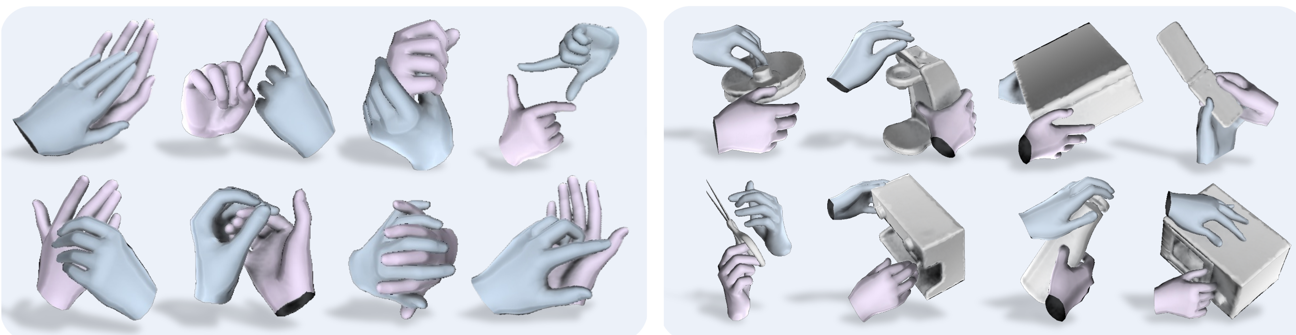
InterHandGen: Two-Hand Interaction Generation via Cascaded Reverse Diffusion

Jihyun Lee[1] Shunsuke Saito[2] Giljoo Nam[2] Minhyuk Sung[1] Tae-Kyun Kim[1,3]
[1] KAIST [2] Codec Avatars Lab, Meta [3] Imperial College London
https://jyunlee.github.io/projects/interhandgen



(a) Generated two-hand interactions. (b) Generated two-hand-object interactions.

Figure 1. **Two-hand synthesis with InterHandGen.** We propose InterHandGen, an approach to generate two-hand interactions with or without an object using a novel cascaded diffusion. To enable high-fidelity and diverse sampling, we decompose the modeling of joint distribution into the modeling of factored unconditional and conditional single-hand distributions.

Abstract

We present InterHandGen, a novel framework that learns the generative prior of two-hand interaction. Sampling from our model yields plausible and diverse two-hand shapes in close interaction with or without an object. Our prior can be incorporated into any optimization or learning methods to reduce ambiguity in an ill-posed setup. Our key observation is that directly modeling the joint distribution of multiple instances imposes high learning complexity due to its combinatorial nature. Thus, we propose to decompose the modeling of joint distribution into the modeling of factored unconditional and conditional ***single*** *instance distribution. In particular, we introduce a diffusion model that learns the single-hand distribution unconditional and conditional to another hand via conditioning dropout. For sampling, we combine anti-penetration and classifier-free guidance to enable plausible generation. Furthermore, we establish the rigorous evaluation protocol of two-hand synthesis, where our method significantly outperforms baseline generative models in terms of plausibility and diversity. We also demonstrate that our diffusion prior can boost the performance of two-hand reconstruction from monocular in-the-wild images, achieving new state-of-the-art accuracy.*

1. Introduction

Two-hand interaction is widely involved in our daily lives. We coordinate our hands closely together when clasping, praying, stretching, or engaging in social interactions. Modeling and understanding two-hand interactions are thus crucial for applications that require capturing human behaviors, such as augmented or virtual reality (AR/VR) and human-computer interaction (HCI). Highlighting this importance, numerous research endeavors have been dedicated to interacting hands reconstruction. With the release of the large-scale two-hand interaction dataset [41], various methods [23, 31, 32, 34, 40, 41, 53, 55, 69, 72] have been proposed mainly for monocular two-hand reconstruction.

The under-explored part in the current two-hand interaction literature is interacting two-hand *generation*. Although there are generative models proposed for other human interaction domains (e.g., hand-object [10, 22, 24, 26, 27, 64] or two-human [35, 43, 58] interaction), directly adapting them for two-hand interaction leads to sub-optimal generations. Compared to hand-object interaction that involves a rigid object, two hands lead to significantly more complex interactions due to the higher degree of freedom in two articulated hands. Additionally, while human-to-human body interaction is typically constrained on a shared ground plane, each joint of two hands has a full 6 DOF to allow more diverse interactions. Motivated by the advancement of unconstrained pose estimation leveraging a strong prior in other domains [43, 46], our goal is to build a highly expressive generative prior for two-hand interaction, which can be effortlessly incorporated into existing learning and optimization frameworks.

In this paper, we introduce InterHandGen, a framework that effectively learns the generative prior of two-hand interaction. The important challenge in two-hand interaction generation lies in its high data complexity caused by the combination of hand articulations. To reduce the complexity of learning such generation target, we propose to reformulate the two-hand distribution modeling into the modeling of single-hand model distribution unconditional and conditional to the other hand, such that:

$$p_\phi(\mathbf{x}_l, \mathbf{x}_r) = p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r | \mathbf{x}_l), \tag{1}$$

where $p_\phi(\cdot)$ is the model distribution, and $\mathbf{x}_l$ and $\mathbf{x}_r$ are left and right hand shapes in interaction, respectively. By leveraging the symmetric nature of the left and right hands, we jointly learn both $p_\phi(\mathbf{x}_l)$ and $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$ in the shared hand parameter domain based on MANO [54] model. In particular, we take a diffusion-based approach [21, 60] and train a single denoising diffusion model via conditioning dropout [20] to model both types of single-hand distribution. This way, the degree of freedom of each generation process is effectively reduced. Importantly, this formulation can be easily extended to two-hand and object interaction generation, by simply adding an object conditioning $\mathbf{c}$ to each of the terms in Equation 1.

In inference time, we sample one hand using the learned model $p_\phi(\mathbf{x}_l)$ and the other hand conditioned on the previously sampled hand using $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$ in a cascaded manner. For conditional sampling, we use classifier-free guidance [20] to achieve a better balance between fidelity and diversity. To avoid sampling a physically implausible state due to penetration, we also introduce anti-penetration guidance that penalizes inter-penetration during the reverse diffusion process. Furthermore, we show how to incorporate the learned two-hand interaction prior into any optimization or learning methods for reducing ambiguity in an ill-posed setup, inspired by Score Distillation Sampling [48] and BUDDI [43].

As there is no established benchmark for two-hand generation, we introduce a new evaluation protocol of two-hand interaction synthesis. In particular, we extend the standard metrics used for generative modeling (e.g., FID [19], KID [7], Diversity [47, 51, 62]) to two-hand interaction by training a tailored feature backbone network. Our experiments show that our approach significantly outperforms the baseline methods on two-hand interaction generation with or without an object. We also show that our diffusion prior is useful for the downstream task of interacting two-hand reconstruction from in-the-wild images, where we set *new state-of-the-art*.

Our main contributions are summarized as follows:

- We propose an effective learning framework to build a generative prior of two-hand interaction. Our cascaded reverse diffusion approach shows significant improvement over baselines in terms of fidelity and diversity.
- Our formulation is general and can be extended to more instances. We show that our approach also achieves superior performance on two-hand interaction with objects.
- Our approach is a drop-in replacement for regularization in optimization or learning problems. By incorporating our prior, we achieve the state-of-the-art performance on interacting two-hand pose estimation from in-the-wild images.
- We provide a comprehensive analysis of two-hand generation with a newly established evaluation protocol. Our code and backbone network weights are publicly available for benchmarking future research.

2. Related Work

In this section, we discuss the related work on interacting two-hand reconstruction, and hand-object and two-human interaction generation. Note that the background on diffusion models can be found in Section 3.1.

Interacting two-hand reconstruction. Various methods have been proposed for interacting two-hand reconstruction from monocular RGB [23, 31, 32, 34, 40, 41, 53, 55, 69, 72], multi-view RGB [4], or depth [42, 45, 61]. To address self-similarity, self-occlusion, and complex articulations of interacting hands, the recent methods mainly exploit attention mechanism [34, 41, 53, 68, 72] and/or interaction-aware shape refinement [32, 53, 55, 69]. Recently, Zuo *et al*. [72] (which is concurrent work to ours) proposes to use a variational autoencoder (VAE) [28] as a prior for monocular two-hand reconstruction. While their approach is specialized for monocular image-based reconstruction using a specific network architecture, our approach can be used for any optimization and learning tasks. In addition, our experiments (Section 4) show that our diffusion-based prior significantly outperforms the vanilla VAE used in [72] for a generation task in all metrics.

Hand-object interaction generation. Most of the methods mainly focus on generating single-hand shapes conditioned on an object [3, 10, 15, 22, 24, 26, 27, 64]. As the existing single-hand and object interaction datasets [2, 9, 14, 17, 18, 30, 38] are mostly limited to grasping [13], the state-of-the-art generation methods actively leverage contact prior [16, 24, 37, 64] or physics simulators [22, 64] to synthesize grasps that cannot be easily broken by applying external force [10, 22, 24]. However, in two-hand interaction, each hand can arbitrarily move by itself, so physical contact between hands does not necessarily occur. Thus, it is non-trivial to directly adapt the existing methods that heavily rely on physical priors. In addition, we consider the recent benchmark (ARCTIC [13]) on *two-hand* and object interaction that captures various bimanual scenarios (e.g., opening a box, operating an espresso machine). Since ARCTIC is also not limited to dense contacts between object and both hands (e.g., grasping), our general approach outperforms the most recent method on single-hand and object interaction synthesis (ContactGen [37]) extended for two-hand and object interaction generation on ARCTIC dataset.

Two-human interaction synthesis. More recently, a few methods for two-human interaction synthesis have been proposed. PriorMDM [58] and InterGen [35] introduce diffusion models for text-driven two-human motion generation. BUDDI [43] (which is concurrent work to ours) proposes an unconditional generation method of interacting two-human shapes. It introduces a transformer-based diffusion model to generate SMPL [39] parameters of two humans *jointly*. In our work, we discover that directly modeling the joint distribution of two hands leads to sub-optimal generation performance due to the high data complexity. Instead, we simplify the learning process by decomposing the joint distribution into conditional and unconditional single-hand distributions and experimentally show that ours yields *significantly* better generation results than BUDDI modified to synthesize two-hand interactions.

3. Method

3.1. Preliminary

Diffusion Models. Diffusion models (e.g., [21, 60]) are a class of generative models that learn to recurrently transform noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into a sample from the target data distribution $\mathbf{z}_0 \sim q(\mathbf{z}_0)$. This denoising process is called the reverse process and can be expressed as:

$$p_\phi(\mathbf{z}_{0:T}) := p(\mathbf{z}_T) \prod_{t=1}^{T} p_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t), \tag{2}$$

where p_ϕ is a model distribution parameterized by ϕ and $\mathbf{z}_1, ..., \mathbf{z}_T$ are latent variables of the same dimensionality as $\mathbf{z}_0$. Conversely, the forward process models $q(\mathbf{z}_{1:T}|\mathbf{z}_0)$ by gradually adding Gaussian noise to the data sample $\mathbf{z}_0$. In this process, the intermediate noisy sample $\mathbf{z}_t$ can be sampled as:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon \tag{3}$$

in variance-preserving diffusion formulation [21]. Here, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a noise variable and $\alpha_{1:T} \in (0, 1]^{\mathrm{T}}$ is a sequence that controls the amount of noise added at each diffusion time t. Given the noisy sample $\mathbf{z}_t$ and t, the diffusion model f_ϕ learns to approximate the reverse process for data generation. The diffusion model parameters ϕ are typically optimized to minimize $\mathbb{E}_{\mathbf{z}_t,\epsilon} \|\epsilon - f_\phi(\mathbf{z}_t,\, t)\|^2$ [21] or $\mathbb{E}_{\mathbf{z}_t,\epsilon} \|\mathbf{z}_0 - f_\phi(\mathbf{z}_t,\, t)\|^2$ [58, 62]. Note that exact formulations vary across the literature, and we kindly refer the reader to the survey papers [11, 67] for a more comprehensive review of diffusion models.

Classifier-Free Guidance (CFG) [20]. CFG is a method proposed to achieve a better trade-off between fidelity and diversity for conditional sampling using diffusion models. Instead of generating a sample using conditional score estimates only, it proposes to mix the conditional and unconditional score estimates to control a trade-off between sample fidelity and diversity:

$$\tilde{f}_\phi(\mathbf{z}_t, t, \mathbf{c}) = (1+w) f_\phi(\mathbf{z}_t, t, \mathbf{c}) - w f_\phi(\mathbf{z}_t, t, \emptyset), \tag{4}$$

where $\mathbf{c}$ is conditioning information and w is a hyperparameter that controls the strength of the guidance. However, Equation 4 requires training both conditional and unconditional diffusion models. To address this, Ho *et al.* [20] introduces conditioning dropout during training, which enables the parameterization of both conditional and unconditional models using a single diffusion network. Conditioning dropout simply sets $\mathbf{c}$ to a null token $\emptyset$ with a chosen probability p_{uncond} to jointly learn the conditional and unconditional scores during network training. Due to its ability to achieve a better balance between fidelity and diversity, CFG is used in many state-of-the-art conditional diffusion models [8, 29, 44, 48, 56, 58, 62].

3.2. Problem Definition and Key Formulation

Our goal is to learn a distribution of 3D interacting two-hand shapes $p_\phi(\mathbf{x}_l, \mathbf{x}_r)$ from the samples from a two-hand data distribution $q(\mathbf{x}_l, \mathbf{x}_r)$. We assume a situation where one left hand $\mathbf{x}_l$ and one right hand $\mathbf{x}_r$ are interacting with each other, following the existing two-hand interaction benchmark [41]. For representing each hand, we use MANO [54] model which is a differentiable statistical model that maps a pose parameter $\theta \in \mathbb{R}^{45}$ and a shape parameter $\beta \in \mathbb{R}^{10}$ to a hand mesh with 3D vertices $\mathbf{V} \in \mathbb{R}^{778\times 3}$ and triangular faces $\mathbf{F} \in \mathbb{R}^{1554\times 3}$. Based on MANO, we parameterize each hand shape as:

$$\mathbf{x}_s = [\theta_s,\, \beta_s,\, \omega_s,\, \tau_s], \tag{5}$$

where $\mathbf{x}_s \in \mathbb{R}^{64}$ represents a 3D hand shape of side $s = \{l, r\}$, and θ_s and β_s are the corresponding MANO pose and shape parameters. $\omega_s \in \mathbb{R}^6$ denotes the root rotation in 6D rotation representation [71], and $\tau_s \in \mathbb{R}^3$ denotes the root translation.

To learn the distribution $p_\phi(\mathbf{x}_l, \mathbf{x}_r)$ that captures plausible two-hand interaction states, one straightforward approach would be to directly model $p_\phi(\mathbf{x}_l, \mathbf{x}_r)$ using a single generative network. However, we observe that the direct learning of joint two-hand distribution leads to suboptimal results, as the target distribution involves highly articulated hand shapes in close interaction, and its combinatorial nature imposes high generation complexity. To address this, our key idea is to decompose the joint two-hand distribution to model the unconditional and conditional single-hand distribution instead, such that:

$$p_\phi(\mathbf{x}_l, \mathbf{x}_r) = p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r|\mathbf{x}_l). \tag{6}$$

Note that the joint distribution of two hands can now be represented by the distribution of a single hand on one side $p_\phi(\mathbf{x}_l)$ and that on the other side $p_\phi(\mathbf{x}_r)$ conditioned on $\mathbf{x}_l$. By decomposing the problem of learning $p_\phi(\mathbf{x}_l, \mathbf{x}_r)$ into two sub-problems of learning unconditional and conditional single-hand distributions, we can effectively reduce the degree of freedom of each generation target. This formulation is general, and can be easily extended to two-hand and object interaction generation, by simply adding an object conditioning $\mathbf{c}$ to each of the terms in Equation 6. In what follows, we explain our novel parameterization of $p_\phi(\mathbf{x}_l)$ and $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$ using diffusion models [21, 60]. Later in the experiments (Section 4), we also show that this simple decomposition leads to significant performance improvement in interacting two-hand generation with or without an object.

3.3. Training

For learning $p_\phi(\mathbf{x}_l)$ and $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$ in Equation 6, one straightforward approach is to separately train unconditional and conditional diffusion networks. However, there is conceptual redundancy embedded in $p_\phi(\mathbf{x}_l)$ and $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$. Both distributions ultimately capture the plausible single-hand shapes, where the differences lie in (1) the side of the hand and (2) whether the distribution is unconditional or conditional. Motivated by multi-task learning [6, 63, 70] that has shown that joint learning of related tasks improves both learning efficiency and accuracy by exploiting the commonalities across tasks, we also introduce a training mechanism that can jointly learn $p_\phi(\mathbf{x}_l)$ and $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$ using a single diffusion network.

Regarding the difference in the side of hand, we pay attention to the observation that shape symmetry exists between left and right hands. The existing MANO model [54] indeed learns a unified hand model in the right-hand space, where the left-hand model is obtained by horizontally flipping the model shape space. Following MANO, we also bring all single-hand generation targets into the shared domain. Since our hand representation is already based on MANO, we follow the same mirroring transformation Γ used in MANO [54] (please refer to the supplementary for details) to map the left-hand generation targets into the shared right-hand MANO parameter space for network training. In particular, our training objective can be written as:

- Learning $p_\phi(\mathbf{x}_r)$ from training samples of $\mathbf{x}_r$ and $\Gamma(\mathbf{x}_l)$;
- Learning $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$ from training samples of $(\mathbf{x}_r, \mathbf{x}_l)$ and $(\Gamma(\mathbf{x}_l), \Gamma(\mathbf{x}_r))$.

This further augments the training data and improves generalization. More importantly, once we normalize the hand side, our training objective becomes learning the unconditional and conditional distributions in the same right-hand MANO parameter space ($p_\phi(\mathbf{x}_r)$ and $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$), rather than learning one unconditional distribution and one conditional distribution in the different hand spaces ($p_\phi(\mathbf{x}_l)$ and $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$). Our new learning objective is now in the form that conditioning dropout [20] (Section 3.1) can be directly applied to parameterize both unconditional and conditional models using a single diffusion network.

Let our diffusion network be D_ϕ that takes a noisy hand parameter $\mathbf{x}_t$, a conditioning hand parameter $\mathbf{x}_l$ and diffusion time t. As shown in Algorithm 1, we can train D_ϕ to enable both conditional hand generation (by taking the other hand parameter $\mathbf{x}_l$ as conditioning input) and unconditional hand generation (by taking $\emptyset$ as conditioning input) via conditioning dropout [20] (Step 3 in Algorithm 1). Later in the experiments (Section 4), we show that training a unified diffusion network for $p_\phi(\mathbf{x}_r)$ and $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$ leads to better generation results than training two separate networks.

Algorithm 1 Training via conditioning hand dropout.

Require: p_{uncond}: probability for conditioning dropout
Require: $\alpha_{1:T}$: diffusion noise scheduling
1: **repeat**
2: Sample $(\mathbf{x}_r, \mathbf{x}_l)$ from $q(\mathbf{x}_r, \mathbf{x}_l)$ or $q(\Gamma(\mathbf{x}_l), \Gamma(\mathbf{x}_r))$
3: $\mathbf{x}_l \leftarrow \emptyset$ with probability p_{uncond}
4: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
▷ *Compute diffused data at time t (Equation 3)*
5: $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_r + \sqrt{1-\alpha_t}\epsilon$
6: Take a gradient step on $\nabla_\phi \left\|\mathbf{x}_r - D_\phi(\mathbf{x}_t, \mathbf{x}_l, t)\right\|^2$
7: **until** converged

3.4. Inference: Cascaded Reverse Diffusion

After training our diffusion network, we can first sample an anchor left-hand $\mathbf{x}_l$ from the learned $p_\phi(\mathbf{x}_r)$ after flipping the model space by Γ [54]. Then, we can sample an interacting right-hand conditioned on the anchor hand $\mathbf{x}_l$ from $p_\phi(\mathbf{x}_r|\mathbf{x}_l)$ in the form of cascaded inference. Our overall inference procedure is described in Algorithm 2. Note that $\mathcal{E}(\cdot)$ denotes a function that computes the added noise ϵ from the diffusion model prediction [58, 62]. In Algorithm 2, we incorporate two types of guidance into the reverse process: (1) classifier-free guidance (CFG) [20] to control a trade-off between fidelity and diversity in conditional sampling (Step 11 in Algorithm 2) and (2) anti-penetration guidance to avoid inter-hand penetration (Step 13 in Algorithm 2). As CFG is already discussed in Section 3.1, we describe our anti-penetration guidance below.

Algorithm 2 Inference via cascaded hand denoising.

Require: w_{cfg}: classifier-free guidance strength
Require: w_{pen}: anti-penetration guidance strength
Require: $\mathcal{L}_{pen}$: penetration loss function
▷ *Sample anchor hand* $\mathbf{x}_l$
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for all** t from T to 1 **do**
3: $\quad \hat{\epsilon} \leftarrow \mathcal{E}(D_\phi(\mathbf{x}_t, \emptyset, t))$
$\quad$ ▷ *DDIM [60] sampling*
4: $\quad \mathbf{x}_{t-1} \leftarrow \sqrt{\alpha_{t-1}}(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\hat{\epsilon}}{\sqrt{\alpha_t}}) + \sqrt{1-\alpha_{t-1}}\hat{\epsilon}$
5: **end for**
6: $\mathbf{x}_l \leftarrow \Gamma(\mathbf{x}_0)$
▷ *Sample interacting hand* $\mathbf{x}_r$ *given anchor hand* $\mathbf{x}_l$
7: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
8: **for all** t from T to 1 **do**
9: $\quad \hat{\epsilon}_{uncond} \leftarrow \mathcal{E}(D_\phi(\mathbf{x}_t, \emptyset, t))$
10: $\quad \hat{\epsilon}_{cond} \leftarrow \mathcal{E}(D_\phi(\mathbf{x}_t, \mathbf{x}_l, t))$
$\quad$ ▷ *Classifier-free guidance [20]*
11: $\quad \hat{\epsilon} \leftarrow (1 + w_{cfg})\hat{\epsilon}_{cond} - w_{cfg}\hat{\epsilon}_{uncond}$
$\quad$ ▷ *DDIM [60] sampling*
12: $\quad \mathbf{x}_{t-1} \leftarrow \sqrt{\alpha_{t-1}}(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\hat{\epsilon}}{\sqrt{\alpha_t}}) + \sqrt{1-\alpha_{t-1}}\hat{\epsilon}$
$\quad$ ▷ *Anti-penetration guidance (Section 3.4)*
13: $\quad \mathbf{x}_{t-1} \leftarrow \mathbf{x}_{t-1} - w_{pen}\nabla_{\mathbf{x}_{t-1}}\mathcal{L}_{pen}(\mathbf{x}_{t-1}, \mathbf{x}_l)$
14: **end for**
15: $\mathbf{x}_r \leftarrow \mathbf{x}_0$

Anti-penetration guidance. Inspired by the existing work on diffusion guidance on image domain [5, 12, 33], we introduce test-time guidance to avoid penetration between the generated two hands. In particular, we move the current interacting hand generation $\mathbf{x}_{t-1}$ towards the negative gradient direction of the penetration loss function $\mathcal{L}_{pen}$ at each denoising step (Step 13 in Algorithm 2). Let $\mathbf{V}_{t-1}, \mathbf{V}_l \in \mathbb{R}^{778\times3}$ denote mesh vertices recovered from the noisy right-hand parameter $\mathbf{x}_{t-1}$ and the conditional left-hand parameter $\mathbf{x}_l$ using MANO [54] layer. In particular, we recover these vertices from clean hand parameter estimated from $t-1$ via DDIM [60] sampling $\frac{\mathbf{x}_{t-1} - \sqrt{1-\alpha_{t-1}}\hat{\epsilon}}{\sqrt{\alpha_{t-1}}}$ to enable more robust loss computation [5, 33]. Then, our penetration loss $\mathcal{L}_{pen}$ is defined as:

$$\mathcal{L}_{pen}(\mathbf{x}_{t-1}, \mathbf{x}_l) = \sum_{i,\,j\,\in\,\mathcal{P}(\mathbf{x}_{t-1},\,\mathbf{x}_l)} ||\mathbf{V}_{t-1}^i - \mathbf{V}_l^j||_2, \quad (7)$$

which is the sum of squared distances between the penetrated vertex $\mathbf{V}_{t-1}^i$ in one hand and its nearest vertex $\mathbf{V}_l^j$ in the other hand. Here, $\mathcal{P}$ denotes a function that returns a set of penetrated vertex indices (i, j) and is defined as:

$$\mathcal{P}(\mathbf{x}_{t-1}, \mathbf{x}_l) = \{(i,\, j) \,|\, -\mathbf{n}_j^{\mathrm{T}} \cdot (\mathbf{V}_{t-1}^i - \mathbf{V}_l^j) > 0\}, \quad (8)$$

where j denotes the vertex index of $\mathbf{V}_l$ that is nearest to $\mathbf{V}_{t-1}^i$, and $\mathbf{n}_j$ is a normal vector at $\mathbf{V}_l^j$. This way, the amount of penetration can be approximated by projecting a vector joining the nearest vertices from the two hands onto the normal vector at the anchor hand, similar to the existing hand-object reconstruction literature [17].

3.5. Generative Prior for Two-Hand Problems

We now explain how our two-hand interaction prior can be easily incorporated into any optimization or learning methods to further boost the accuracy of the downstream problems, such as monocular two-hand reconstruction. Inspired by Score Distillation Sampling (SDS) [48] and BUDDI [43], we treat our pre-trained two-hand diffusion model D_ϕ as a frozen critic that regularizes the current two-hand interaction state $(\mathbf{x}_l, \mathbf{x}_r)$ (e.g., predicted by a reconstruction network) to move to a higher-density region. Our diffusion-based regularization term can be written as:

$$\mathcal{L}_{reg} = ||\, \mathcal{S}\,(D_\phi, \mathbf{x}_l, \mathbf{x}_r) - (\mathbf{x}_l,\, \mathbf{x}_r)\,||_2, \quad (9)$$

where $S(\cdot,\cdot,\cdot)$ denotes a function that performs a single forward-reverse diffusion step [43] that takes as input the current two-hand interaction state $(\mathbf{x}_l, \mathbf{x}_r)$ and outputs the denoised interaction $(\hat{\mathbf{x}}_l, \hat{\mathbf{x}}_r)$ estimated by D_ϕ. Note that we detach the gradients of the diffusion model D_ϕ following [43, 48]. $\mathcal{L}_{reg}$ can be incorporated as an additional regularizer into any loss function during network training or shape optimization in a plug-and-play manner.

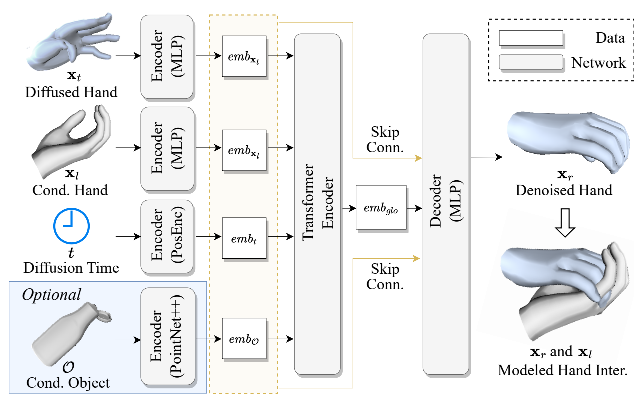


Figure 2. **Our network architecture.** We use self-attention between the embeddings of the inputs (i.e., $\mathbf{x}_t$ $\mathbf{x}_l$, t, and optional $\mathcal{O}$) to estimate the denoised hand parameter $\mathbf{x}_r$.

3.6. Network Architecture

We use a transformer-based architecture for our diffusion model D_ϕ. As shown in Figure 2, we first use two fully connected layers with Swish activation [52] to embed the input hand and conditioning hand parameters (i.e., $\mathbf{x}_t$, $\mathbf{x}_l$). We also embed the diffusion time t using Positional Encoding [21]. Then, we use four-headed self-attention [66] to model the relationship between the input embeddings. Lastly, the updated input embeddings are flattened and fed to eight fully connected layers with ReLU [1] activation and skip connections to estimate the clean hand signal $\mathbf{x}_r$.

Object-conditional generation. To enable two-hand generation conditioned on an object, we can add a global object conditioning $\mathbf{c}$ to model $p_\phi(\mathbf{x}_l, \mathbf{x}_r|\mathbf{c}) = p_\phi(\mathbf{x}_l|\mathbf{c})\, p_\phi(\mathbf{x}_r|\mathbf{x}_l, \mathbf{c})$. To incorporate the object conditioning $\mathbf{c}$, we simply add a PointNet++ [50]-based embedding branch (blue box in Figure 2) for an input object point cloud O. Please refer to the supplementary for more details on our architecture (e.g., layer configurations).

4. Experiments

4.1. Two-Hand Interaction Synthesis

Data. We use InterHand2.6M [41] dataset, which is the most widely used interacting two-hand dataset. Following the existing reconstruction work [32, 34], we use interacting hand (*IH*) samples with *valid* annotation. The resulting dataset consists of 366K training samples, 110K validation samples, and 261K test samples.

Baselines. We first consider VAE used as a two-hand prior for monocular reconstruction in [72]. We also consider BUDDI [43], which is a recently proposed diffusion model that *jointly* generates two human parameters. We modify BUDDI to generate interacting two-hand parameters and denote the resulting model by BUDDI*. We additionally consider our method variations in which the modeling of joint distribution is not decomposed (*Ours w/o Decomposition*) or separate conditional and unconditional networks are trained to model the decomposed single-hand distributions (*Ours w/o Shared Network*). Please refer to the supplementary for the details of the baselines.

Evaluation metrics. As there is no established benchmark for 3D two-hand interaction generation, we build our own evaluation protocol. Following the existing work on human pose and motion generation [51, 62], we extend Fréchet Inception Distance (FID) [19], Kernel Inception Distance (KID) [7], diversity [51, 62] and precision-recall [57] for evaluating the generated two-hand interactions. We also report the mean inter-penetration volume in cm^3 to measure the physical plausibility. Note that FID, KID, and precision-recall are originally proposed for evaluating the feature discrepancy between the generated and the ground truth image distributions. However, there is no pre-trained feature extraction backbone for interacting two-hand shapes unlike in the image [19] or human motion [51, 62] domain. To address this, we train a backbone network to extract 3D two-hand interaction features, whose network weights will be released for benchmarking future research. Inspired by FPD [59] that measures Fréchet distance of the generated 3D objects (e.g., chair, airplane) on PointNet [49] feature space, we train PointNet++ [50] to regress two hand poses in axis-angle representation and their relative root transformation from a 3D two-hand shape represented as a point cloud. Note that, while it is possible to extract two-hand features by specifically leveraging MANO [54] parameter or mesh structure, we aim to propose a more general metric for future work on two-hand interaction generation, that may not be directly reliant on the MANO model. We rename our two-hand-specific metric for FID and KID as Fréchet Hand Interaction Distance (FHID) and Kernel Hand Interaction Distance (KHID), respectively.

Results. As shown in Table 1a, our method significantly outperforms the baselines on most of the metrics. Especially, learning the decomposed two-hand distribution (*rows 5-6*) leads to noticeable performance improvement. While *Ours w/o Shared Network* (*rows 5*) achieves the best precision score, our final method (*rows 6*) achieves significantly better scores on the other metrics. We also notice that ours achieves high scores on both precision and recall with a good balance, while most of the baselines yield a high score on either one of them. Figure 3 qualitatively shows the sampled two-hand interactions using our method, which further demonstrates that our prior captures plausible and diverse two-hand interactions.

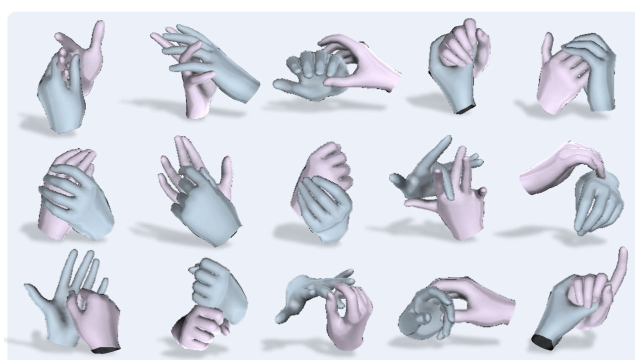


Figure 3. **Two-hand interactions synthesized by InterHandGen.** The sampled interactions are plausible and diverse.

4.2. Object-Conditioned Two-Hand Synthesis

Data. We use the recently released ARCTIC [13] dataset. Unlike the existing hand-object datasets [9, 17, 18, 30, 38] that are mostly limited to single-hand grasps, ARCTIC captures diverse two-hand and object interaction scenarios,

Table 1. **Quantitative comparisons of two-hand interaction synthesis with and without an object. Bold** indicates the best scores, and underline indicates the second best scores. In both experiments, ours significantly outperforms the baselines on most of the metrics. We conduct 20 evaluations and report the average scores, where 10K samples are used in two-hand synthesis and 30K samples (3K samples per object category) are used for two-hand-object synthesis in each evaluation.

(a) **Comparisons on two-hand interaction generation (Section 4.1).**

Method	FHID ↓	KHID $(\times 10^{-2})$ ↓	Diversity ↑	Precision ↑	Recall ↑	PenVol (mm^3) ↓
VAE [72]	8.18	6.23	2.32	0.55	0.02	7.32
BUDDI* [43]	3.48	4.10	<u>2.71</u>	0.56	<u>0.47</u>	<u>0.82</u>
Ours w/o Decomposition	2.09	0.75	2.34	<u>0.86</u>	0.35	3.10
Ours w/o Shared Network	<u>1.32</u>	<u>0.46</u>	2.46	**0.92**	0.42	3.95
Ours	**1.00**	**0.15**	**3.59**	<u>0.86</u>	**0.85**	**0.76**

(b) **Comparisons on object-conditioned two-hand interaction generation (Section 4.2).**

Method	FHID ↓	KHID $(\times 10^{-1})$ ↓	Diversity ↑	Precision ↑	Recall ↑	PenVol (mm^3) ↓
ContactGen* [37]	22.56	1.58	<u>6.70</u>	0.21	0.37	1.80
VAE [72]	21.75	2.12	5.29	0.60	0.17	4.98
BUDDI* [43]	22.51	1.35	6.50	0.28	0.36	<u>1.38</u>
Ours w/o Decomposition	19.84	1.18	6.28	0.40	**0.67**	6.06
Ours w/o Shared Network	<u>17.00</u>	<u>0.97</u>	6.15	**0.74**	<u>0.63</u>	3.85
Ours	**12.91**	**0.55**	**6.77**	<u>0.71</u>	**0.67**	**1.33**

such as opening a box or operating an espresso machine. It contains 339 sequences of interaction with 10 objects. We follow the split protocol (*protocol 1*) released by ARCTIC, resulting in 192K training samples, 25K validation samples, and 25K test samples.

Baselines. We mainly consider the two-hand generation baselines from Section 4.1 modified to additionally take an object conditioning in the same manner as our method (Section 3.6). The baselines were further tuned to perform fair comparisons (please refer to the supplementary for details). We additionally consider ContactGen [37], which is the most recent state-of-the-art method on single-hand and object interaction synthesis. We modify ContactGen to generate two-hand interactions and denote it by ContactGen*.

Evaluation metrics. Similar to Section 4.1, we use FHID, KHID, diversity, precision-recall, and penetration volume. To extract two-hand interaction features relative to an object, we train a PointNet++ [50] backbone network specifically for 3D two-hand and object interactions similar to Section 4.1. Please refer to the supplementary for the details of our backbone network. Note that we compute the metrics per object category and report the average scores.

Results. In Table 1b, our method is shown to outperform the baseline methods on most of the metrics by a large margin. Especially, our method yields significantly better scores on FHID and KHID. One notable observation is that ContactGen* does not achieve good performance on general two-hand and object interaction synthesis, by biasing towards heavy contact cases due to its reliance on the contact prior. In contrast, as shown in Figure 4, ours is capable of generating plausible bimanual hand interactions including loosely contacted cases.

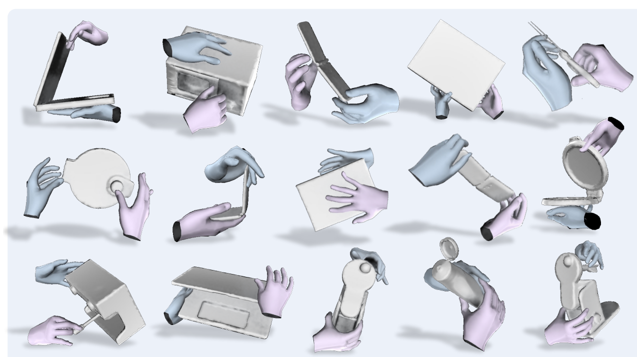


Figure 4. **Object-conditional two-hand interaction synthesized by InterHandGen.** Ours can model plausible and diverse bimanual interactions.

4.3. Monocular Two-Hand Reconstruction

Baseline and Data. We consider InterWild [40] for the baseline, which is the most recent state-of-the-art work proposed for interacting two-hand reconstruction from in-the-wild images. For network training, InterWild uses mixed-batches consisting of motion capture data with full 3D shape supervision (InterHand2.6M [41]) and in-the-wild data with weak 2D keypoints supervision (MSCOCO [25, 36]). In this ill-posed setup, we leverage our diffusion prior to reduce depth ambiguity. In particular, we utilize our pre-trained two-hand diffusion model (used in Section 4.1) to compute the regularization term $\mathcal{L}_{reg}$ defined in Equation 9. We incorporate $\mathcal{L}_{reg}$ into the loss function of InterWild during network training, while other baseline settings (e.g., model architecture) remain unchanged. For testing, we use InterHand2.6M [41] test set and HIC [65] following the original evaluation protocol of InterWild.

Evaluation metrics. We use the same metrics as in InterWild to measure the accuracy of two-hand reconstruction: Mean Per-Joint Position Error (MPJPE), Mean Per-Vertex Position Error (MPVPE), and Mean Relative-Root Position Error (MRRPE) in mm.

Results. As shown in Table 2, our generative prior boosts the reconstruction accuracy of the baseline method in terms of all three metrics, *setting new state-of-the-art on monocular two-hand reconstruction from in-the-wild images*. Especially, it leads to 10% and 18% improvements in MRRPE on InterHand2.6M and HIC datasets, respectively. These results indicate that our generative prior is effective in reducing the shape ambiguity in an ill-posed setup. We also highlight again that our pre-trained prior can be easily incorporated into the existing work in a plug-and-play manner, without a modification of the baseline architecture.

Table 2. **Quantitative comparisons of interacting two-hand reconstruction from in-the-wild images.** Utilizing our generative prior can boost the two-hand reconstruction accuracy.

(a) **Results on InterHand2.6M [41].**

Method	MPVPE ↓	MPJPE ↓	MPRPE ↓
InterWild [40]	13.01	14.83	29.29
InterWild [40] + Ours	**12.10**	**14.53**	**26.56**

(b) **Results on HIC [65].**

Method	MPVPE ↓	MPJPE ↓	MPRPE ↓
InterWild [40]	15.70	16.17	31.35
InterWild [40] + Ours	**15.04**	**15.45**	**26.63**

4.4. Ablation Study

We perform an ablation study to investigate the effectiveness of our self-attention module (*SelfAtt*), classifier-free guidance [20] (*CFG*), and anti-penetration guidance (*APG*). Table 3a compares the generated sample fidelity (measured on FHID and Precision) and diversity with respect to *SelfAtt* and *CFG*. It shows that using *SelfAtt* improves both fidelity and diversity, while *CFG* provides a fidelity-diversity sweet spot as discussed in [20]. In Table 3b, we compare the average penetration volume (in cm^3) and penetration distance (in cm) with and without *APG*. We also measure the proximity ratio, which is the ratio of generated frames that contain close two-hand interactions (where the inter-mesh distance is below $\tau = 2cm$). It is shown that *APG* significantly reduces the amount of shape penetration while not hurting the proximity ratio.

Table 3. **Ablation study results.** We use the same setting as in the two-hand interaction generation experiments (Section 4.1).

(a) **Comparisons on sample fidelity and diversity.** We compare to our method variations in which self-attention (*Ours w/o SelfAtt*) or classifier-free guidance (*Ours w/o CFG*) is not used, respectively.

Method	FHID ↓	Precision ↑	Diversity ↑
Ours w/o SelfAtt	2.87	**0.86**	3.16
Ours w/o CFG	1.12	0.84	**3.61**
Ours	**1.00**	**0.86**	3.59

(b) **Comparisons on inter-penetration.** We compare to our method variation where anti-penetration guidance is not used (*Ours w/o APG*). *PenVol*, *PenDist*, and *ProxRatio* denote penetration volume, penetration distance, and proximity ratio, respectively.

Method	PenVol ↓	PenDist ↓	ProxRatio ↑
Ours w/o APG	6.58	0.40	**0.97**
Ours	**0.76**	**0.04**	**0.97**

5. Conclusion and Future Work

We presented InterHandGen, a diffusion-based framework that learns the generative prior for two-hand interaction with or without an object. Ours provides a theoretical framework to decompose the joint distribution into a sequential modeling problem with unconditional and conditional sampling from a diffusion model. In particular, our experiments show that achieving both diverse and high-fidelity sampling is now possible with the proposed cascaded reverse diffusion. Our approach can be easily extended to more instances, and is easy to integrate into existing learning methods, setting a new state-of-the-art performance on two-hand reconstruction from in-the-wild images.

Limitation and Future Work. Due to the generality of our method, the proposed diffusion prior can be jointly trained with heterogeneous datasets (i.e., a single hand only, a single hand with an object, two hands, and two hands with an object) to build a universal hand prior for all hand-related tasks. Please refer to the supplementary for more discussion. Other future work includes the extension to the temporal dimension and other interaction synthesis problems beyond hands (e.g., animal or human bodies). We hope that our approach will be an important stepping stone towards a unified interaction prior across categories and that it will inspire follow-up work.

Acknowledgements. This work was in part supported by NST grant (CRC 21011, MSIT), KOCCA grant (R2022020028, MCST), and IITP grant (RS-2023-00228996, MSIT). M. Sung acknowledges the support of NRF grant (RS-2023-00209723) and IITP grants (2022-0-00594, RS-2023-00227592) funded by MSIT, Seoul R&BD Program (CY230112), and grants from the DRB-KAIST SketchTheFuture Research Center, Hyundai NGV, KT, NCSOFT, and Samsung Electronics.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375, 2018. 6

[2] Anil Armagan, Guillermo Garcia-Hernando, Seungryul Baek, Shreyas Hampali, Mahdi Rad, Zhaohui Zhang, Shipeng Xie, MingXiu Chen, Boshen Zhang, Fu Xiong, et al. Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3d hand pose estimation under hand-object interaction. In *ECCV*, 2020. 2

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *CVPR*, 2020. 2

[4] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012. 2

[5] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *CVPRW*, 2023. 5

[6] Jonathan Baxter. A model of inductive bias learning. *JAIR*, 2000. 4

[7] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018. 2, 6

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3

[9] 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 *CVPR*, 2021. 2, 6

[10] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 1, 2

[11] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE TPAMI*, 2023. 3

[12] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 5

[13] 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 *CVPR*, 2023. 2, 3, 6

[14] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 2

[15] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. In *IROS*, 2020. 2

[16] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *CVPR*, 2021. 2

[17] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2, 5, 6

[18] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2, 6

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 2, 6

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021. 2, 3, 4, 5, 8

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3, 4, 6

[22] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 2023. 1, 2

[23] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *CVPR*, 2023. 1, 2

[24] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 2021. 1, 2

[25] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 7

[26] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, 2020. 1, 2

[27] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *3DV*, 2021. 1, 2

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2013. 2

[29] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. *CoRR*, abs/2307.07511, 2023. 3

[30] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 2, 6

[31] Jihyun Lee, Junbong Jang, Donghwan Kim, Minhyuk Sung, and Tae-Kyun Kim. Fourierhandflow: Neural 4d hand representation using fourier query flow. In *NeurIPS*, 2023. 1, 2

[32] Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *CVPR*, 2023. 1, 2, 6

[33] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. In *NeurIPS*, 2023. 5

[34] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, 2022. 1, 2, 6

[35] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *CoRR*, abs/2304.05684, 2023. 1, 3

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 7

[37] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *ICCV*, 2023. 2, 3, 7

[38] 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 *CVPR*, 2022. 2, 6

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2015. 3

[40] Gyeongsik Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *CVPR*, 2023. 1, 2, 7, 8

[41] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020. 1, 2, 3, 6, 7, 8

[42] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM TOG*, 2019. 2

[43] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *CoRR*, abs/2306.09337, 2023. 1, 2, 3, 5, 6, 7

[44] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 3

[45] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*, 2012. 2

[46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 2

[47] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, 2021. 2

[48] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2022. 2, 3, 5

[49] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 6

[50] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 6, 7

[51] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data. In *CVPR*, 2023. 2, 6

[52] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *CoRR*, abs/1710.05941, 2017. 6

[53] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *ICCV*, 2023. 1, 2

[54] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM TOG*, 2017. 2, 3, 4, 5, 6

[55] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, 2021. 1, 2

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3

[57] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *NeurIPS*, 2018. 6

[58] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *CoRR*, abs/2303.01418, 2023. 1, 3, 4

[59] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *CVPR*, 2019. 6

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3, 4, 5

[61] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *ACM TOG*, 2017. 2

[62] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2022. 2, 3, 4, 6

[63] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *NeurIPS*, 1995. 4

[64] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *ECCV*, 2022. 1, 2

[65] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 8

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 6

[67] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surv.*, 2022. 3

[68] Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *CVPR*, 2023. 2

[69] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *ICCV*, 2021. 1, 2

[70] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Trans. Knowl. Data En.*, 2021. 4

[71] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3

[72] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *ICCV*, 2023. 1, 2, 6, 7