$ and $\mathbf{O}_{N-1}^{(i)}$ 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 follows a *minimum jerk* principle [115], 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:

$$\mathrm{Jerk}_{\mathrm{pos}} = \frac{1}{T} \sum_{t=1}^{T} \left\| \frac{d^3 \mathbf{p}^t}{dt^3} \right\|_2, \tag{29}$$

where derivatives are approximated using central differences (`np.gradient`). We analogously define the angular jerk $\mathrm{Jerk}_{\mathrm{ang}}$ 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:

$$\mathrm{Acc}_g^{\mathrm{pos}} = \frac{1}{T} \sum_t \left| \frac{d^2 \mathbf{p}_w^t}{dt^2} \right|, \quad \mathrm{Acc}_l^{\mathrm{pos}} = \frac{1}{T} \sum_t \left| \frac{d^2 \tilde{\mathbf{p}}_f^t}{dt^2} \right|, \tag{30}$$

where $|\cdot|$ denotes element-wise absolute value and the result is averaged over all spatial components and both hands. The rotational counterparts $\mathrm{Acc}_g^{\mathrm{rot}}$ and $\mathrm{Acc}_l^{\mathrm{rot}}$ 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.

I Stage-transition inpainting ablation

We ablate the soft inpainting mechanism (Sec. 4.3, 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 9 reports motion-smoothness metrics: positional jerk ($\mathrm{Jerk}_{\mathrm{pos}}$), angular jerk ($\mathrm{Jerk}_{\mathrm{ang}}$), global wrist acceleration ($\mathrm{Acc}_g^{\mathrm{pos}}$), and grasp error (GE).

J Single-stage versus two-stage generation

The single-stage variant generates the same 200-frame sequence in one pass and matches the two-stage model in training data, objective, inference steps, parameter budget, and EgoDex pretraining. The two-stage model improves the contact ratio (26% higher on HOT3D), Phy, and interpenetration volume on both datasets, while AR and SR are comparable and the single-stage model is 40 ms faster (Table 10).

Table 10: Single-stage versus two-stage generation. Both variants share the representation, training data, objective, inference steps, parameter budget, and EgoDex pretraining.

Dataset	Model	IV [cm^3] (↓)	CR [%] (↑)	Phy [%] (↑)	AR (↑)	SR [%] (↑)	Time [s] (↓)
GRAB	Single-stage	11.94	6.75	89.83	0.93	54.12	**0.12**
	Two-stage (ours)	**10.93**	**6.85**	**90.90**	**0.95**	**55.96**	0.16
HOT3D	Single-stage	3.53	1.70	16.42	**0.80**	**6.67**	**0.12**
	Two-stage (ours)	**3.36**	**2.14**	**17.15**	0.78	5.33	0.16

Table 11: Post-grasp dexterity on the 174-instance GRAB reorientation subset.

Method	RRE [deg] (↓)	CT-F1 (↑)
DiffH2O [18]	83.0	0.051
FlowHOI (ours)	**46.4**	**0.079**

K Post-grasp dexterity analysis

We evaluate the 174 of 410 GRAB hand instances with an asymmetric object and more than 30° of ground-truth hand-relative rotation. RRE is the geodesic error of the hand-relative object orientation over post-grasp contact frames, and CT-F1 scores joint-level contact make and release events. FlowHOI reduces RRE by 44% while producing $1.11\times$ the ground-truth rotation range, and it over-segments contact less than DiffH2O ($2.5\times$ vs. $4.5\times$ the ground-truth number of contact events, Table 11).

L Additional qualitative results

We provide additional qualitative results to complement the main paper. Fig. 10 provides qualitative comparisons with DiffH2O [18] and LatentHOI [19] on GRAB [26]. Fig. 11 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.**

M Failure cases

We identify two representative failure modes of our framework (see Fig. 12):

(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.

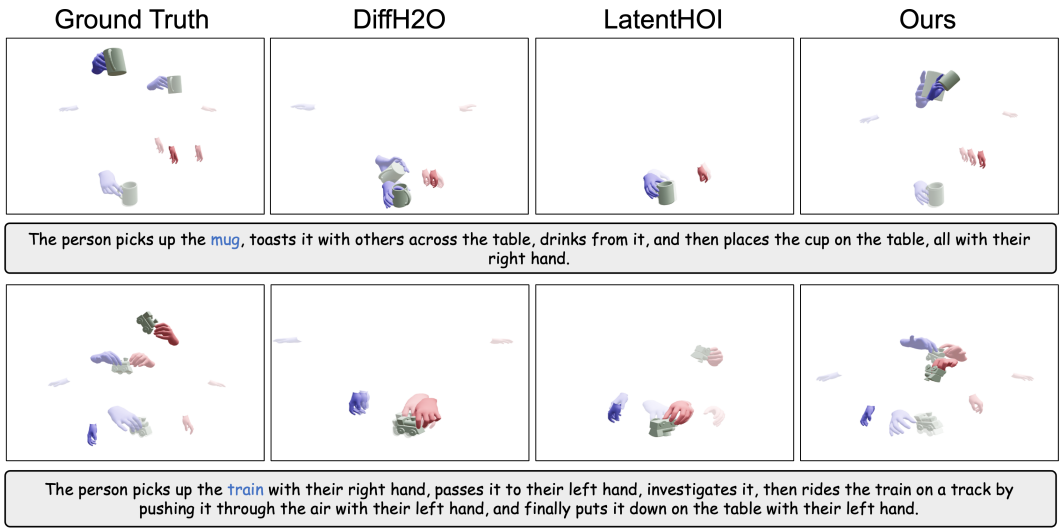

Figure 10: **Qualitative comparison on GRAB.** Comparison with DiffH2O [18] and LatentHOI [19] on diverse actions. Our method produces more physically plausible hand-object interactions with accurate grasping poses and smooth bimanual coordination.

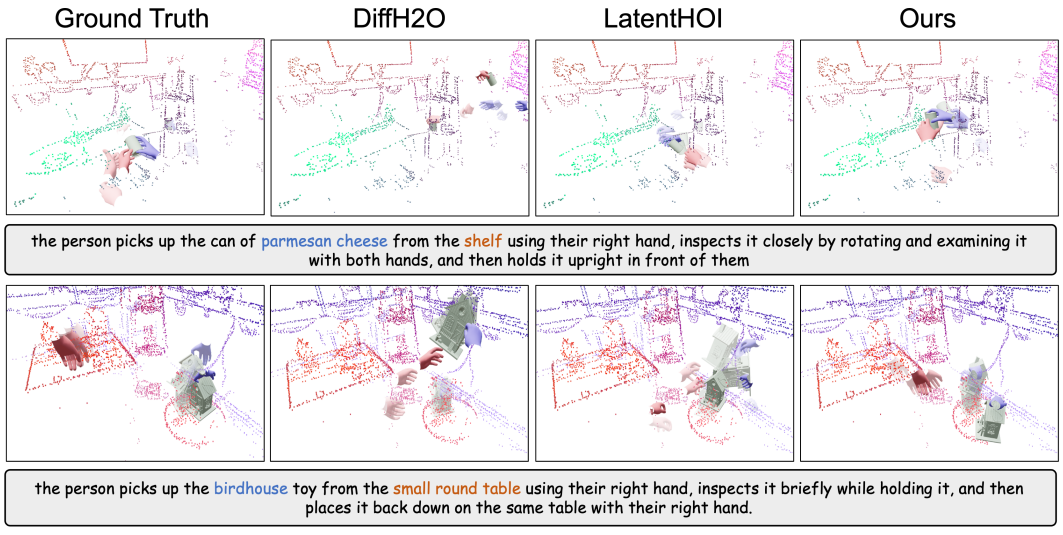

Figure 11: **Additional qualitative results on HOT3D with scene context.** The generated hand trajectories adapt to the surrounding scene layout, correctly interacting with objects on shelves and tables.

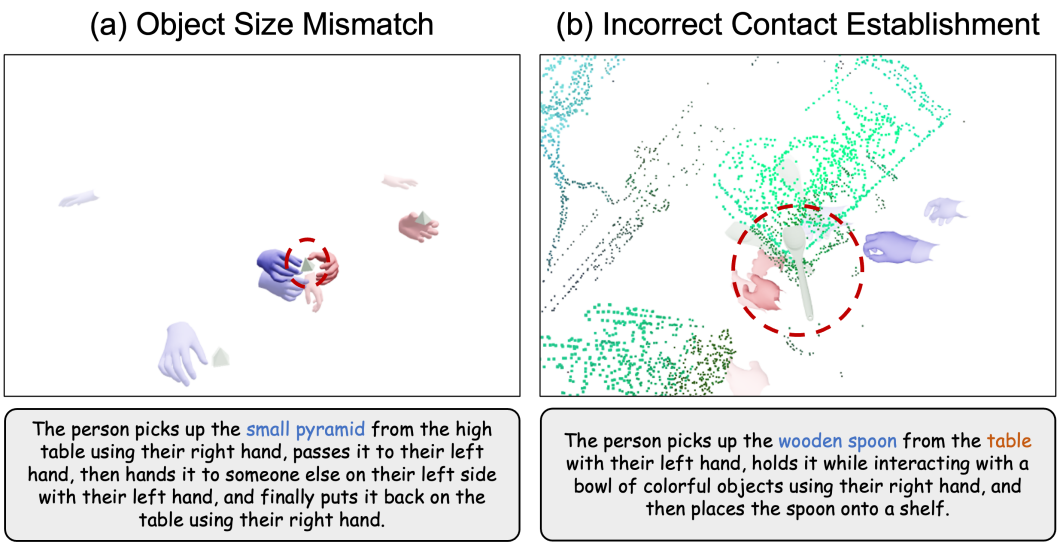

Figure 12: **Representative failure cases.** Red dashed circles highlight the erroneous regions. (a) *Object size mismatch*: the small pyramid floats away from the hand due to a mismatch between object size and generated grasp aperture. (b) *Incorrect contact establishment*: the hand fails to properly wrap around the wooden spoon, resulting in physically implausible finger-object contact. Left: GRAB [26], right: HOT3D [29] with scene context.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three contributions (a two-stage flow-matching HOI generator, semantic grounding via 3D scene tokens with a motion-text alignment loss, and an EgoDex reconstruction pipeline), which match Sec. 4, and the quantitative claims (1.7× SR on GRAB, 40× inference speedup, 5.75× real-robot SR) are supported by Table 1 and Table 2 in Sec. 5. The real-robot result is stated as nominal open-loop executability rather than robustness to perturbations, and the two-stage decomposition is presented as a design choice with a measured contact benefit (Supp. Sec. J).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed explicitly in Sec. 6: (i) reliance on reasonably accurate initial hand and object state estimation, (ii) the kinematic-only nature of generated trajectories that depend on downstream controllers for dynamics, as reflected in the open-loop execution of our four real-robot tasks, and (iii) FlowHOI being an interaction planner rather than a closed-loop controller that cannot react online to perturbations or contact slippage. Sec. 6 also states the scope of the work (contact-rich articulated-hand HOI generation and retargeting under nominal conditions), the dependence on offline refinement, and that repeated release and regrasp, force-level contact modes, deformable objects, and general in-hand manipulation are not established.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical generative-modelling framework and does not introduce new theorems or formal proofs; we rely on the existing flow-matching formulation [73] reviewed in Sec. 3.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 4 describes the full framework, training objectives, and conditioning structure; Sec. 5.2 reports the optimizer, batch size, and integration steps; and Supp. Sec. A, C, E, and F provide complete network, dataset, physics-simulation, and real-robot details.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code, model checkpoints, and the EgoDex-derived HOI training data will be released at `https://github.com/huajian-zeng/flowhoi` with documentation and reproduction instructions; the underlying datasets (GRAB, HOT3D, EgoDex) are already publicly accessible from their respective authors.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Sec. 5.1, Sec. 5.2, and Supp. Sec. A and D report data splits, optimizer and learning-rate schedule, training steps, classifier-free guidance scale, integration steps, and evaluation protocols for both our method and baselines.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For the real-world experiments (Table 2), we report 95% Wilson confidence intervals on the binomial success rates: aggregated over 40 trials, FlowHOI reaches 57.5% $[42.2\%, 71.5\%]$ vs. DiffH2O 10.0% $[4.0\%, 23.0\%]$, and the intervals do not overlap, indicating a statistically significant gap. The simulation results (Table 1) are computed over large test sets (10 trials per sequence over 210 GRAB and 361 HOT3D test sequences, i.e., 2,100 and 3,610 trials respectively); the binomial standard error on SR is therefore at most $\sqrt{0.25/2{,}100} \approx 1.1\%$ on GRAB and below 0.4% on HOT3D, which is much smaller than the gaps between methods. Each generative model is trained once with a fixed seed because of the training cost, so we do not report variability across training seeds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 5.2 states that all training and inference experiments run on a single NVIDIA RTX 6000 Ada GPU (48 GB) with batch size 64; Supp. Sec. A additionally lists per-stage training steps (500K for grasping, 200K for manipulation) and inference time per sequence (≈ 0.16 s).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The research uses publicly released datasets (GRAB, HOT3D, EgoDex) under their original licenses, performs no new data collection from human subjects, and produces no high-risk artefacts that require additional safeguards.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Generating physically plausible hand-object interactions can advance assistive robotics, accessibility, and skill learning from human videos. Potential negative impacts include automation displacement in manipulation-related labor, and the propagation of biases inherited from the MLLM used for sub-instruction extraction; we do not foresee direct misuse pathways comparable to those of generative image or speech models.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained language models, image generators, or scraped datasets that pose a high risk of misuse; the artefacts to be released are limited to a hand-object motion generator and reconstruction pipeline that produce kinematic trajectories for downstream manipulation research.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets are cited at first use and used under their respective licenses. Datasets: GRAB [26] (MPI Software Copyright License for non-commercial scientific research), HOT3D [29] (custom "HOT3D License Agreement" for the dataset; release code under Apache 2.0), and EgoDex [24] (CC BY-NC-ND 4.0). Body/hardware models: MANO [77] (MPI Software Copyright License for non-commercial scientific research), Allegro Hand [103] (commercial hardware product, no software license required). Open-source tools: T5 [84] (Apache 2.0, code and weights), SAM3 [79] (custom Meta SAM License), SAM3D [81] (custom Meta SAM License), DepthAnything3 [80] (code: Apache 2.0; the model weights we use carry CC BY-NC 4.0 where applicable), Concerto [88] (code: Apache 2.0; weights: CC BY-NC 4.0), SceneSplat [89] (CC BY-SA 4.0), pyroki [99] (MIT), DexTrack [101] (BSD 3-Clause), Gaussian-LIC [104] (GPL v3, used only as an external preprocessing tool to produce 3DGS scenes; no GPL-licensed code is incorporated into the FlowHOI codebase). All assets are used in accordance with their respective non-commercial / research-only / permissive terms; we do not redistribute any of the underlying datasets or pretrained weights, and our use is restricted to academic research consistent with each license.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce two new assets: (i) the EgoDex-derived HOI training data produced by our reconstruction pipeline (transition frames, reconstructed object meshes, and aligned MANO trajectories), and (ii) the trained FlowHOI grasping and manipulation model checkpoints. Both are documented in detail in the main paper and supplementary material, including the full data construction procedure (Sec. 4.2, Supp. Sec. C), reconstruction-quality validation against HOT3D ground truth (Supp. Sec. C), network architecture, training hyperparameters, optimizer schedule, and intended use (Supp. Sec. A). Both assets will be released at `https://github.com/huajian-zeng/flowhoi` with full documentation, license terms, and reproduction instructions.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or new research with human subjects; we rely entirely on already-released datasets (GRAB, HOT3D, EgoDex) collected and licensed by their original authors.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not collect data from human subjects ourselves; the datasets used in this work were collected by their original authors under their own institutional approvals and ethical reviews.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use OpenAI GPT-4o (an MLLM) in two non-core auxiliary roles: (i) extracting the grasp-focused sub-instruction T_g from the full action description (Sec. 4.3), and (ii) generating natural-language action descriptions for HOT3D clips (Supp. Sec. D). LLMs are not part of the FlowHOI generative model itself.