HO-Flow: Generalizable Hand-Object Interaction Generation with Latent Flow Matching

Zerui Chen[1], Rolandos Alexandros Potamias[1], Shizhe Chen[2], Jiankang Deng[1], Cordelia Schmid[2], and Stefanos Zafeiriou[1]

[1]Imperial College London, UK [2]Inria, ENS, CNRS, PSL, France

https://zerchen.github.io/projects/hoflow.html

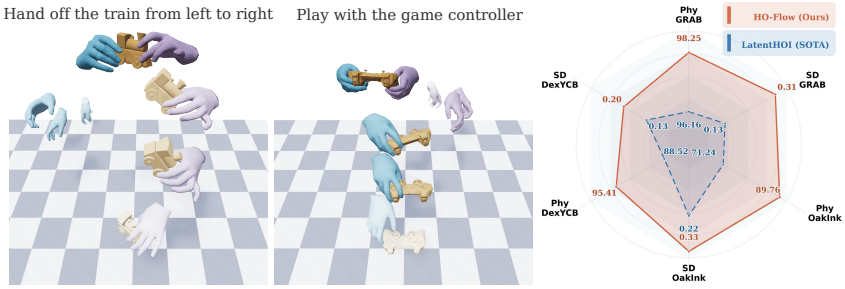

Fig. 1: Our approach can synthesize realistic hand–object interactions for diverse tasks (left), where darker colors represent later time steps. It outperforms the state-of-the-art LatentHOI [26] across three benchmarks (right), namely GRAB [49], DexYCB [5] and OakInk [58]. SD and Phy refer to sample diversity and physical plausibility.

Abstract. Generating realistic 3D hand-object interactions (HOI) is a fundamental challenge in computer vision and robotics, requiring both temporal coherence and high-fidelity physical plausibility. Existing methods remain limited in their ability to learn expressive motion representations for generation and perform temporal reasoning. In this paper, we present HO-Flow, a framework for synthesizing realistic hand-object motion sequences from texts and canonical 3D objects. HO-Flow first employs an interaction-aware variational autoencoder to encode sequences of hand and object motions into a unified latent manifold by incorporating hand and object kinematics, enabling the representation to capture rich interaction dynamics. It then leverages a masked flow matching model that combines auto-regressive temporal reasoning with continuous latent generation, improving temporal coherence. To further enhance generalization, HO-Flow predicts object motions relative to the initial frame, enabling effective pre-training on large-scale synthetic data. Experiments on the GRAB, OakInk, and DexYCB benchmarks demonstrate that HO-Flow achieves state-of-the-art performance in both physical plausibility and motion diversity for interaction motion synthesis.

Keywords: Hand-Object Interaction · Latent Model · Flow Matching

1 Introduction

Generating realistic 3D hand-object interactions (HOI) is a fundamental challenge with wide-ranging applications in character animation, virtual reality, and

robotic manipulation [7,12,32,37]. While previous research has achieved impressive results in synthesizing static grasping poses [21, 25, 28, 35, 66, 69], generating dynamic, temporal HOI sequences from text [11] remains significantly more challenging and relatively under-explored. Extending generative frameworks to the temporal 4D domain introduces substantial difficulties, including maintaining physical plausibility, preserving temporal coherence over long time horizons, and achieving generalization across diverse actions and objects.

A central challenge lies in preventing physical inconsistencies, particularly interpenetration artifacts where the hand geometry passes through the manipulated object during motion. Addressing this issue requires a *highly expressive motion representation* capable of modeling the coupled dynamics between the hand and the object. Early approaches [11] adopt diffusion models [18] to directly generate sequences of raw hand and object poses. However, the high degree of freedom of articulated hand models result in considerable computational cost, and treating poses as unconstrained independent variables often fails to faithfully model the underlying interaction distribution, producing artifacts such as floating hands or implausible contacts. More recent methods, such as LatentHOI [26], alleviate this issue by embedding hand-object poses into a *compact latent space*. Nevertheless, existing latent representations are typically learned frame-independently, which weakens temporal coherence and does not encode fine-grained, contact-aware cues that are necessary for high-fidelity synthesis.

Beyond representation, the generative architecture itself plays a crucial role. *Auto-regressive (AR) Transformers* [20] are effective at modeling long-range temporal dependencies, but they typically require quantizing continuous motions into discrete tokens [51]. Such discretization can suppress contact-rich features that are essential for realistic manipulation. In contrast, *diffusion and flow-based models* [11, 26] operate directly in continuous space, thereby avoiding quantization artifacts. However, these models usually perform holistic denoising over entire sequences, which incurs substantial computational cost. As a result, they often rely on U-Net architectures rather than more expressive Transformer-based models, and tend to produce globally inconsistent motions.

In this work, we present HO-Flow to synthesize realistic hand-object motion sequences based on text prompts. First, we learn an expressive motion representation using a novel *interaction-aware variational autoencoder (Inter-VAE)*. Unlike frame-wise motion encoders, Inter-VAE embeds short-horizon temporal sequences of hand-object interactions in latent spaces. To enrich fine-grained interaction cues, we transform object point clouds into hand-centric coordinates using the hand kinematic chain, and capture both global motion trajectories and contact-rich hand-object coordination in the latent. Building upon this representation, we introduce a masked flow matching model to generate motion latents. This design attends to the full temporal context via the auto-regressive masked transformer while operating in continuous space via a lightweight flow matching head, significantly reducing jitter and outlier frames while promoting long-range coherence and smoothness. To further improve generalization, HO-Flow predicts object motions relative to the initial frame, avoiding dataset-specific

coordinate conventions, which enables effective pre-training on large-scale synthetic data [62] and yields superior transfer to unseen objects. We evaluate HO-Flow on three challenging benchmarks, including GRAB [49], OakInk [58], and DexYCB [5]. Experimental results demonstrate that HO-Flow advances the state of the art in terms of both physical plausibility and motion diversity.
Our contributions are summarized as follows:
• We propose HO-Flow for text-driven hand–object interaction motion generation, combining an expressive motion representation with an efficient temporal generative modeling to synthesize realistic and coherent HOI sequences.
• We introduce Inter-VAE to enhance motion representation learning, which encodes temporal interaction sequences into a unified latent space, capturing both global motion trajectories and fine-grained contact-rich coordination.
• HO-Flow achieves state-of-the-art performance on three challenging benchmarks for hand–object motion generation, benefiting from its expressive motion representation, auto-regressive temporal reasoning, and large-scale pre-training.

2 Related Work

Hand grasp generation. Synthesizing realistic hand grasps for 3D objects is a long-standing problem in computer vision and robotics [3]. Early approaches rely on optimization to find physically plausible grasps given known object geometry [13, 28, 35]. GraspIt! [35], for instance, formulates grasp synthesis as an optimization problem over hand configurations to maximize a physically grounded quality metric. Subsequent works incorporate force-closure constraints [28, 67] and biomechanical contact priors [58, 59] to further improve realism. With the advent of deep learning, data-driven methods train conditional variational autoencoders [23, 31, 49, 55] or transformers [53, 56] to generate hand grasps conditioned on object shape, yielding diverse and plausible configurations. More recent works introduce richer conditioning signals: NL2Contact [66] and SemGrasp [25] leverage natural language and semantic cues to guide synthesis toward task-relevant grasps, while affordance-based methods [21, 54, 61] explicitly model the relationship between object geometry and plausible hand placements. Despite their impressive results, all these methods predict only static hand grasping poses for objects. In this work, we address the more challenging problem of generating dynamic hand-object interaction sequences.
Hand and object motion generation. Generating hand-object interaction motions from task descriptions has attracted growing attention, with approaches broadly divided into optimization-based and learning-based methods. Unlike human motion generation [1, 39, 63], this task emphasizes physical naturalness for interactions. Optimization-based methods leveraging reinforcement learning [45, 48, 52, 57] or physics-based constraints [60, 65] can yield physically plausible results, but require hours of per-sequence optimization and tend to produce monotonous motions, limiting their scalability. Among learning-based approaches, HOI-GPT [20] auto-regressively generates discrete motion tokens via VQ-VAE [51], through quantizing continuous poses into a finite codebook inher-

ently limits motion diversity and fine-grained interaction fidelity. Diffusion models [18] avoid this bottleneck by operating in continuous spaces. DiffH2O [11] directly denoises explicit hand and object pose sequences, but jointly modeling 3D positions and rotations that lie on incompatible manifolds renders the problem ill-posed and limits motion quality. LatentHOI [26] mitigates this by encoding poses into a latent space before applying diffusion, yet its frame-by-frame encoding captures only local grasping poses while neglecting global hand trajectories and object motions, resulting in temporally inconsistent and jittery outputs. In contrast, HO-Flow holistically encodes full hand-object interaction sequences into a unified structured latent space and auto-regressively generates temporally coherent and fine-grained interaction motions.

Diffusion and flow matching models. Diffusion models have emerged as a powerful paradigm for high-fidelity generative modeling across images, videos, and human motions. They define a forward process that gradually perturbs data into noise, and learn a reverse denoising process to sample new instances [18, 46, 47]. This framework has driven substantial progress in image synthesis, especially when combined with improved noise schedules and parameterizations [22, 36], and scalable latent-space formulations for efficient high-resolution generation [43]. Extending diffusion to videos introduces additional challenges in modeling spatio-temporal coherence, and recent works tackle this via temporally-aware architectures and conditioning strategies [2, 17]. Diffusion-based motion generators similarly benefit from operating in continuous spaces and have shown strong performance for text-conditioned human motion synthesis, thanks to their ability to model multi-modal motion distributions and long-range temporal dependencies [50, 64]. In parallel, flow-based continuous-time generative models provide an alternative that learns a velocity field to transport noise into data via ODE integration. Flow matching [27] and rectified flow [29] unify and simplify training objectives for such continuous-time generators, enabling faster sampling with fewer integration steps while retaining high sample quality. Motivated by these advances, we introduce HO-Flow, which uses interaction-aware latents and masked flow matching for efficient and coherent hand-object motion generation.

3 Method

In this section, we introduce HO-Flow (Figure 2), a novel framework that generates Hand–Object interaction motions from task descriptions using an auto-regressive Flow matching model. We first outline the overall architecture (Section 3.1), then detail two core components: (i) an interaction-aware VAE (Section 3.2) that encodes explicit hand–object poses into compact motion tokens, and (ii) a context-aware auto-regressive flow matching model (Section 3.3) that generates coherent interaction sequences conditioned on task specifications.

3.1 Overview

Figure 2 shows HO-Flow, a novel approach for realistic hand-object interaction synthesis conditioned on the text description and the object point cloud.

Realistic hand-object interaction synthesis requires capturing contact-rich dynamics and long-range temporal dependencies beyond frame-wise pose representations. To this end, the proposed method introduces two coupled components: (1) an interaction-aware VAE that compresses hand-object dynamics into structured latents for holistic sequence-level reasoning, and (2) a flow matching model that auto-regressively generates these latents in continuous space for efficient, temporally coherent synthesis.

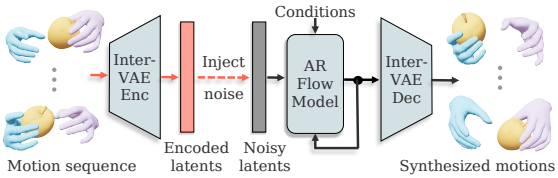

Fig. 2: Overview of HO-Flow. Red and gray blocks indicate encoded and noisy motion latents, respectively. Red arrows are active only during training. The framework synthesizes realistic hand-object interactions using two components: an interaction-aware VAE for compact motion latents with fine-grained interaction features, and an auto-regressive flow-matching model that predicts successive latents for faithful, temporally coherent synthesis.

The interaction-aware VAE employs a symmetric encoder-decoder architecture. The encoder leverages an object point cloud alongside hand and object motion sequences to learn unified latent representations. Within the encoder, a spatial module first extracts per-frame interaction features, which are subsequently aggregated by temporal encoders to produce distinct latent codes for the hand and the object. To regularize the latent space and ensure reconstruction fidelity, dedicated hand and object decoders are trained to recover the original motion sequences from these learned spatial-temporary interaction features.

Following the training of the motion latents, we train an auto-regressive flow matching model. Conditioned on a sparse object point cloud and a natural language description, this model utilizes a masked transformer architecture and auto-regressively aggregates motion context to predict successive motion tokens via a flow matching objective. During inference, the frozen decoders from the interaction-aware VAE decode the generated latent codes back into the space of hand and object poses.

3.2 Interaction-aware Variational Autoencoder

As illustrated in Figure 3, we propose an interaction-aware variational autoencoder (VAE) to compress hand-object interaction sequences into compact, structured latent representations. The VAE takes the object point cloud $\mathbf{p}_o \in \mathbb{R}^{1024\times 3}$, MANO [44] hand poses $\boldsymbol{\theta}_h \in \mathbb{R}^{\mathrm{N}\times 16\times 3}$, and object poses $\mathbf{T}_o \in \mathbb{R}^{\mathrm{N}\times 4\times 4}$ as inputs, where N denotes the

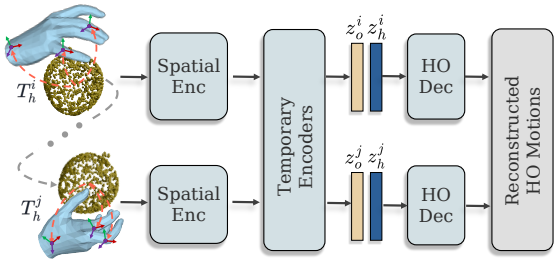

Fig. 3: Interaction-aware VAE captures fine-grained interaction features for latent representations ($\mathbf{z}_o$ and $\mathbf{z}_h$). $\mathbf{T}_h$ is the transformation of hand bones.

number of motion frames. To mitigate the scarcity of bi-manual manipulation data, we follow [26] by focusing on the object manipulation with the right hand and mirroring left-hand data when available.

Kinematic-aware object transformation. To capture fine-grained interaction features, we project the object point clouds into the local coordinate systems of various hand joints. We first transform the object point clouds into the world coordinate system, $\mathbf{p}_o^w \in \mathbb{R}^{\mathrm{N}\times 1024\times 3}$, using the pose sequence $\mathbf{T}_o$. We then derive the global transformation $\mathbf{T}_{h,t}^i \in \mathbb{R}^{4\times 4}$ for the i-th hand joint using its pose $\boldsymbol{\theta}_{h,t}^i$ and the MANO kinematic tree:

$$\mathbf{T}_{h,t}^i = \prod_{j\in A(i)} \left[\begin{array}{c|c} \exp\left(\boldsymbol{\theta}_{h,t}^j\right) & \boldsymbol{\phi}_h^j \\ \hline \mathbf{0} & 1 \end{array}\right], \tag{1}$$

where $A(i)$ is the ordered set of ancestors of the i-th joint and $\exp(\cdot)$ is the Rodrigues formula converting axis-angle $\boldsymbol{\theta}_{h,t}^j$ to a rotation matrix. The translation $\boldsymbol{\phi}_h^j$ is the joint's offset relative to its parent as defined in the MANO template.

Multi-view interaction features. By traversing the kinematic chain, we obtain global transformations for all joints. We then transform the world-frame object point clouds $\mathbf{p}_o^w$ into each joint's local frame via $(\mathbf{T}_{h,t}^i)^{-1}$ and concatenate them to form $\mathbf{p}_o^h \in \mathbb{R}^{N\times 1024\times(16\times 3)}$, encoding rich hand-relative geometric context:

$$\mathbf{p}_o^h = \bigoplus_{i=0}^{15} \widetilde{\mathrm{H}}((\mathbf{T}_{h,t}^i)^{-1} \cdot \mathrm{H}(\mathbf{p}_o^w)), \tag{2}$$

where $\bigoplus$ denotes concatenation, $\mathrm{H}(\cdot)$ denotes conversion to homogeneous coordinates, and $\widetilde{\mathrm{H}}(\cdot)$ denotes inverse projection back to Euclidean space.

Spatial-Temporal encoding. The concatenated point clouds $\{\mathbf{p}_o, \mathbf{p}_o^w, \mathbf{p}_o^h\}$ are processed by a spatial encoder based on the PointNet++ architecture [41], which shares weights across different frames and utilizes three set abstraction layers to extract features $\mathbf{f}_s \in \mathbb{R}^{N\times 768}$. These features are subsequently fused with $\boldsymbol{\theta}_h$ and $\mathbf{T}_o$ via two MLP layers to yield hand and object spatial features $\mathbf{f}_h, \mathbf{f}_o \in \mathbb{R}^{256}$, respectively. Finally, temporal encoders, comprised of 1D convolutional layers with a total stride of 4, aggregate these features across the sequence to produce compact latent codes $\mathbf{z}_h, \mathbf{z}_o \in \mathbb{R}^{\frac{\mathrm{N}}{4}\times 32}$.

Motion Reconstruction. To optimize the latent space, dedicated hand and object decoders upsample the latent features to the original sequence length and utilize convolutional layers to reconstruct the hand and object motion trajectories. We supervise the VAE using reconstruction losses over both hand and object motions, combined with KL regularization on the latent posteriors. The overall objective is formulated as:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{mesh}} + \mathcal{L}_{\text{obj-rot}} + \mathcal{L}_{\text{obj-trans}} + \beta\left(\mathcal{L}_{\text{KL}}^h + \mathcal{L}_{\text{KL}}^o\right), \tag{3}$$

where we utilize ℓ_1 losses for all reconstruction terms and set β to 1e-4. $\mathcal{L}_{\text{pose}}$ penalizes errors in the 6D rotation representation [68] for all MANO joints. $\mathcal{L}_{\text{trans}}$ supervises the hand translation predicted relative to the object translation to

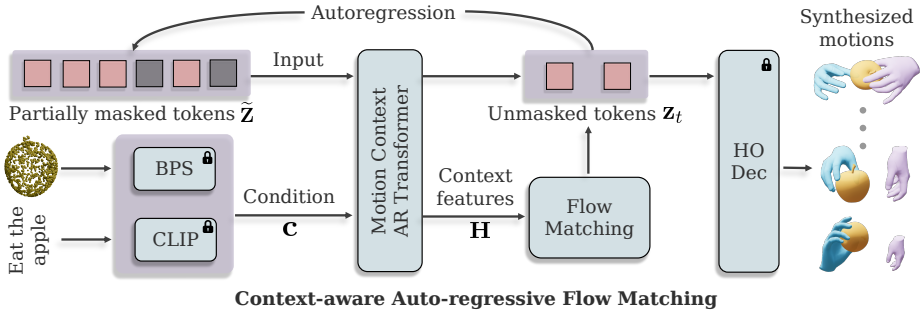

Fig. 4: Illustration of our flow matching model. Gray and pink blocks indicate masked and unmasked tokens, respectively. Conditioned on object point clouds and task descriptions, the model aggregates motion context to autoregressively predict successive motions, thereby reducing uncertainty and enabling high-fidelity interaction synthesis.

encourage stable interaction geometry. $\mathcal{L}_{\text{mesh}}$ enforces vertex-level consistency by reconstructing MANO meshes. For the object, $\mathcal{L}_{\text{obj-rot}}$ and $\mathcal{L}_{\text{obj-trans}}$ supervise the object's 6D rotation and 3D translation. Following variational auto-encoding [24], the KL terms regularize the temporally-strided latent sequences:

$$\mathcal{L}_{\text{KL}}^{(\cdot)} = \text{KL}\big(q(\mathbf{z}_{(\cdot)} \mid \mathbf{p}_o, \boldsymbol{\theta}_h, \mathbf{T}_o) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big), \quad (\cdot) \in \{h, o\}. \tag{4}$$

3.3 Auto-regressive Flow Matching

As illustrated in Figure 4, inspired by recent advances in motion generation [34, 70], we employ autoregressive flow matching to generate interaction-motion latents conditioned on the object geometry $\mathbf{p}_o$ and natural-language task description s. Leveraging the structured latent space $\mathbf{Z}$ learned by our interaction-aware VAE, our model estimates $p(\mathbf{Z} \mid \mathbf{p}_o, s)$ to synthesize temporally coherent and realistic hand-object interactions.

Given an input sequence of N frames, the temporal stride of the VAE yields a latent sequence of length $L = \text{N}/4$. For each temporal step t, we concatenate the object and hand latents for both the right hand and its mirrored left counterpart to form a unified motion token:

$$\begin{aligned} \mathbf{z}_t &= \left[\mathbf{z}_{o,t}^r \parallel \mathbf{z}_{h,t}^r \parallel \mathbf{z}_{o,t}^l \parallel \mathbf{z}_{h,t}^l\right] \in \mathbb{R}^{128}, \\ \mathbf{Z} &= [\mathbf{z}_1, \ldots, \mathbf{z}_L] \in \mathbb{R}^{\frac{\text{N}}{4} \times 128}, \end{aligned} \tag{5}$$

where $\mathbf{z}_{o,t}^{(\cdot)}$, $\mathbf{z}_{h,t}^{(\cdot)} \in \mathbb{R}^{32}$ denote the object and hand latent codes, respectively.

Condition encoding. We encode the task description s using a frozen CLIP [42] text encoder to obtain a semantic embedding $\mathbf{e}_{\text{text}} \in \mathbb{R}^{512}$. To represent the object geometry $\mathbf{p}_o$ as a fixed-length descriptor, we follow previous practice [11, 26] to compute a Basis Point Set (BPS) [40] encoding, resulting in distance features $\mathbf{e}_{\text{bps}} \in \mathbb{R}^{4096}$. These features are projected into a shared conditioning space via linear layers and summed to form the final condition vector:

$$\mathbf{c} = \mathbf{W}_{\text{text}}\mathbf{e}_{\text{text}} + \mathbf{W}_{\text{bps}}\mathbf{e}_{\text{bps}} \in \mathbb{R}^d, \tag{6}$$

where we set d to 1024 in the implementation.

Context-aware transformer. To capture long-range temporal dependencies while facilitating progressive generation, we adopt a masked auto-regressive transformer (MAR) architecture over the latent tokens. The sequence of motion tokens $\mathbf{Z}$ is embedded, augmented with sinusoidal positional encodings, and processed by a stack of transformer blocks. Each block utilizes adaptive LayerNorm (AdaLN) [38] modulation conditioned on $\mathbf{c}$ to yield contextual features $\mathbf{H} = \mathrm{MAR}(\widetilde{\mathbf{Z}}, \mathbf{c}) \in \mathbb{R}^{\frac{\mathrm{N}}{4} \times d}$, where $\widetilde{\mathbf{Z}}$ represents the partially masked input sequence during training or inference.

Flow matching in latent space. For each masked latent position $t \in \mathcal{M}$, we learn a conditional flow matching model to map a prior noise distribution to the target token $\mathbf{z}_t$. We define a probability path as a straight-line linear interpolation between noise $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the ground-truth latent $\mathbf{x}_1 = \mathbf{z}_t$:

$$\mathbf{x}_\tau = \tau \mathbf{x}_1 + (1 - \tau)\mathbf{x}_0, \quad \tau \in [0, 1]. \tag{7}$$

The corresponding conditional vector field is the time-derivative of this path, which serves as the ground-truth supervision target for our model:

$$\mathbf{u}_t(\mathbf{x}_\tau) = \frac{d\mathbf{x}_\tau}{d\tau} = \mathbf{x}_1 - \mathbf{x}_0. \tag{8}$$

We parameterize a velocity field $v_\phi(\mathbf{x}_\tau, \tau; \mathbf{h}_t)$ using a lightweight SiT-style MLP [33] conditioned on the contextual features $\mathbf{h}_t$. The model is trained to regress the constant velocity $\mathbf{u}_t$ by minimizing the flow matching objective. This formulation effectively guides noise along a deterministic trajectory toward the hand-object interaction manifold, enabling efficient inference via ODE solvers.

In training, we pre-compute latent targets $\mathbf{Z}$ from the dataset with the frozen VAE. The model is trained to inpaint masked tokens using a masked modeling strategy. Specifically, we sample a subset of token positions $\mathcal{M}$ according to a cosine mask-ratio schedule. These positions are corrupted using a BERT-style strategy: 80% are replaced by a learnable `[MASK]` token, 10% by random Gaussian noise, and 10% are left unchanged. For each masked position $t \in \mathcal{M}$, we minimize the flow matching objective:

$$\mathcal{L}_{\mathrm{AR}} = \mathbb{E}_{\mathcal{M}, \tau, \mathbf{x}_0} \left[\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \| v_\phi(\mathbf{x}_\tau, \tau; \mathbf{h}_t) - (\mathbf{z}_t - \mathbf{x}_0) \|_2^2 \right], \tag{9}$$

where $\mathbf{h}_t$ is the contextual feature extracted by the MAR transformer from the corrupted sequence. We apply classifier-free guidance [19] by randomly dropping the text condition during training (20% probability) and maintain an exponential moving average (EMA) of model parameters for stable inference.

4 Experimental Evaluation

We conduct comprehensive experiments on three hand-object motion benchmarks to validate the effectiveness of HO-Flow in synthesizing both bi-manual and single-hand manipulation motions across diverse objects.

4.1 Datasets

GRAB [49]: The dataset contains bimanual manipulation tasks. Following the object-based split in [26], we reserve 4 objects for testing and 47 for training. To focus on the manipulation phase, training sequences begin at the first contact frame, and sequences are padded or truncated to a maximum of 160 frames. The unseen test split consists of 17 (text prompt, object) pairs.
OakInk [58]: To evaluate out-of-distribution generalization, we use a challenging novel-object split [26] from this dataset, selecting 100 objects across 20 categories. These unseen shapes are paired with intents from GRAB, resulting in 212 evaluation pairs. As no motion data is used for training on this split, we use the model trained on GRAB for this evaluation setup.
DexYCB [5]: For single-hand grasping, we employ an object-based split by reserving 4 unseen objects out of 20 for testing. Sequences start from the first frame and are padded to a maximum of 96 frames.
GraspXL [62]: It provides over five million synthetic right-hand manipulation trajectories across half a million diverse objects, generated via RL in a physics-based simulator. While physically plausible, it is restricted to relocation tasks and lacks motion diversity. We therefore leverage this large-scale data for pre-training to learn interaction priors before fine-tuning on more complex motions.

4.2 Evaluation metrics

We evaluate our method in terms of latent quality, physical plausibility and diversity of interaction motions, following prior works [8–11, 23, 26]. Please refer to the supplementary material for details of different evaluation metrics.

To evaluate the quality of the learned latent representation, we report mean joint error (E_j) and mean vertex error (E_v) in mm for the hand, and mean translation error (E_o) in mm and chamfer distance (CD_o) in cm^2 for the object, where CD_o also reflects orientation error and is robust to object symmetries. We further assess temporal smoothness for both hand and object trajectories using acceleration and jerk, denoted as Acc_h, Acc_o, $Jerk_h$, and $Jerk_o$, respectively, which are computed as the second- and third-order finite differences of the corresponding trajectories.

To evaluate the physical plausibility and diversity of generated interaction motions, we adopt the following evaluation metrics for motion generation.
Contact ratio (CR). We compute the percentage of hand vertices in contact with the object surface and report the average ratio (%) over the sequence.
Interpenetration (IV/ID). To quantify collisions, we report the hand-object intersection volume (IV, cm^3) and the maximum penetration depth (ID, cm), both averaged over frames with non-zero interpenetration.
Physics score (IVU/Phy). We report the interpenetration volume normalized by the estimated contact area (IVU) and the percentage of frames, Phy (%), in which the object moves while maintaining hand-object contact.
Sample diversity (SD). We compute sample diversity as the average pairwise ℓ_2 distance among multiple generations conditioned on the same input.

Table 1: Ablation experiments for interaction-aware autoencoder on GRAB dataset. Our proposed model that captures both spatial and temporal interaction features yields the best performance in encoding hand and object motions.

	Object kine.	Hand kine.	Sep. latents	Pre-train	$E_v^r \downarrow$	$E_j^r \downarrow$	$E_v^l \downarrow$	$E_j^l \downarrow$	$E_o \downarrow$	$CD_o \downarrow$
R1	×	×	×	×	23.93	24.34	19.14	19.55	16.34	1.36
R2	✓	×	×	×	22.15	22.56	20.51	20.58	8.93	0.71
R3	✓	✓	×	×	21.69	21.97	16.59	16.90	5.19	0.37
R4	✓	✓	✓	×	17.29	17.61	13.05	13.33	3.70	0.26
R5	×	×	×	✓	17.84	18.22	14.76	15.08	11.52	0.86
R6	✓	✓	✓	✓	**9.13**	**9.41**	**8.03**	**8.31**	**2.33**	**0.12**

4.3 Implementation details

Model architecture. We implement the two-stage HO-Flow framework described in Section 3. In the interaction-aware VAE, temporal encoders and decoders comprise two stride-2 downsampling and upsampling stages with a hidden dimension of 512. For the context-aware autoregressive model, we encode text features using a frozen CLIP ViT-B/32 and employ a single-layer transformer with 16 attention heads to efficiently encode the context features $\mathbf{h}_t$. Additional architectural details are provided in the supplementary material.

Training and inference Details. We optimize all models using AdamW [30] with linear warmup and cosine learning rate decay. The interaction-aware VAE is first pre-trained on GraspXL [62] for 100k iterations with a batch size of 32 and a base learning rate of 2×10^{-4} (decayed to 2×10^{-5}) across 4 NVIDIA H100 GPUs. Then, it is fine-tuned on the specific dataset under the same setting but with a base learning rate of 1×10^{-4}. To enhance robustness, we apply random global 3D rotations to the whole manipulation sequence as data augmentation. Once converged, the VAE is frozen to pre-compute latent targets for the generative stage. The latent flow matching model is also first pre-trained on GraspXL and subsequently fine-tuned on specific dataset for 300k iterations with a batch size of 32 on a single NVIDIA H100 GPU, following the same warmup and decay schedules used for Inter-VAE. We employ an Exponential Moving Average (EMA) with a decay of 0.9999 for stability. During inference, we generate samples using 18 steps with a classifier-free guidance weight of 1.5.

4.4 Ablation studies

We conduct comprehensive ablation studies on the GRAB dataset to evaluate the effectiveness of each component of our proposed HO-Flow approach.

Interaction-aware variational autoencoder. Table 1 presents an ablation study of the proposed components within our VAE model. The baseline (**R1**), which utilizes canonical object point clouds alongside hand and object poses, fails to capture the spatial interactions in world coordinates, leading to unsatisfactory performance. By incorporating 6D object poses and processing object point

Table 2: Ablation experiments for the auto-regressive flow matching model on GRAB dataset. Our approach achieves the best performance over other variants. Seq. indicates whether the model takes a temporal sequence as input. r and l denote right and left.

	Representation		Auto-reg.	Pre-train	$IV_r \downarrow$	$IV_l \downarrow$	$ID_r \downarrow$	$ID_l \downarrow$	$CR_r \uparrow$	$CR_l \uparrow$	IVU $\downarrow$	Phy $\uparrow$	SD $\uparrow$
	Latent	Seq.											
R1	×	✓	×	×	10.35	2.67	1.14	0.48	11.68	3.73	0.13	84.89	0.24
R2	✓	×	×	×	8.13	1.94	0.98	0.43	**12.54**	0.71	0.13	88.76	0.12
R3	✓	✓	×	×	6.99	1.82	0.82	0.36	10.27	3.07	0.11	92.69	0.26
R4	✓	✓	✓	×	5.48	1.35	0.63	0.30	11.26	2.86	0.08	97.94	0.30
R5	✓	✓	✓	✓	**5.31**	**1.24**	**0.61**	**0.26**	11.38	**3.96**	**0.07**	**98.25**	**0.31**

clouds directly in the world coordinate space (**R2**), we observe a significant reduction in object reconstruction errors (E_o and CD_o). Transforming these point clouds into various hand-joint coordinate systems (**R3**) further enhances the extraction of hand-object interaction features, yielding consistent improvements across both hand and object metrics. In **R4**, we transition from a shared latent code to separate latent representations for the hand and object. This decoupling facilitates the extraction of more discriminative motion features, further boosting accuracy. **R5** further pre-trains **R1** on synthetic GraspXL data [62], leading to substantial performance improvements. Finally, pre-training Inter-VAE using GraspXL (**R6**) significantly improves generalization to unseen objects and achieves the best performance across all metrics. Notably, **R6** yields much larger gains over **R5** across all metrics, suggesting that our interaction-aware model design is particularly effective at leveraging synthetic interaction priors.

Auto-regressive latent flow matching. We conduct ablation studies on several variants of the generative model architecture, with results summarized in Table 2. The baseline **R1** directly generates hand and object motions in the original pose space. This design leads to limited physical fidelity, achieving the lowest physical plausibility score (84.89 in Phy) and severe right-hand interpenetration (10.35 in IV_r). We then consider **R2**, which adopts the latent representation of LatentHOI [26]. Specifically, it encodes motion frames independently and uses a VAE to embed local hand poses into a latent space while keeping global hand and object motions in explicit pose space. Without temporal latent modeling, **R2** struggles to capture holistic hand-object interaction patterns across the sequence, resulting in poor temporal coherence and physical realism (*e.g.*, 8.13 in IV_r and 88.76 in Phy), as well as reduced motion diversity (0.12 in SD). Compared with **R2**, **R3** introduces our proposed interaction-aware VAE (Section 3.2) to jointly encode hand-object interaction features in both spatial and temporal domains. This expressive latent representation improves overall physical fidelity (Phy increases to 92.69), reduces penetration and distance errors (*e.g.*, IV_r drops to 6.99, ID_r drops to 0.82), and substantially improves diversity (SD increases to 0.26). Finally, **R4** incorporates our context-aware transformer (Section 3.3) with an auto-regressive formulation, where each token is predicted conditioned on the generated motion context rather than predicted in parallel. This design

Table 3: Comparison with state-of-the-art methods on GRAB benchmark. HO-Flow demonstrates strong generalization ability to faithfully interact with diverse objects. pt denotes whether this model has been pre-trained on GraspXL data.

Methods	$IV_r \downarrow$	$IV_l \downarrow$	$ID_r \downarrow$	$ID_l \downarrow$	$CR_r \uparrow$	$CR_l \uparrow$	IVU ↓	Phy ↑	SD ↑
IMoS [15]	10.38	-	1.25	-	4.61	-	0.53	84.88	0.00
MDM [50]	9.12	2.61	1.24	0.51	8.21	1.29	0.19	89.81	0.18
MLD [6]	9.62	3.14	1.06	0.49	10.23	0.87	0.24	85.68	0.20
LatentHOI [26]	6.38	1.66	0.77	0.29	**11.94**	1.11	0.10	96.16	0.13
HO-Flow, wo pt	5.48	1.35	0.63	0.30	11.26	2.86	0.08	97.94	0.30
HO-Flow, w pt	**5.31**	**1.24**	**0.61**	**0.26**	11.38	**3.96**	**0.07**	**98.25**	**0.31**

Table 4: Comparison with state-of-the-art methods on OakInk benchmark. HO-Flow achieves superior performance in terms of both motion quality and diversity. pt denotes whether this model has been pre-trained on GraspXL data.

Methods	$IV_r \downarrow$	$IV_l \downarrow$	$ID_r \downarrow$	$ID_l \downarrow$	$CR_r \uparrow$	$CR_l \uparrow$	IVU ↓	Phy ↑	SD ↑
Text2HOI [4]	15.19	11.54	2.14	1.39	**11.24**	**6.33**	0.26	82.58	0.21
MDM [50]	8.46	2.47	1.69	0.34	4.97	1.02	0.20	60.89	0.23
MLD [6]	9.15	4.25	1.79	0.56	5.36	0.77	0.29	46.41	0.32
LatentHOI [26]	7.22	3.11	1.10	0.37	7.80	1.73	0.14	71.24	0.22
HO-Flow, wo pt	5.82	2.05	0.66	0.30	8.76	2.34	0.11	83.62	0.31
HO-Flow, w pt	**4.10**	**1.72**	**0.45**	**0.26**	8.85	3.78	**0.09**	**89.76**	**0.33**

further refines generation, yielding a sharp reduction in penetration metrics (IV_r from 6.99 to 5.48, IVU from 0.11 to 0.08) and boosting physical plausibility to 97.94. **R5** further incorporates pre-training on GraspXL, achieving the best performance across metrics (*e.g.*, 98.25 in Phy, 5.31 in IV_r, and 0.31 in SD).

4.5 Comparison with state of the art

In this section, we compare our proposed HO-Flow approach with state-of-the-art approaches on three mainstream hand-object motion benchmarks.

GRAB. Table 3 provides a comprehensive quantitative evaluation on GRAB benchmark. Compared with LatentHOI, HO-Flow substantially reduces unnatural interactions, achieving the lowest interpenetration volume (IV) and intersection depth (ID) for both hands. In particular, HO-Flow improves physical plausibility (Phy) from 96.16% to 98.25% and significantly increases semantic diversity (SD), more than doubling LatentHOI's score. We also evaluate temporal smoothness using acceleration and jerk, the second- and third-order trajectory differences. Table 6 shows that HO-Flow achieves lower acceleration and jerk for both hands and objects than prior methods, indicating fewer high-frequency artifacts. Overall, these results demonstrate that our flow-based approach generates

Table 5: Comparison with state-of-the-art methods on DexYCB benchmark. HO-Flow also shows strong performance for synthesizing single-hand manipulation sequences. pt denotes whether this model has been pre-trained on GraspXL data.

Methods	IV ↓	ID ↓	CR_r ↑	IVU ↓	Phy ↑	SD ↑
MDM [50]	7.78	2.10	8.87	0.12	86.22	0.13
LatentHOI [26]	7.70	2.01	11.98	0.13	88.52	0.13
HO-Flow, wo pt	6.84	1.82	12.06	0.11	90.77	0.20
HO-Flow, w pt	**6.37**	**1.20**	**13.83**	**0.10**	**95.41**	**0.20**

Table 6: Comparison of temporal smoothness on GRAB. Acc and jerk are in mm/N^2 and mm/N^3 (lower is better).

Methods	Acc_h	$Jerk_h$	Acc_o	$Jerk_o$
MDM [50]	4.65	3.87	4.12	3.06
MLD [6]	4.31	3.52	4.63	3.44
LatentHOI [26]	3.87	3.03	3.22	2.39
HO-Flow	**3.08**	**2.38**	**2.59**	**1.68**

Table 7: User study with 10 evaluators across three benchmarks. Values indicate the preference ratio (higher is better).

Methods	GRAB	OakInk	DexYCB
MDM [50]	0.16	0.12	0.12
MLD [6]	0.11	0.10	0.13
LatentHOI [26]	0.23	0.22	0.23
HO-Flow (Ours)	**0.50**	**0.56**	**0.52**

physically grounded and diverse hand-object interactions, preserving realistic contact while effectively suppressing unnatural interpenetrations.

OakInk. We further evaluate the generalization of HO-Flow model on the out-of-distribution OakInk benchmark, which features a much broader set of unseen objects than prior datasets. As reported in Table 4, HO-Flow achieves consistently strong performance in both motion quality and diversity. In particular, our method generalizes well to novel object geometries, yielding a clear reduction in hand-object interpenetrations and intersection depths compared with the recent LatentHOI [26]. Moreover, HO-Flow improves physical plausibility (Phy) to 89.76%, outperforming the text-driven Text2HOI [4] and diffusion-based approaches [6,50]. Despite the increased difficulty of interacting with diverse unseen objects, HO-Flow maintains the highest sample diversity (SD), indicating that it captures the underlying distribution of hand-object interactions without sacrificing realism. Notably, in this more challenging setting with more diverse objects, pre-training on GraspXL provides a larger gain in performance, further improving physical fidelity and reducing penetration-related interaction artifacts.

DexYCB. Table 5 further demonstrates the effectiveness of HO-Flow in synthesizing single-hand manipulation sequences on the DexYCB benchmark. Compared to MDM [50] and the state-of-the-art LatentHOI [26], our approach shows a significant reduction in mesh penetration, decreasing the Intersection Depth (ID) from 2.01cm to 1.20cm. Furthermore, HO-Flow achieves a notable leap in physical plausibility (Phy), reaching 95.41% compared to LatentHOI's 88.52%, while simultaneously improving semantic diversity (SD). These results confirm that our model not only excels in bimanual manipulation synthesis but also pos-

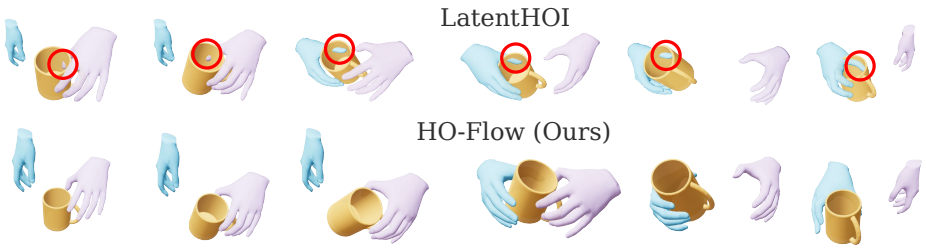

Fig. 5: Qualitative comparison with the state-of-the-art LatentHOI approach [26]. Our HO-Flow model can generate more realistic hand-object interactions with more natural contacts and significantly less penetrations.

sesses a strong generative capability to produce realistic and diverse interaction sequences for single-hand scenarios.

Interactions with articulated objects. We further demonstrate that HO-Flow can be adapted to generate hand interactions with articulated objects on ARCTIC [14]. Following Text2HOI [4], we augment the object branch of Inter-VAE with an additional articulation-angle prediction, so that the generated motion can drive articulated object parts. We follow metrics used in Text2HOI [4] to use top-3 action accuracy (Acc.↑), feature distribution distance (FID↓), diversity (Div.→), text-conditioned multimodality (MMod.↑), and physical realism (Phys.↑). Table 8 shows that HO-Flow outperforms previous methods under different evaluation metrics.

Table 8: Comparison with previous methods on ARCTIC dataset.

Methods	Acc.	FID	Div.	MMod.	Phys.
T2M [16]	0.52	0.36	0.33	0.08	0.01
MDM [50]	0.56	0.30	0.50	0.26	0.70
IMoS [15]	0.82	0.18	0.57	0.27	0.76
Text2HOI [4]	0.92	0.13	0.58	0.32	0.88
HO-Flow	**0.94**	**0.12**	**0.59**	**0.33**	**0.90**

User study. We conduct a user study with 10 human evaluators. For each benchmark, we randomly select 30 test samples and present rendered videos from MDM [50], MLD [6], LatentHOI [26], and HO-Flow as four-choice questions with randomized method order. Evaluators select the best result based on visual quality, interaction realism, and alignment with the input text prompt. As shown in Table 7, our method is consistently preferred across all benchmarks. Since each evaluator rates multiple videos, the resulting judgments may not be fully independent due to evaluator-specific preferences. We therefore interpret the preference ratios as evaluator-level evidence rather than relying on error bars that scale with the raw number of ratings. The large margins in Table 7 nevertheless show a clear preference for our HO-Flow approach.

Qualitative performance. Figure 5 qualitatively compares HO-Flow with the state-of-the-art LatentHOI [26], showing that our model synthesizes more realistic interaction motions with much fewer hand-object penetrations. Figure 6 provides qualitative examples of HO-Flow on objects from OakInk and DexYCB datasets, in addition to those on GRAB shown in Figure 5. We observe that HO-Flow synthesizes natural, realistic hand-object interactions in both bi-manual and single-hand settings, across a wide range of objects and tasks.

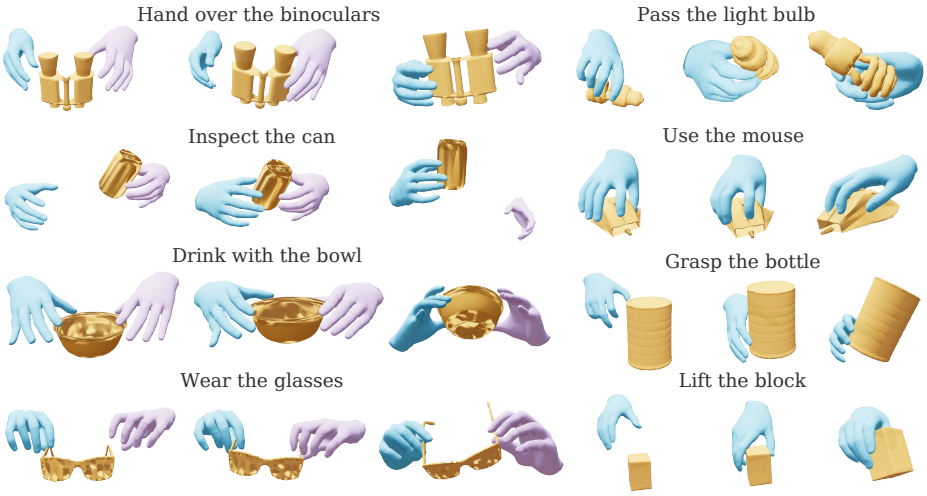

Fig. 6: Qualitative results of HO-Flow on OakInk and DexYCB benchmarks. Our approach can faithfully synthesize hand and object motions with natural interactions.

5 Conclusion

This work presents HO-Flow, a novel framework for generating realistic 3D hand-object interaction motion sequences. It aims to improve physical plausibility, long-horizon temporal coherence, and generalization across diverse actions and objects. HO-Flow combines an expressive interaction-aware motion representation with efficient hand-object interaction synthesis in the continuous space. Inter-VAE encodes short-horizon hand-object interaction sequences into unified latent codes, capturing both global motion trajectories and fine-grained, contact-rich coordination through hand-centric geometric cues. On top of this representation, a masked flow matching model generates motion latent tokens with auto-regressive temporal reasoning, reducing jitter, outliers, and interpenetration. Predicting object motion relative to the initial frame further improves robustness to dataset-specific coordinate conventions and supports large-scale pre-training on synthetic data. Experiments on GRAB, OakInk, and DexYCB show that HO-Flow advances the state of the art in terms of both physical plausibility and motion diversity for text-conditioned hand-object motion generation.

Acknowledgement: Stefanos Zafeiriou and Rolandos Alexandros Potamias was supported by the EPSRC Project GNOMON (EP/X011364/1) and the Turing AI Fellowship (EP/Z534699/1). Jiankang Deng was supported by the NVIDIA Academic Grant. The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "France 2030" program, reference ANR-23-IACL-0008 (PR[AI]RIEPSAI projet), the ANR project 3D-GEM (ANR-25-CE23-7777-01), the ANR project VideoPredict (ANR-21-FAI1-0002-01) and the Paris Île-de-France Region in the frame of the DIM AI4IDF. Cordelia Schmid would like to acknowledge the support by the Körber European Science Prize.

References

1. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: SINC: Spatial composition of 3D human motions for simultaneous action generation. In: ICCV (2023)
2. Blattmann, A., Rombach, R., Oktay, D., Ommer, B.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
3. Bohg, J., Morales, A., Asfour, T., Kragic, D.: Data-driven grasp synthesis—a survey. IEEE TRO (2013)
4. Cha, J., Kim, J., Yoon, J.S., Baek, S.: Text2HOI: Text-guided 3D motion generation for hand-object interaction. In: CVPR (2024)
5. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: DexYCB: A benchmark for capturing hand grasping of objects. In: CVPR (2021)
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)
7. Chen, Z., Chen, S., Arlaud, E., Laptev, I., Schmid, C.: ViViDex: Learning vision-based dexterous manipulation from human videos. In: ICRA (2025)
8. Chen, Z., Chen, S., Schmid, C., Laptev, I.: gSDF: Geometry-driven signed distance functions for 3D hand-object reconstruction. In: CVPR (2023)
9. Chen, Z., Hasson, Y., Schmid, C., Laptev, I.: AlignSDF: Pose-aligned signed distance fields for hand-object reconstruction. In: ECCV (2022)
10. Chen, Z., Potamias, R.A., Chen, S., Schmid, C.: HORT: Monocular hand-held objects reconstruction with transformers. In: ICCV (2025)
11. Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauser, E., Ma, S., Tekin, B.: DiffH2O: Diffusion-based synthesis of hand-object interactions from textual descriptions. In: SIGGRAPH Asia (2024)
12. Dasari, S., Gupta, A., Kumar, V.: Learning dexterous manipulation from exemplar object trajectories and pre-grasps. In: ICRA (2023)
13. Du, Y., Weinzaepfel, P., Lepetit, V., Brégier, R.: Multi-finger grasping like humans. In: IROS (2022)
14. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: CVPR (2023)
15. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: IMoS: intent-driven full-body motion synthesis for human-object interactions. In: Computer Graphics Forum (2023)
16. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3D human motions from text. In: CVPR (2022)
17. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Video diffusion models. In: NeurIPS (2022)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv:2207.12598 (2022)
20. Huang, M., Chu, F.J., Tekin, B., Liang, K.J., Ma, H., Wang, W., Chen, X., Gleize, P., Xue, H., Lyu, S., et al.: HOIGPT: Learning long-sequence hand-object interaction with language models. In: CVPR (2025)
21. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: ICCV (2021)
22. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022)

23. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 3DV (2020)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
25. Li, K., Wang, J., Yang, L., Lu, C., Dai, B.: SemGrasp: Semantic grasp generation via language aligned discretization. In: ECCV (2024)
26. Li, M., Christen, S., Wan, C., Cai, Y., Liao, R., Sigal, L., Ma, S.: LatentHOI: On the generalizable hand object motion generation with latent hand diffusion. In: CVPR (2025)
27. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
28. Liu, T., Liu, Z., Jiao, Z., Zhu, Y., Zhu, S.C.: Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. IEEE RA-L (2021)
29. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
31. Lu, J., Kang, H., Li, H., Liu, B., Yang, Y., Huang, Q., Hua, G.: UGG: Unified generative grasping. In: ECCV (2024)
32. Luo, Z., Cao, J., Christen, S., Winkler, A., Kitani, K.M., Xu, W.: OmniGrasp: Grasping diverse objects with simulated humanoids. In: NeurIPS (2024)
33. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV (2024)
34. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In: CVPR (2025)
35. Miller, A.T., Allen, P.K.: Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine (2004)
36. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
37. Paliwal, B., Etukuru, H., Liang, W., Abbeel, P., Shafiullah, N.M.M., Malik, J.: Do as i do: Dexterous manipulation data from everyday human videos. arXiv:2606.19333 (2026)
38. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
39. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: ECCV (2022)
40. Prokudin, S., Lassner, C., Romero, J.: Efficient learning on point clouds with basis point sets. In: ICCV (2019)
41. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS (2017)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
44. Romero, J., Tzionas, D., Black, M.J.: Embodied Hands: Modeling and capturing hands and bodies together. TOG (2017)
45. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:1707.06347 (2017)

46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
47. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
48. Sutton, R.S.: Reinforcement learning: An introduction. A Bradford Book (2018)
49. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: ECCV (2020)
50. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
51. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017)
52. Wan, W., Geng, H., Liu, Y., Shan, Z., Yang, Y., Yi, L., Wang, H.: UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In: ICCV (2023)
53. Wei, Y.L., Jiang, J.J., Xing, C., Tan, X.T., Wu, X.M., Li, H., Cutkosky, M., Zheng, W.S.: Grasp as you say: Language-guided dexterous grasp generation. In: NeurIPS (2024)
54. Wei, Y.L., Lin, M., Lin, Y., Jiang, J.J., Wu, X.M., Zeng, L.A., Zheng, W.S.: AffordDexGrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. In: ICCV (2025)
55. Wu, Z., Potamias, R.A., Zhang, X., Zhang, Z., Deng, J., Luo, S.: CEDex: Cross-embodiment dexterous grasp generation at scale from human-like contact representations. In: ICRA (2026)
56. Xu, G.H., Wei, Y.L., Zheng, D., Wu, X.M., Zheng, W.S.: Dexterous grasp transformer. In: CVPR (2024)
57. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., et al.: UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In: CVPR (2023)
58. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: OakInk: A large-scale knowledge repository for understanding hand-object interaction. In: CVPR (2022)
59. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: CPF: Learning a contact potential field to model the hand-object interaction. In: ICCV (2021)
60. Ye, Q., Li, H., Liu, Q., Jiang, S., Zhou, T., Huo, Y., Chen, J.: Contact2motion: Contact guided dexterous grasp motion generation with synergy embedded optimization. IJRR (2025)
61. Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: CVPR (2023)
62. Zhang, H., Christen, S., Fan, Z., Hilliges, O., Song, J.: GraspXL: Generating grasping motions for diverse objects at scale. In: ECCV (2024)
63. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2M-GPT: Generating human motion from textual descriptions with discrete representations. In: CVPR (2023)
64. Zhang, M., Li, Z., Wang, Q., Shi, J., Li, Y., Tan, P.: MotionDiffuse: Text-driven human motion generation with diffusion model. arXiv:2208.15001 (2022)
65. Zhang, W., Dabral, R., Golyanik, V., Choutas, V., Alvarado, E., Beeler, T., Habermann, M., Theobalt, C.: BimArt: A unified approach for the synthesis of 3D bimanual interaction with articulated objects. In: CVPR (2025)
66. Zhang, Z., Wang, H., Yu, Z., Cheng, Y., Yao, A., Chang, H.J.: NL2Contact: Natural language guided 3D hand-object contact modeling with diffusion model. In: ECCV (2024)

67. Zhong, Y., Jiang, Q., Yu, J., Ma, Y.: DexGrasp Anything: Towards universal robotic dexterous grasping with physics awareness. In: CVPR (2025)
68. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)
69. Zhu, H., Zhao, T., Ni, X., Wang, J., Fang, K., Righetti, L., Pang, T.: Should we learn contact-rich manipulation policies from sampling-based planners? IEEE RA-L (2025)
70. Zuo, R., Potamias, R.A., Ververas, E., Deng, J., Zafeiriou, S.: Signs as tokens: A retrieval-enhanced multilingual sign language generator. In: ICCV (2025)