TriDi: Trilateral Diffusion of 3D Humans, Objects, and Interactions

Ilya A. Petrov[1,2], Riccardo Marin[3,4], Julian Chibane[1,5], and Gerard Pons-Moll[1,2,5]

[1]University of Tübingen, Germany , [2]Tübingen AI Center, Germany,
[3]Technical University of Munich, Germany, [4]Munich Center for Machine Learning, Germany,
[5]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{i.petrov, gerard.pons-moll}@uni-tuebingen.de
riccardo.marin@tum.de, jchibane@mpi-inf.mpg.de

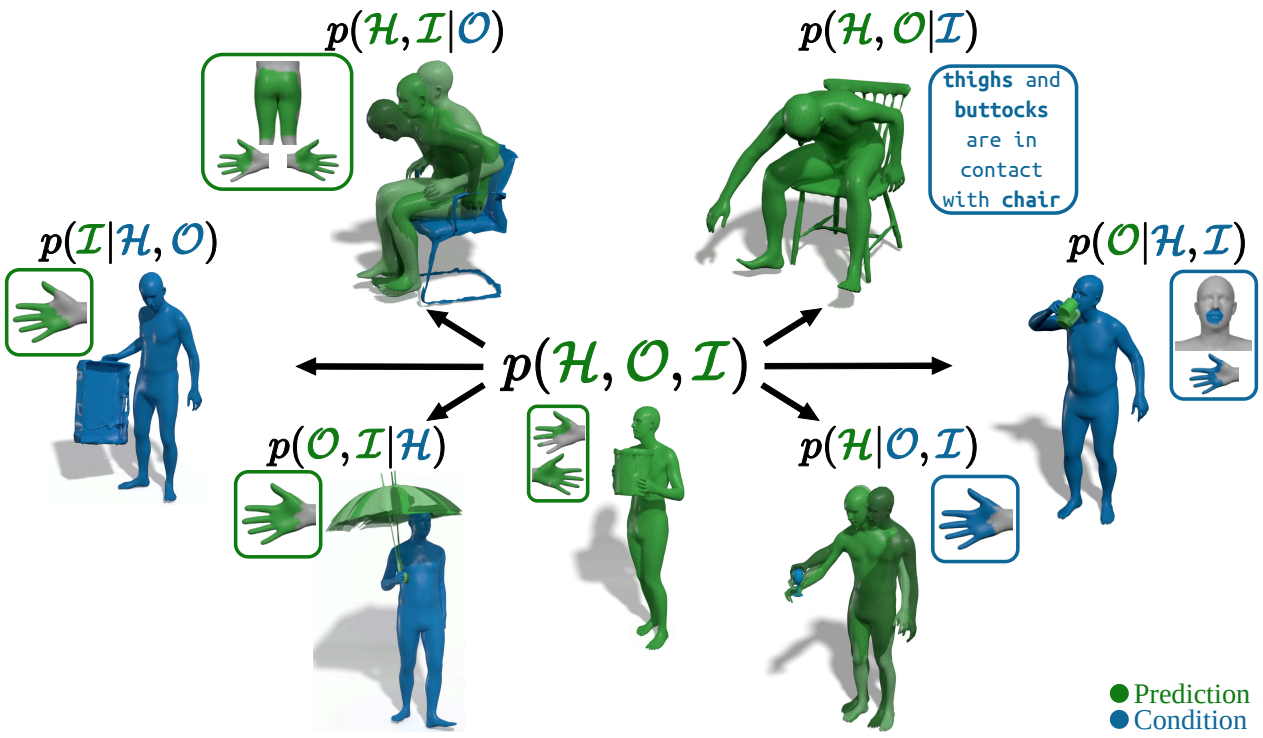

Figure 1. **TriDi**. We present TriDi, the first joint probabilistic model of human pose ($\mathcal{H}$), object ($\mathcal{O}$) and human-object interaction ($\mathcal{I}$). The joint model unifies these three modalities, capturing mutual dependencies between them, and allows for sampling in *seven* conditioning configurations, covering the use cases treated in isolation by previous works. The colors on the image encode prediction and condition.

Abstract

Modeling 3D human-object interaction (HOI) is a problem of great interest for computer vision and a key enabler for virtual and mixed-reality applications. Existing methods work in a one-way direction: some recover plausible human interactions conditioned on a 3D object; others recover the object pose conditioned on a human pose. Instead, we provide the first unified model - TriDi *which works in any direction. Concretely, we generate Human, Object, and Interaction modalities simultaneously with a new three-way diffusion process, allowing to model* seven *distributions with one network. We implement TriDi as a transformer attending to the various modalities' tokens, thereby discovering conditional relations between them. The user can control the interaction either as a text description of HOI or a contact map. We embed these two representations into a shared latent space, combining the practicality of text descriptions with the expressiveness of contact maps. Using a single network, TriDi unifies all the special cases of prior work and extends to new ones, modeling a family of* seven *distributions. Remarkably, despite using a single model, TriDi generated samples surpass one-way specialized baselines on*

GRAB and BEHAVE in terms of both qualitative and quantitative metrics, and demonstrating better diversity. We show the applicability of TriDi to scene population, generating objects for human-contact datasets, and generalization to unseen object geometry. The project page is available at: https://virtualhumans.mpi-inf.mpg.de/tridi/.

1. Introduction

Humans constantly interact with objects around them – they lean on tables, carry backpacks, or touch keyboards. Different *objects* afford different kinds of *human poses*, and vice-versa, different poses support only certain types of objects. Furthermore, given a *human* and an *object*, many *interactions* are possible. For example, we can sit on a chair, lift it, push it, or carry it, and each interaction will require different contacts. We argue that a comprehensive model should capture such interplay of objects, humans, and interactions, regardless of the modality considered as an input condition. Such a joint model is much more flexible than one-way models, giving rise to many applications: generating humans that fit a given object, objects that fit a human pose, unconditional generation, or even automatically annotating the interaction of existing 3D datasets. This versatility is needed in such applications as content creation, AR/VR, ergonomics, and manufacturing.

However, existing works have modeled human-object interaction as the posteriors of human given the object [41, 42, 67, 78, 97] or object given human[58, 88]. Following this paradigm requires a tailored model for each conditioning case and, thus, a specialized design choices, training procedure, and architecture. Such an approach is impractical and difficult to scale. Instead of modeling each individual conditional distribution, we shift this paradigm and design a single compact architecture that models the joint and conditional distributions of *human*, *object*, and *interactions*. By design, we can sample from a joint unconditional distribution of humans, objects, and interactions, as well as from all possible conditional combinations.

We propose **TriDi**, a unified 3D human-object interaction model capturing the joint distribution of *humans*, *objects*, and *interactions*. TriDi produces samples from every conditional distribution arising from the combination of three in addition to the joint distribution, giving rise to $2^3 - 1 = 7$ possible modes of operation, see Fig. 1.

TriDi performs a three-way diffusion building on the UniDiffuser paradigm [2], implemented through token-wise attention, enabling to capture fine-grained relations. Since most interactions imply contact, prior work represents interaction through body contact maps [27, 70]. In contrast to text prompts, this representation is difficult for users to control. Hence, we propose to unify textual descriptions and body contact maps by a joint embedding space. This results in a novel representation that is useful to guide the model and intuitive to the user. To double the effective training data and remove right-handed biases, we augment the HOI by exploiting the left-right symmetry of the interactions.

We demonstrate the flexibility of TriDi, which nicely encapsulates uni-directional methods published in different papers using a single network. Beyond savings in terms of model size and ease of use, TriDi surpasses uni-directional baselines tailored to specific conditioning cases. Moreover, TriDi performs on par or better than the same model trained on one-way conditional tasks (e.g., generating human and interactions conditioned on the object), demonstrating the effectiveness of the joint modeling. TriDi is general and capable of synthesizing a static 3D HOI starting from different inputs, covering all the previous works' use cases plus new ones (Fig. 1). We demonstrate how TriDi can populate scenes with realistic interactions, generalize to novel geometry, and open new applications such as generating objects that fit observed humans and interactions in images.

In summary, our contributions are:

- We formulate TriDi, the first joint model for $P(\mathcal{H}, \mathcal{O}, \mathcal{I})$, modeling it as the three-variable joint distribution and covering a total of *7* modes of operation, rendering prior works as special use cases of our model.
- We propose a novel representation of interaction by jointly embedding body contact maps and textual descriptions, resulting in intuitive control for the user while providing detailed guidance to the model.
- We will release our code, providing the community with a tool for scene population, generation from partial observations, and other tasks that involve 3D HOI.

2. Related Work

From object to human. Modeling 3D HOI from the objects has been studied from diverse perspectives. At a macro scale, studying humans in the context of 3D scenes is prominent [25, 27, 30, 53, 73, 74, 90, 96, 99]. These works are instrumental for downstream tasks like synthetic dataset generation [4, 34, 57, 81]. Dynamic motions in scenes can be conditioned by object's 3D location [8, 26], control points [97], milestones [42, 59], physical properties [98], or text descriptions [43, 65]. Such a high-level perspective on HOI should be complemented by modeling interactions on an object level. Hence, [67, 78] synthesize the motion towards a static object, and [7, 44] focus on manipulation interactions. These works consider temporal sequences, which are demanding to capture, thus limiting the scaling beyond the settings seen at training time. Synthesizing hand-object interactions presents several challenges [17, 18], which originated specialized methods [12, 32, 46, 48, 75, 87, 92]. Producing accurate prediction raised the demand for hand-object refinement [52, 68, 101, 102], but those methods are limited by smaller objects and hand-held interactions. TriDi works

with single frames, models contact beyond the hands, and thus supports human synthesis involving diverse objects.

From human to object. Reasoning about objects from humans is a less explored direction, despite the applicability in AR/VR, where humans often interact with objects without a physical counterpart. [86, 89] generate scenes satisfying the observed human motion. Object Pop-up [58] regresses an object position from a 3D human point cloud, disregarding the uncertainty behind this ill-posed task. An interesting self-supervised approach regresses the heatmap for plausible object center location [24], while the follow-up work [38] studies objects' affordances. TriDi models a joint distribution of HOI, naturally allowing for the uncertainty in predictions while retaining downstream applications.

Contacts modeling. Contact is the physical medium of many human-object interactions. In practice, contact maps are a good proxy to promote realism [27, 55, 78]; however, their capture is often complicated by manual annotation [70] or the need for specialized hardware [5]. Contact is represented in a range of ways, e.g., as distances [15], proximity [96], or maps on the body [70] and the object [16]. Contact is often modeled on the hands [5, 6, 21], with recent works considering the full body [27, 70]. An alternative is to represent the interaction through text [15, 65, 90]. While more interpretable and controllable, this representation limits the possibility of spatial reasoning for the methods. In our work, we combine text and contact maps in a shared latent space, inheriting the advantage of both.

Joint modeling. A number of works focus on reconstructing interactions with single objects external data such as images [9, 47, 55, 70, 76, 77, 79, 81, 85, 95], videos [80, 88, 100], and multi-view recording setups [35, 93]. These works are backed by recent HOI data collections [3, 16, 31, 33, 49, 84]. Modeling hand-object interactions jointly requires tailored methods [37]. FLEX [69] combines grasp with full-body generation to fit HOI samples in the scene constraints. IMoS [19] and InterDiff [82] start from past observations to forecast the continuation of a 3D HOI sequence. CG-HOI [15] synthesizes human and object motion from text, training only on one dataset at a time. These methods rely on strong conditioning: temporal HOI sequence, deterministic future, and text. Using single-frame data, although challenging and ambiguous, is more general and scalable. Finally, we find it exciting to mention recent works in nascent fields: compositional shape generation including human and object [10, 11, 13, 40], and modeling multiple human and object interactions [39, 54, 63, 91, 94], suggesting more complex synergies in HOI. TriDi models HOI jointly, covering all the use cases of previous works tailored to the specific conditioning.

3. Background

Probabilistic Diffusion. A Diffusion process [28, 64] is divided into a forward process that progressively noises the original data sample $\mathbf{z}_0$, and a backward process that recovers the sample $\mathbf{z}_0$ from the noise using a learned model.

Formally, the forward process follows a Markov chain of T steps; it produces a series of time-dependent distributions $q(\mathbf{z}_t|\mathbf{z}_{t-1})$: $q(\mathbf{z}_{1:T}|\mathbf{z}_0) = \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1})$. At every timestamp, we inject noise into the distribution until the final $\mathbf{z}_T$ converges to a sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $\beta_0 = 0$, and $\beta_t \in (0, 1)$:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}). \tag{1}$$

We follow the formulation of Denoising Diffusion Probabilistic Model (DDPM) [28] to obtain a closed-form expression for $\mathbf{z}_t$ (formulation is provided in the Sup. Mat.).

The inference is then performed by reversing the process, starting from $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ and recovering samples from the original distribution. Instead of recovering the added noise ϵ for each timestep, we follow the formulation of [61] and recover the original sample $\mathbf{z}_0$. To achieve this, we parametrize the reverse process by a denoising neural network $\mathcal{D}_\psi$ that is trained to recover the original sample $\mathbf{z}_0$ from the noised sample $\mathbf{z}_t$ at timestep t given the condition c. Defining for brevity $\mathbb{E}_p \equiv \mathbb{E}_{\mathbf{z}_0 \sim p_{data}}$, $\mathbb{E}_t \equiv \mathbb{E}_{t \sim \mathcal{U}\{0,\dots,T\}}$, and $\mathbb{E}_q \equiv \mathbb{E}_{\mathbf{z}_t \sim q(\mathbf{z}_t|\mathbf{z}_0)}$ we obtain the training objective (inference formulation is provided in the Sup. Mat.):

$$\min_\psi \mathbb{E}_p \, \mathbb{E}_t \, \mathbb{E}_q \, \|\mathcal{D}_\psi(\mathbf{z}_t; c, t) - \mathbf{z}_0\|). \tag{2}$$

Multimodal diffusion. While the previous formulation handles the generation of a single modality, data often constitutes a composition of multiple modalities, e.g., $\mathbf{z}_0 = (\mathbf{x}_0, \mathbf{y}_0) \sim p(x, y)$. Hence, we are naturally interested in modeling this joint distribution together with the marginals $p(y)$ and $p(x)$, as well as conditional ones $p(x|y)$ and $p(y|x)$. UniDiffuser [2] proposes a network $\mathcal{D}_\psi(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}; t^x, t^y)$ dedicated to recovering $\mathbf{z}_0$ given a noisy sample from the joint distribution.

Adapting the definitions from Eq. 2 to two modalities: $\mathbb{E}_p \equiv \mathbb{E}_{(\mathbf{x}_0, \mathbf{y}_0) \sim p(x,y)}$, $\mathbb{E}_t \equiv \mathbb{E}_{(t^x, t^y) \sim \mathcal{U}\{0,\dots,T\}^2}$, and $\mathbb{E}_q \equiv \mathbb{E}_{\mathbf{x}_{t^x} \sim q(\mathbf{x}_{t^x}|\mathbf{x}_0), \mathbf{y}_{t^y} \sim q(\mathbf{y}_{t^y}|\mathbf{y}_0)}$ we obtain the following training objective:

$$\min_\psi \mathbb{E}_p \, \mathbb{E}_t \, \mathbb{E}_q \, \|\mathcal{D}_\psi(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}; t^x, t^y) - (\mathbf{x}_0, \mathbf{y}_0)\|. \tag{3}$$

The benefit of minimizing the objective in Eq. 3 is that the resulting network captures all the desired distributions. Namely, setting $t^y = T$ allows to model the marginal distribution $p(x)$, on the other hand, $t^y = 0$ corresponds to conditional distribution $p(x|y)$. we note that in its original formulation, UniDiffuser is designed to consider text and images as two diffusion modalities.

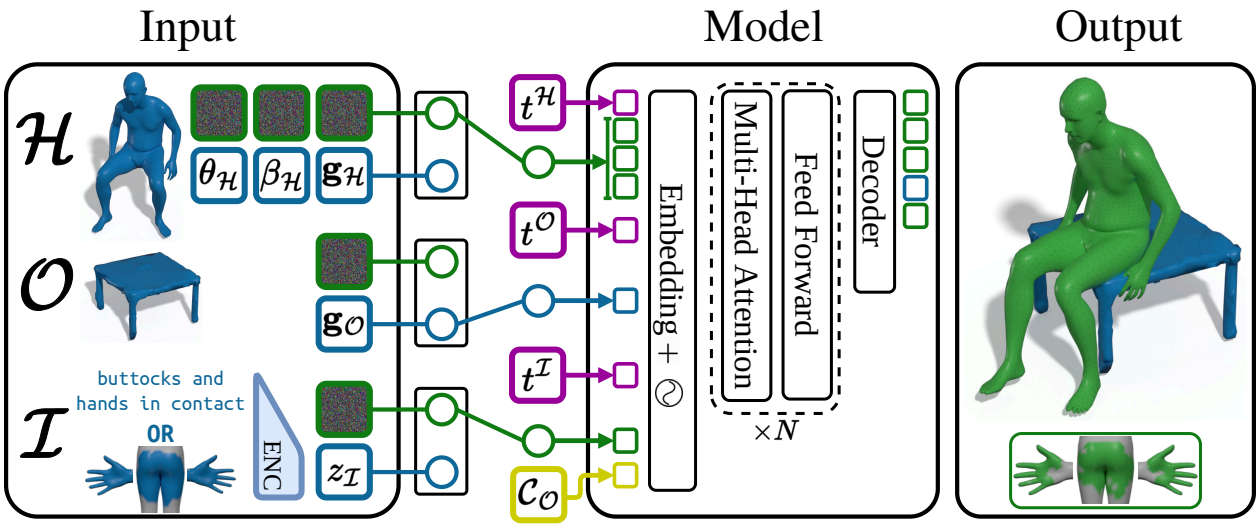

Figure 2. **TriDi Overview.** TriDi is a Trilateral Diffusion for Human $\mathcal{H}$ (pose $\theta_\mathcal{H}$, identity $\beta_\mathcal{H}$, and 6-DoF global pose $\mathbf{g}_\mathcal{H}$), Object $\mathcal{O}$ (6-DoF global pose $\mathbf{g}_\mathcal{O}$) and Interaction $\mathcal{I}$ (Contact-Text latent $\mathbf{z}_\mathcal{I}$). In this figure the model is configured to sample $p(\mathcal{H}, \mathcal{I}|\mathcal{O})$. One of the *seven* operating modes is chosen by adjusting the **timestamp** to be 0 for a **given condition** ($t^\mathcal{O}$ above) and T for the **desired prediction** ($t^\mathcal{H}$ and $t^\mathcal{I}$ above), and supplying an **object class condition** ("Table" above).

4. Method

Overview. Our goal is to model the three-variable joint distribution of Human $\mathcal{H}$, Object $\mathcal{O}$, and Interaction $\mathcal{I}$, taking as input only the object class with a canonical representation and, optionally, conditions from the three modalities. Previous works focus on one-way cases, with fixed conditional modality, e.g., human from an object, ($P(\mathcal{H}|\mathcal{O})$, [42]) or human and object from a text ($P(\mathcal{H}, \mathcal{O}|\mathcal{I})$, [15]). In contrast, we want to model $P(\mathcal{H}, \mathcal{O}, \mathcal{I})$, providing *a unified model for Human-Object Interaction*. To achieve this, we introduce TriDi, a transformer based model that operates on tokenized representations of $\mathcal{H}$, $\mathcal{O}$, and $\mathcal{I}$ (an overview is presented in Fig.2). The following sections define the representations for HOI (Sec. 4.1), introduce our trilateral diffusion formulation (Sec. 4.2), and discuss training details (Sec. 4.3).

4.1. Modalities representations

Human and object. Following SMPL+H body model [50, 62] we decompose the human as:

$$\mathcal{H} = (\theta_\mathcal{H}, \beta_\mathcal{H}, \mathbf{g}_\mathcal{H}), \tag{4}$$

where $\mathbf{g}_\mathcal{H} \in \mathbb{R}^9$ is a 6-DoF global pose, and $\theta_\mathcal{H} \in \mathbb{R}^{51\times3}$ and $\beta_\mathcal{H} \in \mathbb{R}^{10}$ are the pose and shape parameters respectively of a template function that maps them to a triangular mesh. In TriDi we rely on a decimated version of SMPL with vertices $\mathbf{V}_\mathcal{H} \in \mathbb{R}^{690}$, reducing the computations while retaining the capability to recover the full template mesh.

For objects, the canonical geometry is given as input by the user and serves as conditioning for our model. We represent it as $\mathcal{C}_\mathcal{O} = (\mathbf{f}_\mathcal{O}, \mathbf{y}_\mathcal{O})$, consisting of $\mathbf{f}_\mathcal{O} \in \mathbb{R}^{1024}$ PointNeXt [60] features and a one-hot class encoding vector $\mathbf{y}_\mathcal{O}$. Our model diffuses objects' 6-DoF global pose $\mathbf{g}_\mathcal{O} \in \mathbb{R}^9$:

$$\mathcal{O} = (\mathbf{g}_\mathcal{O}). \tag{5}$$

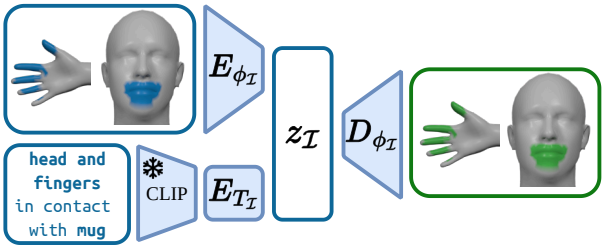

Figure 3. **Architecture of Contact-Text Interactions model**. We train a mapping from the contact map $E_{\phi_\mathcal{I}}$ and CLIP embedding $E_{T_\mathcal{I}}$ to a joint latent space $\mathbf{z}_\mathcal{I}$ that is used to represent the interaction $\mathcal{I}$. Jointly, we train the decoder $D_{\phi_\mathcal{I}}$ that maps the latent back to the contact map.

Interactions. Representing interaction $\mathcal{I}$ is particularly challenging as we want to combine the intuitiveness of text descriptions with the expressiveness of contact maps. Our solution is to learn a compact latent representation that encodes both in a joint space. Given a set of pairs $(T_\mathcal{I}, \phi_\mathcal{I})$, where $T_\mathcal{I}$ is a text description and $\phi_\mathcal{I} \in \{0,1\}^{690}$ is a contact map defined on $\mathbf{V}_\mathcal{H}$, we simultaneously train two encoders $E_{\phi_\mathcal{I}}(\phi_\mathcal{I}) = \mathbf{z}_\mathcal{I} \in \mathbb{R}^{128}$ for contact maps and $E_{T_\mathcal{I}}(\text{CLIP}(T_\mathcal{I})) = \mathbf{z}_\mathcal{I}$ for CLIP embedding of $T_\mathcal{I}$, as well as decoder $D_{\phi_\mathcal{I}}$ mapping the latent space back to the contact map $\phi_\mathcal{I}$. We optimize them with the following loss:

$$\begin{aligned} L_{CT}(T_\mathcal{I}, \phi_\mathcal{I}) = &\text{BCE}(D_{\phi_\mathcal{I}}(E_{\phi_\mathcal{I}}(\phi_\mathcal{I})), \phi_\mathcal{I}) + \\ &\text{BCE}(D_{\phi_\mathcal{I}}(E_{T_\mathcal{I}}(T_\mathcal{I})), \phi_\mathcal{I}) + \\ &\|E_{T_\mathcal{I}}(T_\mathcal{I}) - E_{\phi_\mathcal{I}}(\phi_\mathcal{I})\|_2, \end{aligned} \tag{6}$$

where BCE is the Binary-Cross Entropy loss, and the loss terms are auto encoding loss, text-to-contact map encoding loss, and latent space similarity loss. A sketch of this module can be seen in Fig.3. Thus, the interactions are represented via a compact code from the unified latent space:

$$\mathcal{I} = (\mathbf{z}_\mathcal{I}). \tag{7}$$

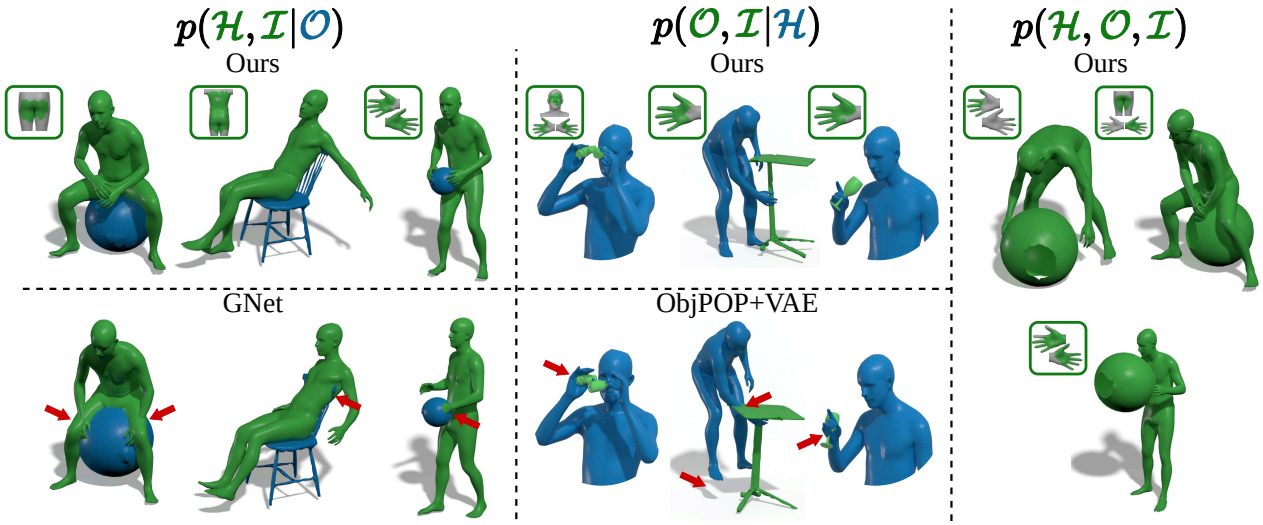

Figure 4. **Comparison with baselines**. In the two left-most columns, we show three samples for $p(\mathcal{H},\mathcal{I}|\mathcal{O})$ and $p(\mathcal{O},\mathcal{I}|\mathcal{H})$ from BEHAVE and GRAB test sets. TriDi's generations are better aligned with the condition, causing less interpenetration (e.g., for basketball), respecting fine-grained details (e.g., for smaller objects), and demonstrating more diversity for limbs not restricted by contacts (e.g., for yoga ball). On the right, TriDi is the *only model* that can sample from $p(\mathcal{H},\mathcal{O},\mathcal{I})$.

4.2. TriDi: Trilateral Diffusion for HOI

Diffusion formulation. To model the joint distribution of Human $\mathcal{H}$, Object $\mathcal{O}$, and Interaction $\mathcal{I}$ we formulate a three-way diffusion. For brevity, we define:

$$\begin{aligned}\mathbb{E}_p &\equiv \mathbb{E}_{(\mathcal{H}^0,\mathcal{O}^0,\mathcal{I}^0)\sim p(\mathcal{H},\mathcal{O},\mathcal{I})},\\ \mathbb{E}_t &\equiv \mathbb{E}_{(t^{\mathcal{H}},t^{\mathcal{O}},t^{\mathcal{I}})\sim\mathcal{U}\{0,\dots,T\}^3},\\ \mathbb{E}_q &\equiv \mathbb{E}_{\mathcal{H}^{t^{\mathcal{H}}}\sim q(\mathcal{H}^{t^{\mathcal{H}}}|\mathcal{H}^0),\mathcal{O}^{t^{\mathcal{O}}}\sim q(\mathcal{O}^{t^{\mathcal{O}}}|\mathcal{O}^0),\mathcal{I}^{t^{\mathcal{I}}}\sim q(\mathcal{I}^{t^{\mathcal{I}}}|\mathcal{I}^0)}\end{aligned} \tag{8}$$

Hence, the parameters ψ of a model TriDi_ψ are optimized by minimizing the objective (extending Eq. 3):

$$\begin{aligned}\min_\psi \mathbb{E}_p\mathbb{E}_t\mathbb{E}_q\|\text{TriDi}_\psi(\mathcal{H}^{t^{\mathcal{H}}},\mathcal{O}^{t^{\mathcal{O}}},\mathcal{I}^{t^{\mathcal{I}}};t^{\mathcal{H}},t^{\mathcal{O}},t^{\mathcal{I}};\mathcal{C}_{\mathcal{O}})\\ -(\mathcal{H}^0,\mathcal{O}^0,\mathcal{I}^0)\|_2.\end{aligned} \tag{9}$$

In practice, we build our method on top of a transformer [72] architecture with an additional embedding layer for all the input modalities that maps them into a common token space. Formally, TriDi_ψ is defined as:

$$\begin{aligned}\text{TriDi}_\psi :(\theta^{t^{\mathcal{H}}}_{\mathcal{H}},\beta^{t^{\mathcal{H}}}_{\mathcal{H}},\mathbf{g}^{t^{\mathcal{H}}}_{\mathcal{H}},\mathbf{g}^{t^{\mathcal{O}}}_{\mathcal{O}},\mathbf{z}^{t^{\mathcal{I}}}_{\mathcal{I}};t^{\mathcal{H}},t^{\mathcal{O}},t^{\mathcal{I}},\mathcal{C}_{\mathcal{O}})\mapsto\\ (\hat{\mathcal{H}},\hat{\mathcal{O}},\hat{\mathcal{I}})\equiv(\hat{\theta}_{\mathcal{H}},\hat{\beta}_{\mathcal{H}},\hat{\mathbf{g}}_{\mathcal{H}},\hat{\mathbf{g}}_{\mathcal{O}},\hat{\mathbf{z}}_{\mathcal{I}}).\end{aligned} \tag{10}$$

We remark that the only required conditioning for TriDi is the object representation $\mathcal{C}_{\mathcal{O}}$, while other inputs are optional depending on the operating mode. To help the network learn the relation between the different modalities, the triplet $\mathcal{H}$, $\mathcal{O}$, $\mathcal{I}$ is tokenized, and we use token-level self-attention to attend fine-grained interaction among the three modalities.

Guidance. Despite explicitly modeling the interaction modality, the diffusion predictions do not always satisfy the contact. For 3D HOI, this is a hard constraint to respect in order to avoid floating objects and interpenetrations. In order to enforce contacts through the denoising process, we adopt a classifier-based guidance [14] that perturbs the model's prediction on every diffusion step following the feedback of a supervising function $\mathcal{F}$.

Our idea is to force the human to be in contact with the object where the contact map is active. The contact map $\hat{\phi}_{\mathcal{I}} = D_{\phi_{\mathcal{I}}}(\hat{\mathbf{z}}_{\mathcal{I}})$ predicted by TriDi enables such use of self-supervised guidance at each diffusion step. We formulate the supervising function as:

$$\mathcal{F}(\hat{\mathcal{H}},\hat{\mathcal{O}},\hat{\mathcal{I}}) = \sum_{j\in|\mathbf{V}_{\mathcal{H}}|}|\hat{\phi}_{\mathcal{I}j}\hat{\mathbf{d}}_j|, \tag{11}$$

where $\hat{\mathbf{d}}\in\mathbb{R}^{690}$ contains for every vertex of the predicted human $\hat{\mathbf{V}}_{\mathcal{H}}$ the distance to the closest vertex of the predicted object $\hat{\mathbf{V}}_{\mathcal{O}}$:

$$\hat{\mathbf{d}}_j = \min_{i\in|\hat{\mathbf{V}}_{\mathcal{O}}|}\|\hat{\mathbf{V}}^j_{\mathcal{H}}-\hat{\mathbf{V}}^i_{\mathcal{O}}\|_2. \tag{12}$$

We adopt the reconstruction guidance formulation of [29], where the predicted sample $(\hat{\mathcal{H}},\hat{\mathcal{O}},\hat{\mathcal{I}}) = \text{TriDi}_\psi(\mathcal{H},\mathcal{O},\mathcal{I};t^{\mathcal{H}},t^{\mathcal{O}},t^{\mathcal{I}},\mathcal{C}_{\mathcal{O}})$ is directly modified on each denoising step. The reconstruction guidance with scale λ is thus formulated as:

$$(\hat{\mathcal{H}},\hat{\mathcal{O}},\hat{\mathcal{I}}) := (\hat{\mathcal{H}},\hat{\mathcal{O}},\hat{\mathcal{I}}) - \lambda\nabla_{\mathcal{H}^{t^{\mathcal{H}}},\mathcal{O}^{t^{\mathcal{O}}},\mathcal{I}^{t^{\mathcal{I}}}}\mathcal{F}(\hat{\mathcal{H}},\hat{\mathcal{O}},\hat{\mathcal{I}}). \tag{13}$$

4.3. Training

Contact-Text Labeling. To train our method, we must collect the contact maps $\phi_{\mathcal{I}}$ and the text descriptions $T_{\mathcal{I}}$, which are unavailable for many 3D HOI datasets. We define a simple automatic annotation procedure: for every training sample, we obtain the human-object distances $\mathbf{d}$ as described in Equation 12 and threshold them to obtain a binary contact map $\phi_{\mathcal{I}}$. We detect which of the 24 body parts of the human template contains at least one vertex in contact,

Method	BEHAVE $\mathcal{H},\mathcal{I}\|\mathcal{O}$			$\mathcal{O},\mathcal{I}\|\mathcal{H}$		
	1-NNA ($\rightarrow$ 50)	COV↑	MMD↓	1-NNA ($\rightarrow$ 50)	COV↑	MMD↓
ObjPOP [58] + cVAE	-	-	-	$81.36^{\pm 0.2}$	$35.02^{\pm 0.1}$	$0.329^{\pm 0.003}$
GNet [67]	$80.01^{\pm 0.4}$	$40.71^{\pm 0.4}$	$1.789^{\pm 0.036}$	-	-	-
s-TriDi-OI **(Ours)**	-	-	-	$65.06^{\pm 0.5}$	$50.49^{\pm 0.1}$	$\underline{0.167}^{\pm 0.001}$
s-TriDi-HI **(Ours)**	$\underline{69.51}^{\pm 0.2}$	$\underline{46.97}^{\pm 0.4}$	$\underline{1.358}^{\pm 0.010}$	-	-	-
TriDi **(Ours)**	$\mathbf{67.89}^{\pm \mathbf{0.3}}$	$\mathbf{47.81}^{\pm \mathbf{0.2}}$	$\mathbf{1.352}^{\pm \mathbf{0.005}}$	$\mathbf{63.72}^{\pm \mathbf{0.3}}$	$\mathbf{51.71}^{\pm \mathbf{0.1}}$	$\mathbf{0.166}^{\pm \mathbf{0.001}}$

Method	GRAB $\mathcal{H},\mathcal{I}\|\mathcal{O}$			$\mathcal{O},\mathcal{I}\|\mathcal{H}$		
	1-NNA ($\rightarrow$ 50)	COV↑	MMD↓	1-NNA ($\rightarrow$ 50)	COV↑	MMD↓
ObjPOP [58] + cVAE	-	-	-	$82.09^{\pm 0.3}$	$37.52^{\pm 0.8}$	$0.483^{\pm 0.061}$
GNet [67]	$89.64^{\pm 0.8}$	$39.33^{\pm 1.2}$	$1.422^{\pm 0.087}$	-	-	-
s-TriDi-OI **(Ours)**	-	-	-	$\underline{66.78}^{\pm 0.8}$	$\underline{48.27}^{\pm 0.1}$	$\mathbf{0.252}^{\pm \mathbf{0.012}}$
s-TriDi-HI **(Ours)**	$\mathbf{82.65}^{\pm \mathbf{0.1}}$	$\mathbf{42.87}^{\pm \mathbf{0.2}}$	$\mathbf{0.917}^{\pm \mathbf{0.004}}$	-	-	-
TriDi **(Ours)**	$\underline{82.71}^{\pm 0.5}$	$\underline{42.76}^{\pm 0.3}$	$\underline{0.930}^{\pm 0.012}$	$\mathbf{65.02}^{\pm \mathbf{0.7}}$	$\mathbf{48.84}^{\pm \mathbf{1.2}}$	$\underline{0.268}^{\pm 0.011}$

Table 1. **Quality of Generated Distribution**. TriDi is the only one operating in all the modalities and shows better capability in covering data distribution, improving up to 47%.

and we use this information to compose the labels following one of the predefined templates, e.g., "`[parts] are in contact with [object]`".

Augmentation. The lack of interaction variability in the datasets has been one of the main challenges for us. Often, interactions are performed by a single person, statistically introducing a bias toward right-handed interactions. Surprisingly, no previous human-object interaction modeling method addresses this problem. Hence, we mirror every sample through the ZY plane, doubling the training data. While the lack of perfect symmetry causes small artifacts, we demonstrate that these are negligible, and the augmentation is highly beneficial for generalization.

Losses. During training, TriDi_ψ takes as input the object class condition $\mathcal{C}_\mathcal{O}$, three timesteps $(t^\mathcal{H}, t^\mathcal{O}, t^\mathcal{I})$, and a noisy version of the tokenized representation of human $\theta_\mathcal{H}^{t^\mathcal{H}}$, $\beta_\mathcal{H}^{t^\mathcal{H}}$ and $\mathbf{g}_\mathcal{H}^{t^\mathcal{H}}$, object $\mathbf{g}_\mathcal{O}^{t^\mathcal{O}}$, and interaction $\mathbf{z}_\mathcal{I}^{t^\mathcal{I}}$, generating the predictions $\hat{\theta_\mathcal{H}}, \hat{\beta_\mathcal{H}}, \hat{\mathbf{g}_\mathcal{H}}, \hat{\mathbf{g}_\mathcal{O}}, \hat{\mathbf{z}_\mathcal{I}}$. The learning is supervised by the ground truth representations $\theta_\mathcal{H}, \beta_\mathcal{H}, \mathbf{g}_\mathcal{H}, \mathbf{g}_\mathcal{O}, \mathbf{z}_\mathcal{I}$ and templates vertex positions $\mathbf{V}_\mathcal{H}, \mathbf{V}_\mathcal{O}$. We also incorporate the supervision on distances $\mathbf{d}$, fostering spatial alignment. We report the loss details in Sup. Mat.

5. Experiments

In this section, we compare with one-way methods, assessing the quality of our generation in terms of distribution and spatial consistency. Comparing with specialized approaches in Section 5.1 is challenging for TriDi since it is designed as a unified framework, not privileging any particular modality. We also demonstrate the utility of our representation formulation and the usability of our interaction representation. In Section 5.2, we ablate the components of our method, providing insights into their specific contribution. Finally, we demonstrate applications arising from TriDi in Section 5.3. In the Sup. Mat., we also include an analysis of the running time, experiments with unseen geometries, and a user study to validate the generation quality.

Datasets For our comparisons, we train TriDi and baselines on the union of BEHAVE [3] and GRAB [66], following the train-test split provided by Object Pop-up[58]. We also explore the scalability of TriDi by extending the training to InterCap [31] and OMOMO [42]. We provide descriptions of these datasets and their sampling in Sup. Mat.

5.1. Comparison to one-way methods

Metrics. We evaluate the *quality of the generated distributions* by comparing the generated samples $g \in S_g$ with the reference ones $r \in S_r$ coming from the test sets (with $|S_g| = |S_r|$), reporting the statistics across *three* sampling runs. We report three measures: The Coverage (**COV**) [1] matches every sample of S_g with the closest sample of S_r and counts the percentage of samples in S_r that are associated with at least one generated sample (100 indicates perfect overlap). The Minimum Matching Distance (**MMD**) [1] measures the average distance of samples in the reference set to their closest neighbors in the generated set, quantifying the misalignment of the distributions. The 1-nearest neighbor accuracy (**1-NNA**) [83] measures the leave-one-out accuracy over the union of $S_r \cup S_g$; the optimal value is 50. To evaluate the *Geometrical Consistency of Generation* of humans, we report the Mean Per Joint Position Error (**MPJPE**) that measures in mm. the error in predicting body joints. We also report its value after applying Procustes Analysis (**MPJPE-PA**) [20], alleviating the effect of rotation and scale. For the object we employ the vertex-to-vertex ($\mathbf{E_{v2v}}$) and the object center ($\mathbf{E_c}$) errors, together with contact accuracy ($\mathbf{Acc_{cont}}$). For TriDi, we measure the error for the contact predicted directly by the method and the one calculated from the generated 3D HOI. We refer to Sup. Mat. for metrics' rigorous definitions.

Baselines. TriDi is the first approach *trained only once* and addressing all the seven human-object interaction combinations. We compare with single-frame methods specialized in different modalities, posing a challenge to our general setup. For $P(\mathcal{H}, \mathcal{I}|\mathcal{O})$ we rely on **GNet** [67], while for $P(\mathcal{O}, \mathcal{I}|\mathcal{H})$ we choose Object Pop-up [58] (**ObjPOP**). Since this latter is a regressive method, we provide a further baseline by substituting its object's center MLP with a cVAE (**ObjPOP** [58]**+cVAE**). As in [58], we integrate a Nearest-Neighbor baseline (**NN**), which uses the input condition modality to query the training set and return the output associated with the most similar frame. For objects, the similarity is in terms of $\mathbf{g}_{\mathcal{O}}$, and for humans, it is the distance of body joints after centering the root. Additionally, to evaluate the benefits of the joint model, we train two variants of TriDi that work only in a single configuration: **s-TriDi-OI** for $P(\mathcal{O}, \mathcal{I}|\mathcal{H})$ and **s-TriDi-HI** for $P(\mathcal{H}, \mathcal{I}|\mathcal{O})$.

Comparison: Quality of Generated Distribution. For every method, we generate as many samples as the one contained in the test sets of BEHAVE [3] and GRAB [66], and we compare the generated and the test distributions relying on the metrics described above. We consider cases when the condition modality is the object $\mathcal{H}, \mathcal{I}|\mathcal{O}$ or the human $\mathcal{O}, \mathcal{I}|\mathcal{H}$. We exclude NN and ObjPOP from this comparison since they do not have variance in prediction. In Tab.1, we report the mean and the variance along *three* sampling runs. Despite being more general, TriDi outperforms the specialized baselines on all the metrics, obtaining improvement up to *47%*. The consistently higher COV and lower MMD indicate we better cover the real data distribution, while a 1-NNA close to 50 suggests this is not due to memorizing. We report a qualitative comparison in Fig.4. Notably, TriDi performs better or on par with s-TriDi-HI and s-TriDi-OI , indicating that joint training benefits the generalization capabilities of the model. To compare the methods further, we conducted a user study that collected 40 responses. In summary, our method's output is frequently preferred w.r.t. the baselines' ones ($\sim$ 80% of the cases) and on par with the GT samples($\sim$ 46% of the cases); further details are provided in Sup. Mat.

Comparison: Geometrical Consistency of Generation. Comparing generations' spatial consistency with ground truth is not straightforward since a condition may lead to multiple solutions. Hence, in Tab 2, we report the error considering the best out of *three* samples (in the case of NN, we consider 3-NN). The improvement over all the errors indicates that our better distribution representation comes with high precision and an understanding of spatial relations. For GRAB, we notice a drastic improvement in object center and orientation, even over the regressive ObjPOP. Considering that our contact is always more accurate, we conclude

Method	BEHAVE $\mathcal{H}, \mathcal{I}\|\mathcal{O}$			BEHAVE $\mathcal{O}, \mathcal{I}\|\mathcal{H}$		
	MPJPE↓	MPJPE-PA↓	Acc_{cont} ↑	E_{v2v} ↓	E_c ↓	Acc_{cont} ↑
NN	30.5	14.2	95.0/NA	33.2	22.0	95.4/NA
ObjPOP [58]	-	-	-	**27.5**	22.6	95.2/NA
ObjPOP [58] + cVAE	-	-	-	35.2	23.5	93.6/NA
GNet [67]	35.6	14.6	94.6/NA	-	-	-
s-TriDi-OI (**Ours**)	-	-	-	27.9	15.6	98.8/**96.2**
s-TriDi-HI (**Ours**)	21.0	12.5	**95.6**/**96.5**	-	-	-
TriDi (**Ours**)	**20.8**	**12.3**	95.5/**96.5**	28.0	**15.3**	**95.9**/96.1

Method	GRAB $\mathcal{H}, \mathcal{I}\|\mathcal{O}$			GRAB $\mathcal{O}, \mathcal{I}\|\mathcal{H}$		
	MPJPE↓	MPJPE-PA↓	Acc_{cont} ↑	E_{v2v} ↓	E_c ↓	Acc_{cont} ↑
NN	18.9	13.0	97.1/NA	13.1	11.8	97.8/NA
ObjPOP [58]	-	-	-	9.4	7.7	98.1/NA
ObjPOP [58] + cVAE	-	-	-	13.6	12.3	97.4/NA
GNet [67]	26.7	15.5	96.6/NA	-	-	-
s-TriDi-OI (**Ours**)	-	-	-	**6.9**	**4.9**	**99.0**/**98.7**
s-TriDi-HI (**Ours**)	16.0	11.6	**98.0**/98.1	-	-	-
TriDi (**Ours**)	**15.3**	**11.1**	**98.0**/**98.3**	**6.9**	5.0	**99.0**/98.2

Table 2. **Geometrical Consistency of Generation**. TriDi shows a high level of consistency both for human and object predictions. Our contact prediction indicates the networks have also learned to reason based on the interaction modality. For contacts, we show both the accuracy of contacts inferred from $\mathcal{H}$ and $\mathcal{O}$ meshes, as well as diffused contacts $\mathcal{I}$ (when available).

that our predictions produce more realistic samples. We notice that the predicted contact is better or on par with the contact inferred from meshes, suggesting the network has developed an understanding of the $\mathcal{I}$ modality. We show in Fig. 5 the flexibility of our $\mathcal{I}$ representation using varied text descriptions. In Sup. Mat. we show the qualitative examples of our method adapting to object geometries unseen at training time.

5.2. Ablations

We perform an ablation to analyze the role of the augmentation, $\mathcal{I}$ modality modeling, and the guidance (we report quantitative evaluation in Sup. Mat. First, our full model obtains the best performance in most metrics. Our augmentation has a valuable effect in improving the 1-NNA, suggesting a better distribution. The guidance and $\mathcal{I}$ modality plays a crucial role in geometrical consistency, both for the human and the object. The complexity of considering three modalities instead of two seems, in general, beneficial.

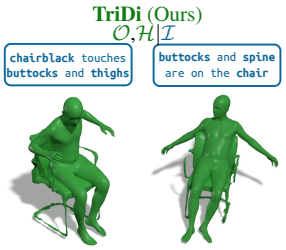

Figure 5. **Text Interaction.** TriDi supports text conditioning for $\mathcal{I}$ modality, providing user control on the contact.

5.3. Applications

In this section, we describe applications in populating scenes, interaction reconstructions, and sequences with keyframing. The results are obtained by the same TriDi model evaluated in Section 5.1. We show generalization to

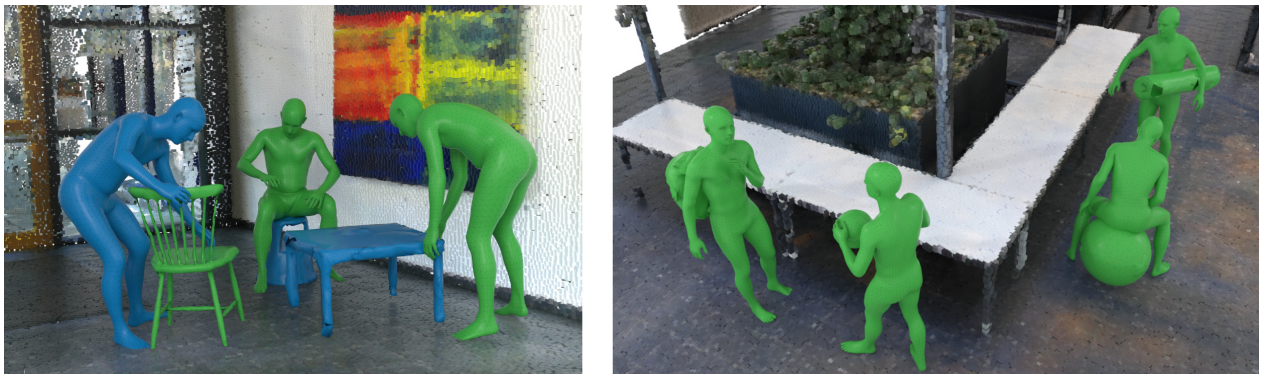

Figure 6. **Scene populating**. Using 3D scans from HPS [22], we validate the practicality of TriDi for scene population in various conditioning cases. On the left, we demonstrate conditional synthesis of human-object interactions. On the right, TriDi is used for the joint generation of humans and objects.

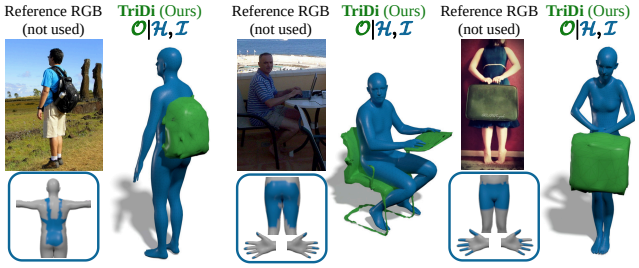

Figure 7. **Interaction reconstruction**. DECO [70] annotates human $\mathcal{H}$ and contact $\mathcal{I}$ for the RGB image, while our TriDi recovers the object $\mathcal{O}$, showing generalization on unseen data distributions.

unseen geometries and report samples from a more powerful model trained on more datasets in the Sup. Mat., showing scalability to a variety of objects (e.g., vacuum, umbrella, skateboard) and interactions (e.g., feet).

Populating scenes. Populating scenes is interesting for several downstream tasks like those from AR/VR, or for synthetic data generations. Here, we demonstrate one possible way to populate scenes with TriDi. We first place a virtual object or a human on the ground of a scene, the publicly available HPS dataset[22], and then run TriDi to generate the complementary modality. Also, we can directly sample $p(\mathcal{H}, \mathcal{O}, \mathcal{I})$ and populate with both humans and objects. Results are shown in Fig. 6.

Interaction reconstruction. Our method can also be used to reconstruct interactions from images indirectly. In Fig.7, we provide an example from the DAMON dataset of DECO [70]. DAMON provides contact annotations for images from the HOT [9] dataset along with SMPL parameters estimated by CLIFF [45]. Remarkably, TriDi generalizes to such cases despite not being trained on the DECO dataset. More examples are included in Sup. Mat.

Sequences with keyframing. Perhaps the most widespread mechanism to generate sequences in animation is via keyframing. Here, we can use TriDi to automatically generate the keyframes of interaction and use an off-the-shelf in-betweener to generate a sequence. Given a 4D human sequence at 30 fps, we sample at 1 fps, generating the object and interpolate frames in between using slerp for angles and linear for translations. We provide an example in our Sup. Mat. video.

6. Conclusions

In this work, we proposed TriDi, the first joint model for Human, Object, and Interaction, modeling it as a three-variable joint distribution and handling a total of seven different operation modes. Such versatility of TriDi renders prior works as special use cases of the proposed method. TriDi employs an original Contact-Text Interaction representation that combines the interpretability of text with the guidance from the contact information. Quantitative comparisons demonstrated the superiority of the proposed method both in terms of distribution quality and spatial consistency, with an improvement up to *47%* over the baseline methods. Finally, we demonstrated the applicability of the proposed method to scene population, interaction reconstruction from partial data, and generalization to novel object geometry. This paves the way for unified HOI usage in content creation, data generation, and AR/VR in the future.

Limitations and future work. There are several exciting avenues for future work. We consider scaling beyond single human and object interaction a promising direction to enable the modeling of realistic social situations with increasing complexity. The recent advancements in data capturing [34, 39, 94] open up a possibility of blending scene and object conditioning, laying the foundation for advanced HOI models. As a data-driven method, TriDi is sensitive to the skewness of the data distribution, expressing more variety towards frequent objects. Although TriDi shows generalization to unseen geometries (e.g., chairs and stools), we do not expect it to support objects with significantly novel functionality (e.g., wheelchairs, bicycles, bowling balls).

Acknowledgements Special thanks to Garvita Tiwari, Nikita Kister, and Xianghui Xie for the helpful discussions. We also thank RVH team for their help with proofreading the manuscript. This work is funded by the Deutsche Forschungsgemeinschaft - 409792180 (EmmyNoether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting I. A. Petrov. R. Marin has been supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101109330. The project was made possible by funding from the Carl Zeiss Foundation. J. Chibane is a fellow of the Meta Research PhD Fellowship Program - area: AR/VR Human Understanding.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 6

[2] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 2, 3

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and Method for Tracking Human Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 3, 6, 7

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 2

[5] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 3

[6] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pages 361–378. Springer, 2020. 3

[7] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 464–473. IEEE, 2024. 2

[8] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5887–5895, 2021. 2

[9] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17100–17110, 2023. 3, 8

[10] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. In *European Conference on Computer Vision*. Springer, 2024. 3

[11] Xinhua Cheng, Tianyu Yang, Jianan Wang, Yu Li, Lei Zhang, Jian Zhang, and Li Yuan. Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. *arXiv preprint arXiv:2310.11784*, 2023. 3

[12] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions. *arXiv:2112.03028 [cs]*, 2022. 2

[13] Sisi Dai, Wenhao Li, Haowen Sun, Haibin Huang, Chongyang Ma, Hui Huang, Kai Xu, and Ruizhen Hu. Interfusion: Text-driven generation of 3d human-object interaction. In *European Conference on Computer Vision*. Springer, 2024. 3

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 5

[15] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024. 3, 4

[16] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[17] Thomas Feix, Ian M. Bullock, and Aaron M. Dollar. Analysis of Human Grasping Behavior: Object Characteristics and Grasp Type. *IEEE Transactions on Haptics*, 7(3):311–323, 2014. 2

[18] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The GRASP Taxonomy of Human Grasp Types. *IEEE Transactions on Human-Machine Systems*, 46(1):66–77, 2016. 2

[19] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12. Wiley Online Library, 2023. 3

[20] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 6

[21] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing Contact to Improve Grasps. In *2021*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1471–1481, 2021. 3

[22] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 8

[23] Vladimir Guzov, Ilya A Petrov, and Gerard Pons-Moll. Blendify–python rendering framework for blender. *arXiv preprint arXiv:2410.17858*, 2024. 1

[24] Sookwan Han and Hanbyul Joo. Chorus: Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15835–15846, 2023. 3

[25] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 2

[26] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 2

[27] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D Scenes by Learning Human-Scene Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 2, 3

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 1

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 5

[30] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 2

[31] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022. 3, 6

[32] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021. 2

[33] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023. 3

[34] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 2, 8

[35] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2022. 3

[36] Tero Karras. Maximizing parallelism in the construction of bvhs, octrees, and k-d trees. In *Proceedings of the Fourth ACM SIGGRAPH/Eurographics Conference on High-Performance Graphics*, pages 33–37, 2012. 5

[37] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning Implicit Representations for Human Grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344, 2020. 3

[38] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In *European Conference on Computer Vision*. Springer, 2024. 3

[39] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions, 2024. 3, 8

[40] Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. Ncho: Unsupervised learning for neural 3d composition of humans and objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14817–14828, 2023. 3

[41] 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. *arXiv preprint arXiv:2307.07511*, 2023. 2

[42] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2, 4, 6

[43] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*. Springer, 2024. 2

[44] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3035–3044, 2024. 2

[45] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 8

[46] C Karen Liu. Dextrous manipulation from a grasping pose. In *ACM SIGGRAPH 2009 papers*, pages 1–6. 2009. 2

[47] Siqi Liu, Yong-Lu Li, Zhou Fang, Xinpeng Liu, Yang You, and Cewu Lu. Primitive-based 3d human-object interaction modelling and programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3711–3719, 2024. 3

[48] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters*, 7(1):470–477, 2021. 2, 5

[49] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 3

[50] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 4

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[52] Haowen Luo, Yunze Liu, and Li Yi. Physics-aware hand-object interaction denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2341–2350, 2024. 2

[53] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. *arXiv preprint arXiv:2304.02061*, 2023. 2

[54] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *arXiv preprint arXiv:2306.09337*, 2023. 3

[55] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2024. 3

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[57] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 2

[58] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4726–4736, 2023. 2, 3, 6, 7

[59] Huaijin Pi, Sida Peng, Minghui Yang, Xiaowei Zhou, and Hujun Bao. Hierarchical generation of human-object interactions with diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15061–15073, 2023. 2

[60] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35:23192–23204, 2022. 4

[61] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[62] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 4

[63] Roman Shapovalov, Yanir Kleiman, Ignacio Rocco, David Novotny, Andrea Vedaldi, Changan Chen, Filippos Kokkinos, Ben Graham, and Natalia Neverova. Replay: Multi-modal multi-view acted videos for casual holography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20338–20348, 2023. 3

[64] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[65] Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2024. 2, 3

[66] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *Computer Vision – ECCV 2020*, pages 581–600. Springer International Publishing, Cham, 2020. 6, 7, 3

[67] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13273, 2022. 2, 6, 7

[68] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using spatial cues and latent consistency. In *2024 International Conference on 3D Vision (3DV)*, pages 933–943. IEEE, 2024. 2

[69] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023. 3

[70] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the

wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8001–8013, 2023. 2, 3, 8, 4

[71] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118: 172–193, 2016. 5, 6

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[73] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 2

[74] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12206–12215, 2021. 2

[75] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366. IEEE, 2023. 2, 5, 6

[76] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *2022 International Conference on 3D Vision (3DV)*, pages 353–362. IEEE, 2022. 3

[77] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2021. 3

[78] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[79] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3

[80] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4757–4768, 2023. 3

[81] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10015, 2024. 2, 3

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

[83] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4541–4550, 2019. 6

[84] Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Siyuan Huang. F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. In *European Conference on Computer Vision*. Springer, 2024. 3

[85] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16284–16295, 2024. 3

[86] Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. Scene synthesis from human motion. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[87] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012. 2

[88] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-Aware Object Placement for Visual Environment Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3959–3970, 2022. 2, 3

[89] Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12965–12976, 2023. 3

[90] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*. Springer, 2024. 2, 3

[91] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023. 3

[92] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 2

[93] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8845, 2023. 3

[94] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 516–526, 2024. 3, 8

[95] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European conference on computer vision*, pages 34–51. Springer, 2020. 3

[96] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity Learning of Articulation and Contact in 3D Environments. In *2020 International Conference on 3D Vision (3DV)*, pages 642–651, 2020. 2, 3

[97] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[98] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. Force: Dataset and method for intuitive physics guided human-object interaction. *arXiv preprint arXiv:2403.11237*, 2024. 2

[99] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6194–6204, 2020. 2

[100] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–741, 2024. 3

[101] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 2

[102] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20634–20643, 2024. 2

[103] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 1

TriDi: Trilateral Diffusion of 3D Humans, Objects, and Interactions

Supplementary Material

Abstract

This supplementary material provides summary of notation used in the text in Sec. 7. We report further implementation details of TriDi, description of text labels annotation, insights on symmetry augmentation, and training losses in Sec. 8. In Sec. 9, we include details on the conducted user study, qualitative results on unseen data, ablation results, qualitative results on GRAB, BEHAVE, OMOMO, and InterCap, as well as extended qualitative comparison with the baselines. In Sec. 10, we include a discussion on the broader impacts of our work. Details on all four datasets used in the experiments are summarized in Sec. 11. Sec. 12 introduces an optional post-processing refinement procedure that increases the realism of the generated interactions. Finally, in Sec. 13, we provide full definition of the error metrics. In the attached video, we show results of the keyframing animation discussed in the main text, as well as additional qualitative examples, and we encourage the reader to look at the video.

7. Background and Notation

Background. We follow the formulation of Denoising Diffusion Probabilistic Model (DDPM) [28] to obtain a closed-form expression for $\mathbf{z}_t$ given the original sample $\mathbf{z}_0$. Let $\alpha_i = 1 - \beta_i$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}),$$
$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{14}$$

An iterative denoising process with denoising network $\mathcal{D}_\psi$ is defined by the following:

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathcal{D}_\psi(\mathbf{z}_t; c, t) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon, \tag{15}$$

where $\hat{\mathbf{z}}_0 = \mathcal{D}_\psi(\mathbf{z}_t; c, t)$.

Notation. Tab. 3 defines symbols used in our work.

Symbol	Description	Domain
$\mathcal{H}$	Human Modality	$(\theta_\mathcal{H}, \beta_\mathcal{H}, \mathbf{g}_\mathcal{H})$
$\theta_\mathcal{H}$	Human Pose	$\mathbb{R}^{51\times3}$
$\beta_\mathcal{H}$	Human Identity	$\mathbb{R}^{10}$
$\mathbf{V}_\mathcal{H}$	Human Template's Vertices	$\mathbb{R}^{690}$
$\mathbf{g}_\mathcal{H}$	Human Global Pose in 6-DoF	$\mathbb{R}^{9}$
$\mathbf{d}$	Human to Object vertex distance	$\mathbb{R}^{690}$
$\mathcal{O}$	Object Modality	$(\mathbf{g}_\mathcal{O})$
$\mathbf{g}_\mathcal{O}$	Object Global Pose in 6-DoF	$\mathbb{R}^{9}$
$\mathcal{C}_\mathcal{O}$	Object Information for conditioning	$(\mathbf{f}_\mathcal{O}, \mathbf{y}_\mathcal{O})$
$\mathbf{f}_\mathcal{O}$	PointNext features object	$\mathbb{R}^{1024}$
$\mathbf{y}_\mathcal{O}$	one-hot encoding of the class	$\{0,1\}^{40}$
$\mathbf{V}_\mathcal{O}$	Object Template's Vertices	$\mathbb{R}^{1500}$
$\mathcal{I}$	Interaction	$(\mathbf{z}_\mathcal{I})$
$T_\mathcal{I}$	Interaction Textual Label	text
$\mathbf{z}_\mathcal{I}$	Interaction latent representation	$\mathbb{R}^{128}$
$\phi_\mathcal{I}$	Interaction contact map	$\{0,1\}^{690}$
$E_{\phi_\mathcal{I}}$	Interaction Encoder (Contact Map)	$\phi_\mathcal{I} \mapsto \mathbf{z}_\mathcal{I}$
$D_{\phi_\mathcal{I}}$	Interaction Decoder (Contact Map)	$\mathbf{z}_\mathcal{I} \mapsto \phi_\mathcal{I}$
$E_{T_\mathcal{I}}$	Interaction Encoder (Textual Label)	$T_\mathcal{I} \mapsto \mathbf{z}_\mathcal{I}$

Table 3. **Notation Table**. The main notation used in our paper.

8. Implementation details

The denoising network has a total of 15M parameters, and it is trained end-to-end. We use a batch size of 1024, a learning rate of $1e-4$ with a cosine scheduler, and warm up the training during the first 50k steps. The parameters are optimized with AdamW [51]. We train for a total of 300k steps. All the experiments are performed on a machine with RTX4090 GPU. The training of the model takes approximately 20 hours. The contact encoder-decoder network with 1.7M parameters is trained separately for 70 epochs, converging on the same machine in $\sim$ 1 hour. The inference for one example with diffusion guidance takes around 3.07 seconds. Since TriDi works per-frame the inference can be majorly sped up using batching, e.g. inference time for 1024 examples in one batch is 38.79 s. All models are implemented in PyTorch [56] framework. Following [103] we convert all rotations $(\theta_\mathcal{H}, \mathbf{g}_\mathcal{H}, \mathbf{g}_\mathcal{O})$ to 6-d representations before passing them to the network. We rely on blendify [23] for visualization.

We implement diffusion reconstruction guidance within DDPM pipeline and apply it for the last 200 out of 1000 iterations of the denoising process with weight $\lambda = 2.0$.

Text labels annotation. During training, we use a set of predefined templates to generate text labels on the fly, making the encoder $E_{T_\mathcal{I}}$ more robust to diverse text inputs. The template is selected randomly from a pool (provided in Listing 1) based on which body parts are in contact with the object and the object's class. For example, if a person sits on a chair, then the text label is selected from a set of *1. Generic templates* and *2.2 Sitting templates*.

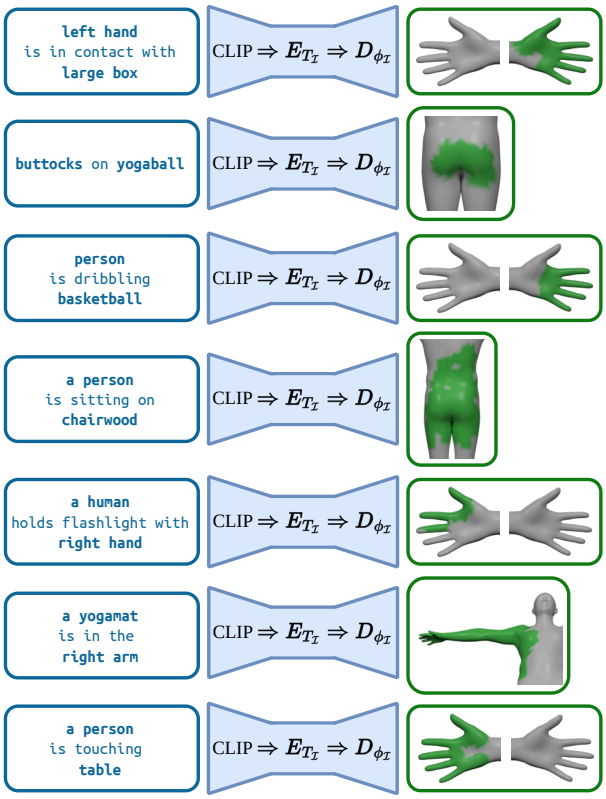

Figure 8. **Contact maps**. Examples of contact maps decoded from text queries.

Augmentation. During training, we apply the symmetry augmentation randomly mirroring samples through ZY plane. As a result, the model exhibits less bias towards right-handed interactions. Qualitative examples in Fig. 9 for both cases of sampling from $p(\mathcal{H}, \mathcal{I}|\mathcal{O})$ and $p(\mathcal{O}, \mathcal{I}|\mathcal{H})$ demonstrate how TriDi generates left- and right-handed interactions given the same condition.

Losses The objective function used to train our network is the weighted combination of the following losses:

$$\begin{aligned} L_n^\mathcal{H} &= \|\theta_\mathcal{H} - \hat{\theta}_\mathcal{H}\|_1 + \|\beta_\mathcal{H} - \hat{\beta}_\mathcal{H}\|_1 + \|\mathbf{g}_\mathcal{H} - \hat{\mathbf{g}}_\mathcal{H}\|_1 \\ L_n^\mathcal{O} &= \|\mathbf{g}_\mathcal{O} - \hat{\mathbf{g}}_\mathcal{H}\|_1 \\ L_n^\mathcal{I} &= \|\mathbf{z}_\mathcal{I} - \hat{\mathbf{z}}_\mathcal{I}\|_2 \\ L_v^\mathcal{H} &= \|\mathbf{V}_\mathcal{H} - \hat{\mathbf{V}}_\mathcal{H}\|_2 \\ L_v^\mathcal{O} &= \|\mathbf{V}_\mathcal{O} - \hat{\mathbf{V}}_\mathcal{O}\|_2 \\ L_v^\mathcal{I} &= \|\mathbf{d} - \hat{\mathbf{d}}\|_2 \end{aligned} \tag{16}$$

The resulting loss function is:

$$\begin{aligned} L_{TriDi} = &\lambda_n^\mathcal{H} L_n^\mathcal{H} + \lambda_n^\mathcal{O} L_n^\mathcal{O} + \lambda_n^\mathcal{I} L_n^\mathcal{I} + \\ &\lambda_v^\mathcal{H} L_v^\mathcal{H} + \lambda_v^\mathcal{O} L_v^\mathcal{O} + \lambda_v^\mathcal{I} L_v^\mathcal{I} \end{aligned} \tag{17}$$

with weighting coefficients set to: $\lambda_n^\mathcal{H} = \lambda_v^\mathcal{O} = 2, \lambda_n^\mathcal{O} = \lambda_n^\mathcal{I} = 1, \lambda_v^\mathcal{H} = 6, \lambda_v^\mathcal{I} = 4$.

```
1. Generic templates:
  - <body parts> <is / are> in contact
    with <object class>
  - <object class> is in contact
    with <body parts>
  - <body parts> touch(-es) <object class>
  - <object class> <touches> <body parts>

2. Interaction specific templates:
2.1 Basketball template
  - a person is dribbling basketball

2.2 Sitting templates
  - <body parts> <is / are> on <object class>
  - a person <is / sits> on <object class>

2.3 Hands-only templates
  - <object class> is in <body parts>
  - <body parts> <hold(-s) / grab(-s)>
    <object class>
  - a person is <holding / grabbing / carrying>
    <object class>
```

Listing 1. **Text labels**. All templates used during training.

9. Additional Experiments

User study This section introduces details on the user study that was used to evaluate TriDi. We have designed and run a user study, asking participants to rate the quality of the generated interactions. We compared TriDi against one baseline method and ground-truth data in two generation modes: $p(\mathcal{H}, \mathcal{I}|\mathcal{O})$ and $p(\mathcal{O}, \mathcal{I}|\mathcal{H})$. We used GNet and ObjPOP+cVAE as the baselines, and randomly selected 10 queries for the generation (5 from each of BEHAVE and GRAB) for each mode. In every question we show users three randomly shuffled samples: ground-truth data, TriDi, and corresponding baseline. The participants were asked to rate the quality of each sample based on the realism of human-object interaction, and the amount of interpenetration between human and object. The rating scale consisted of three options: *Worst*, *Moderate*, and *Best*, with ratings being non-exclusive (i.e., more than one sample can have a similar rating). Example interface of the user study is provided in the Fig. 10. As a result, we have collected 40 responses. We summarize the results in the Tab. 4, comparing the ratings assigned to the samples by users. On average, results of TriDi were preferred to the baselines in 81.4% of the cases and preferred to the ground-truth examples in 46.65% of the cases. This suggests that the results of TriDi are more appreciable than the baselines and produce a realism comparable to captured data.

Generalization to unseen data . We provide qualitative examples of TriDi on eight unseen objects in two sampling

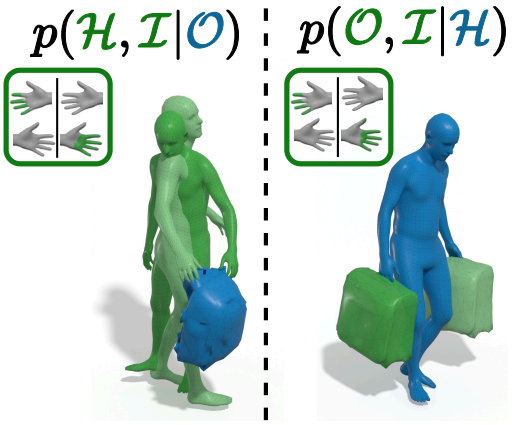

Figure 9. **Qualitative examples**. Results demonstrating the effectiveness of the symmetry augmentation. TriDi generates left- and right-handed interactions given the same condition.

Mode	Rating comparison	Result in %
$p(\mathcal{H}, \mathcal{I}\|\mathcal{O})$	TriDi > GNet	82.0%
	TriDi > GT data	50.3%
$p(\mathcal{O}, \mathcal{I}\|\mathcal{H})$	TriDi > ObjPOP+cVAE	80.8%
	TriDi > GT data	43.0%

Table 4. **User study**. Summary of the user study results.

modes in Fig. 11. The model is able to generate realistic interactions for objects with known functionality. We also include more examples for interaction reconstruction on the DAMON dataset in Figure 12.

Ablations Here, we report the quantitative evaluations of our ablations described in the main paper. Table 5 covers the quality of the generated distributions, while Table 6 covers geometrical consistency of the generation.

Qualitative results This section includes additional qualitative results on BEHAVE (Figure 16) and GRAB (Figure 17), and introduces examples from InterCap (Figure 14) and OMOMO (Figure 15).

Comparison with baselines In Fig. 18 we provide an extended comparison with baselines, showing 3 generated samples per same input.

10. Broader Impacts

Our method provides an invaluable tool for general content creation and supports analysis of different disciplines like behavioral sciences or ergonomic studies. Since our method studies human interaction, analysis of subjects' behavior may be included in surveillance applications, leading to privacy issues. However, at the present date, acquiring

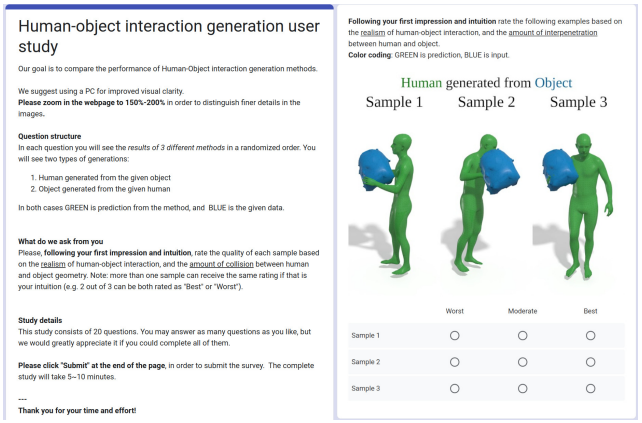

Figure 10. **User study**. The interface of the user study.

the 3D data used in our method cannot be easily done without the consensus of the target subject.

11. Datasets

BEHAVE. BEHAVE [3] captures 8 subjects interacting with 20 different objects, represented as SMPL+H meshes and global configuration, respectively. We downsample the $30fps$ train sequences to $10fps$ and consider the official $1fps$ test subset.

GRAB. We use the subset of GRAB [66] introduced in [58]. This subset includes 10 subjects interacting with 20 objects. The $120fps$ train and test sequences are downsampled to $1fps$. The test set consists of interactions performed by subjects 9 and 10.

InterCap. We downsample the original $30fps$ sequences to $10fps$ and follow the train-test split provided by VisTracker [80]: Data from subjects 1-8 is used for training, and sequences from subjects 9 and 10 are used for evaluation.

OMOMO. This dataset captures 17 humans interacting with 15 objects. We employ the official split, using the first 15 subjects for training and subjects 16,17 for testing, and downsample all the sequences to $10fps$.

12. Post-processing refinement

Motivation. In some cases, TriDi's samples may miss perfect plausibility of fine grained details, especially for smaller objects. Such behavior is naturally caused by a lack of detailed hand modeling in the majority of the training data. To counter this problem, we introduce a post-processing refinement. We demonstrate qualitative exam-

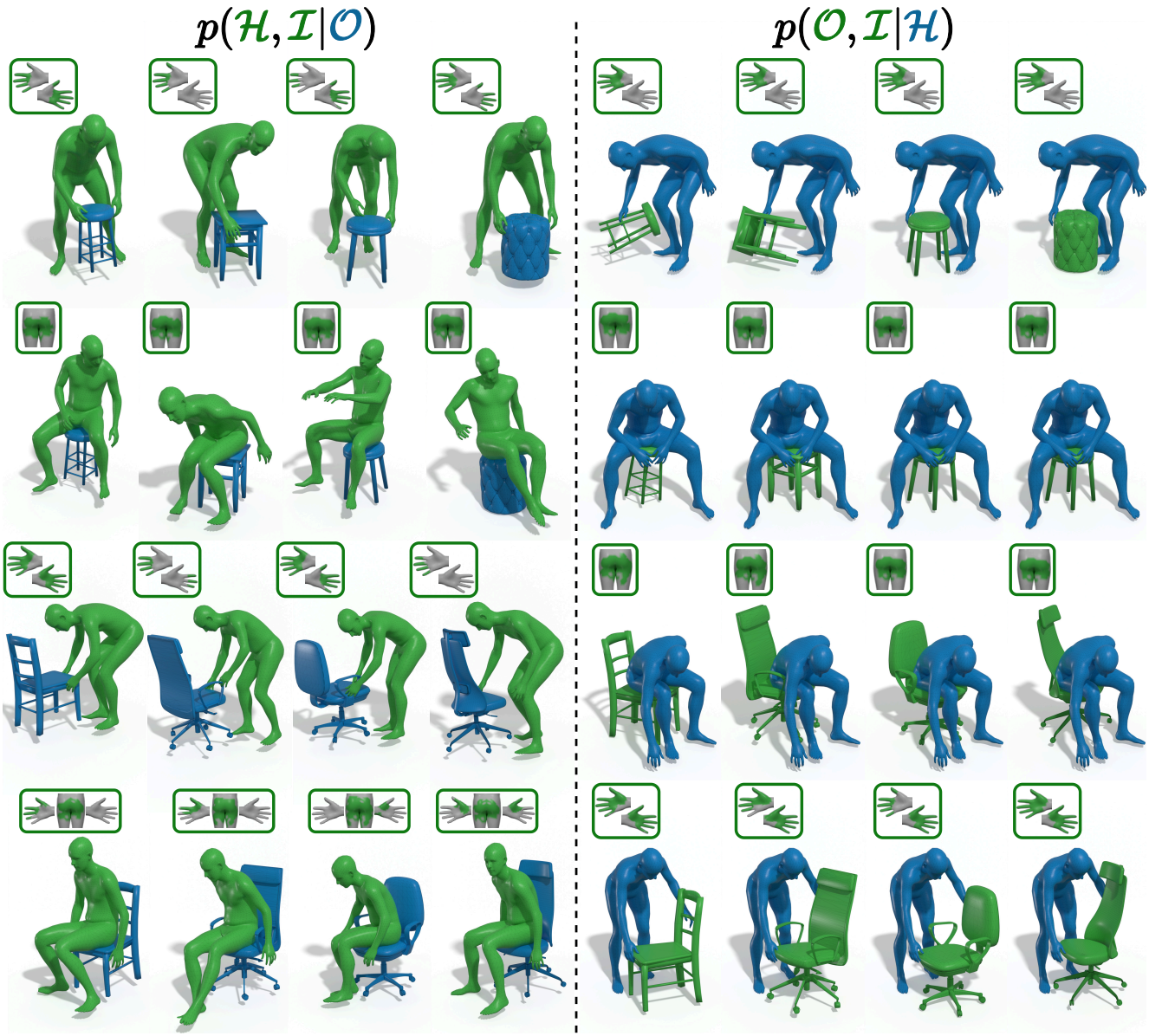

Figure 11. **Generalization to unseen geometry**. TriDi samples from $p(\mathcal{H},\mathcal{I}|\mathcal{O})$ and $p(\mathcal{O},\mathcal{I}|\mathcal{H})$ with unseen objects.

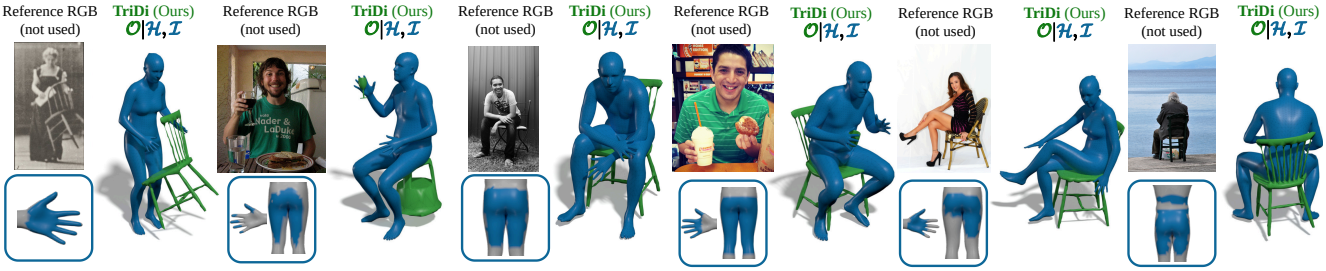

Figure 12. **Interaction reconstruction**. DECO [70] annotates human $\mathcal{H}$ and contact $\mathcal{I}$ for the RGB image, while our TriDi recovers the object $\mathcal{O}$, showing generalization on unseen data distributions.

Method	BEHAVE $\mathcal{H},\mathcal{I}\|\mathcal{O}$			BEHAVE $\mathcal{O},\mathcal{I}\|\mathcal{H}$		
	1-NNA ($\rightarrow$ 50)	COV↑	MMD↓	1-NNA ($\rightarrow$ 50)	COV↑	MMD↓
TriDi	**67.89**$^{\pm 0.3}$	47.81$^{\pm 0.2}$	**1.352**$^{\pm 0.005}$	**63.72**$^{\pm 0.3}$	**51.71**$^{\pm 0.1}$	**0.166**$^{\pm 0.001}$
NoGuide	68.04$^{\pm 0.5}$	**48.87**$^{\pm 0.2}$	1.355$^{\pm 0.002}$	63.80$^{\pm 0.4}$	51.62$^{\pm 0.3}$	0.167$^{\pm 0.001}$
$(\mathcal{H},\mathcal{O})$	68.19$^{\pm 0.4}$	48.57$^{\pm 0.1}$	1.373$^{\pm 0.006}$	65.18$^{\pm 0.5}$	50.85$^{\pm 0.2}$	**0.166**$^{\pm 0.001}$
NoAug	69.74$^{\pm 0.3}$	46.21$^{\pm 0.3}$	1.409$^{\pm 0.009}$	69.39$^{\pm 0.3}$	46.20$^{\pm 0.3}$	0.184$^{\pm 0.002}$
Method	**GRAB** $\mathcal{H},\mathcal{I}\|\mathcal{O}$			**GRAB** $\mathcal{O},\mathcal{I}\|\mathcal{H}$		
	1-NNA ($\rightarrow$ 50)	COV↑	MMD↓	1-NNA ($\rightarrow$ 50)	COV↑	MMD↓
TriDi	82.71$^{\pm 0.5}$	42.76$^{\pm 0.3}$	0.930$^{\pm 0.012}$	**65.02**$^{\pm 0.7}$	48.84$^{\pm 1.2}$	0.268$^{\pm 0.011}$
NoGuide	82.99$^{\pm 0.5}$	41.74$^{\pm 1.0}$	0.957$^{\pm 0.007}$	65.64$^{\pm 0.4}$	47.98$^{\pm 1.3}$	0.269$^{\pm 0.012}$
$(\mathcal{H},\mathcal{O})$	**82.40**$^{\pm 1.0}$	42.53$^{\pm 1.2}$	0.996$^{\pm 0.014}$	66.58$^{\pm 1.7}$	**49.23**$^{\pm 0.4}$	**0.262**$^{\pm 0.002}$
NoAug	83.05$^{\pm 1.0}$	**43.78**$^{\pm 0.6}$	**0.878**$^{\pm 0.012}$	67.38$^{\pm 0.3}$	46.11$^{\pm 0.3}$	0.275$^{\pm 0.006}$

Table 5. **Ablation - Quality of Generated Distribution**. Impact of augmentation, $\mathcal{I}$ diffusion, and guidance.

Method	BEHAVE $\mathcal{H},\mathcal{I}\|\mathcal{O}$			BEHAVE $\mathcal{O},\mathcal{I}\|\mathcal{H}$		
	MPJPE↓	MPJPE-PA↓	Acc_{cont} ↑	E_{v2v} ↓	E_{center} ↓	Acc_{cont} ↑
TriDi	**20.8**	**12.3**	95.5/**96.5**	**28.0**	**15.3**	95.9/96.1
NoGuide	21.5	12.4	**96.0**/**96.5**	28.1	15.4	**96.2**/**96.2**
$(\mathcal{H},\mathcal{O})$	21.9	12.7	**96.0** / NA	28.4	15.6	96.1 / NA
NoAug	23.2	12.9	95.4 / 96.2	31.0	17.8	95.5 / 96.0
Method	**GRAB** $\mathcal{H},\mathcal{I}\|\mathcal{O}$			**GRAB** $\mathcal{O},\mathcal{I}\|\mathcal{H}$		
	MPJPE↓	MPJPE-PA↓	Acc_{cont} ↑	E_{v2v} ↓	E_{center} ↓	Acc_{cont} ↑
TriDi	15.3	11.1	98.0/98.3	**6.9**	**5.0**	**99.0**/98.2
NoGuide	16.2	11.3	97.5 / 98.3	9.0	7.5	98.2 / 98.3
$(\mathcal{H},\mathcal{O})$	17.3	11.8	97.3 / NA	9.5	7.9	98.0 / NA
NoAug	**14.1**	**10.4**	**98.2**/**98.4**	7.2	5.2	98.9 / **98.5**

Table 6. **Ablation - Geometrical Consistency of Generation**. Impact of augmentation, $\mathcal{I}$ diffusion, and guidance.

ples of post-processing refinement in Fig. 13 to show extended capabilities of TriDi. The proposed refinement procedure is able to correct mistakes in fine-grained grasps leading to increased realism of predictions. In the following paragraphs we provide details on the post-processing refinement. We remark that all the qualitative and quantitative results in the main paper and supplementary are obtained without the refinement for a fairer comparison.

Refinement implementation. We take inspiration from DexGraspNet [75] to design an optimization procedure refining the generated hands. The original refinement minimizes the error term:

$$E_{fc} + w_{dis}E_{dis} + w_{pen}E_{pen} + w_{spen}E_{spen} + w_{prior}E_{prior} \tag{18}$$

where E_{fc} is a force closure term proposed in [48] that encourages the closed grasp, E_{dis} and E_{pen} are, respectively, attraction and repulsion terms, enforcing contact and penalizing penetration, E_{spen} is a self-penetration term, E_{prior} is a hand prior term penalizing unrealistic pose configurations. We refer to [75] for detailed definition of the energies. We add two more terms to the original energy to adapt the method to our use case. Firstly, we want the final result to don't deviate too much from the initial prediction of TriDi, thus we introduce regularization:

$$E_{reg} = \|\hat{\theta_{\mathcal{H}}} - \tilde{\theta_{\mathcal{H}}}\|_2 \tag{19}$$

where $\hat{\theta_{\mathcal{H}}}$ is human pose predicted by TriDi and $\tilde{\theta_{\mathcal{H}}}$ is the refined human pose. Secondly, we want to explicitly penalize intersections between hands and objects. To achieve this we introduce a term inspired by [36, 71] that detects the collision between hand and object meshes, penalizing the quantity:

$$E_{isect} = \sum_{(\mathbf{f}_{\mathcal{H}},\mathbf{f}_{\mathcal{O}})\in C} \left[\sum_{\mathbf{v}_{\mathcal{H}}\in\mathbf{f}_{\mathcal{H}}} \| - \Psi_{\mathbf{f}_{\mathcal{O}}}(\mathbf{v}_{\mathcal{H}})\|^2 + \sum_{\mathbf{v}_{\mathcal{O}}\in\mathbf{f}_{\mathcal{O}}} \| - \Psi_{\mathbf{f}_{\mathcal{H}}}(\mathbf{v}_{\mathcal{O}})\|^2 \right] \tag{20}$$

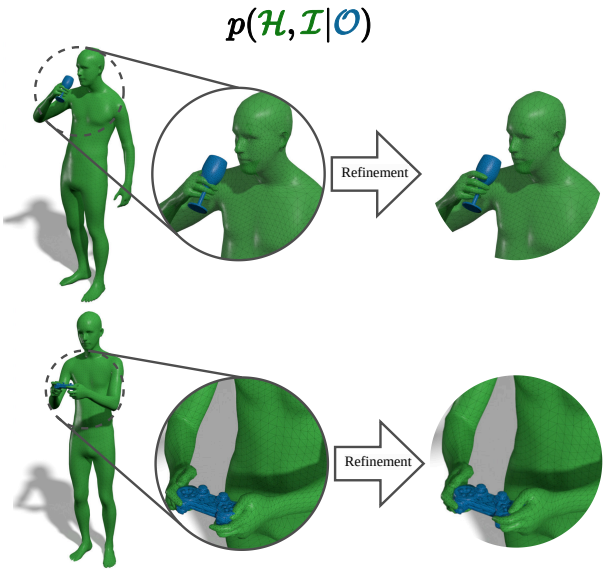

Figure 13. **Post-processing refinement result**. Example results demonstrating the effectiveness of the post-processing refinement. Optionally, TriDi results can be refined using an optimization procedure that improves fine hand details.

where $\mathbf{v}_{\mathcal{H}} \in \mathbf{V}_{\mathcal{H}}$ and $\mathbf{f}_{\mathcal{H}} \in \mathbf{F}_{\mathcal{H}}$ are vertices and faces of the human mesh, $\mathbf{v}_{\mathcal{O}} \in \mathbf{V}_{\mathcal{O}}$ and $\mathbf{f}_{\mathcal{O}} \in \mathbf{F}_{\mathcal{O}}$ are vertices and faces of the object mesh, C is a set of pairs of collided faces, $\Psi_{\mathbf{f}} : \mathbb{R}^3 \rightarrow \mathbb{R}_+$ is a cone distance field from the face $\mho$ (full definition can be found in [71]).

Since TriDi deals with full bodies, the optimization procedure is split into two stages: first, to fix the global positioning of the hand (optimization w.r.t. shoulder, elbow, and wrist joints), next to fix the fine details (optimization w.r.t. fingers). Therefore, we obtain the following energy terms:

$$\begin{aligned} E_{stage\_1} =& w_{dis}E_{dis} + w_{pen}E_{pen} + \\ & w_{reg}E_{reg} + w_{isect}E_{isect} \\ E_{stage\_2} =& E_{fc} + w_{dis}E_{dis} + w_{pen}E_{pen} + \\ & w_{spen}E_{spen} + w_{prior}E_{prior} + \\ & w_{reg}E_{reg} + w_{isect}E_{isect} \end{aligned} \tag{21}$$

where weights are $w_{dis} = 0.2$, $w_{pen} = 100$, $w_{reg} = 20$, $w_{isect} = 400$ for the first stage, and $w_{dis} = w_{pen} = w_{isect} = 100$, $w_{spen} = 10$, $w_{prior} = 0.5$, $w_{reg} = 10$ for the second stage. Optimization setup follows [75] with 1000 iterations for the first stage and 2000 iterations for the second stage.

13. Error Metrics

Quality of Generated Distribution. To evaluate our fitting to the target distribution, we use three metrics. The *Coverage (COV)*[1]:

$$COV(S_g, S_r) = \frac{|\{\arg\min_{r \in S_r} D(g,r) | g \in S_g\}|}{|S_r|}, \tag{22}$$

where $D(g,r)$ is $L2$ distance between corresponding feature vectors, namely, root-centered body joints for humans and concatenated global position and orientation for objects.

Minimum Matching Distance (MMD)[1]:

$$MMD(S_g, S_r) = \frac{1}{|S_r|} \sum_{r \in S_r} \min_{g \in S_g} D(g,r) \tag{23}$$

We employ the same definition of $D(\cdot,\cdot)$ as for COV.

1-Nearest Neighbor Accuracy (1-NNA) [83]. Given a generated sample g, The idea is to evaluate how a 1-NN classifier trained on $S_{-g} = S_r \cup S_g - \{g\}$ would classify the sample g. Namely, 1-NNA evaluates the leave-one-out accuracy over the union dataset:

$$\begin{aligned} &\text{1-}NNA(S_g, S_r) = \\ &\frac{\sum_{X \in S_g} \mathbb{1}[N_X \in S_g] + \sum_{Y \in S_r} \mathbb{1}[N_Y \in S_r]}{|S_g| + |S_r|}, \end{aligned} \tag{24}$$

where N_X is the nearest neighbor of X in S_{-X}, $\mathbb{1}[\cdot]$ is the indicator function. We define nearest neighbors according to the aforementioned distance metrics $D(\cdot,\cdot)$.

Geometrical Consistency of Generation. The E_{v2v} error measures the average L2 distance between the position of the predicted object vertices and the ones of the ground truth:

$$E_{v2v} = \frac{1}{|\mathbf{V}_{\mathcal{O}}|} \sum_{i \in |\mathbf{V}_{\mathcal{O}}|} \|\mathbf{V}^i_{\mathcal{O}} - \hat{\mathbf{V}}^i_{\mathcal{O}}\|_2 \tag{25}$$

The E_c error measures the average L2 distance between the position of the predicted object center and the one of the ground truth:

$$E_c = \left\| \frac{1}{|\mathbf{V}_{\mathcal{O}}|} \sum_{i \in |\mathbf{V}_{\mathcal{O}}|} \mathbf{V}^i_{\mathcal{O}} - \frac{1}{|\hat{\mathbf{V}}_{\mathcal{O}}|} \sum_{i \in |\hat{\mathbf{V}}_{\mathcal{O}}|} \hat{\mathbf{V}}^i_{\mathcal{O}} \right\|_2. \tag{26}$$

We complement the reconstruction metrics with the contact accuracy metric Acc_{cont}:

$$Acc_{cont} = \frac{1}{|\mathbf{V}_{\mathcal{H}}|} \sum_{i \in |\mathbf{V}_{\mathcal{H}}|} \mathbb{1}[\hat{\phi}^i_{\mathcal{I}} = \phi^i_{\mathcal{I}}], \tag{27}$$

where $\mathbb{1}$ is an indicator function.

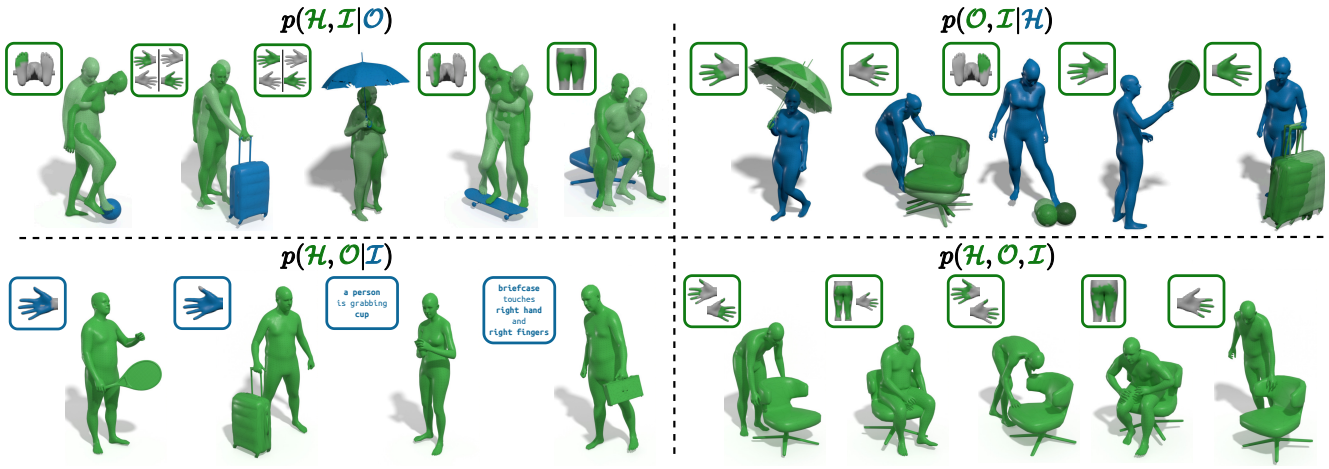

Figure 14. **Qualitative results of TriDi on InterCap**.

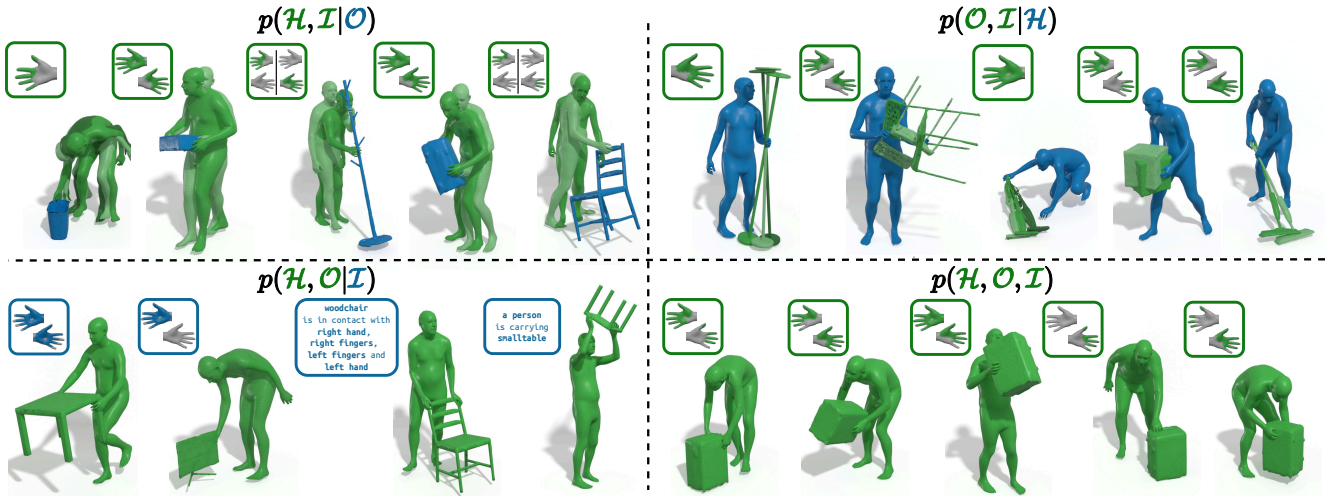

Figure 15. **Qualitative results of TriDi on OMOMO**.

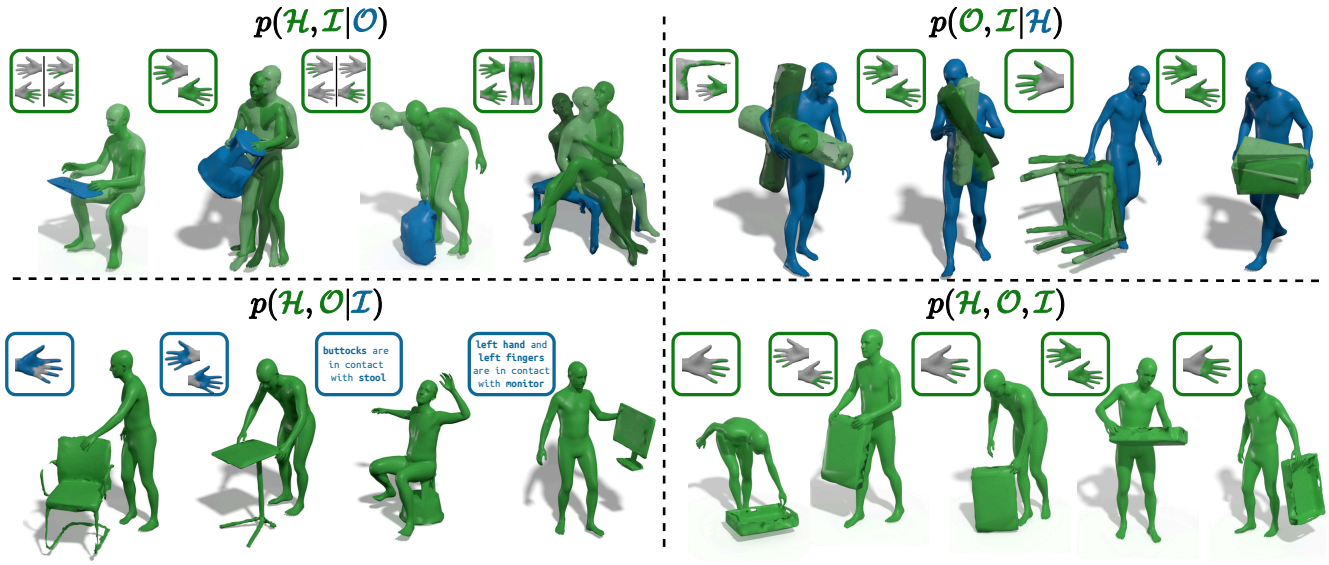

Figure 16. **Qualitative results of TriDi on BEHAVE**.

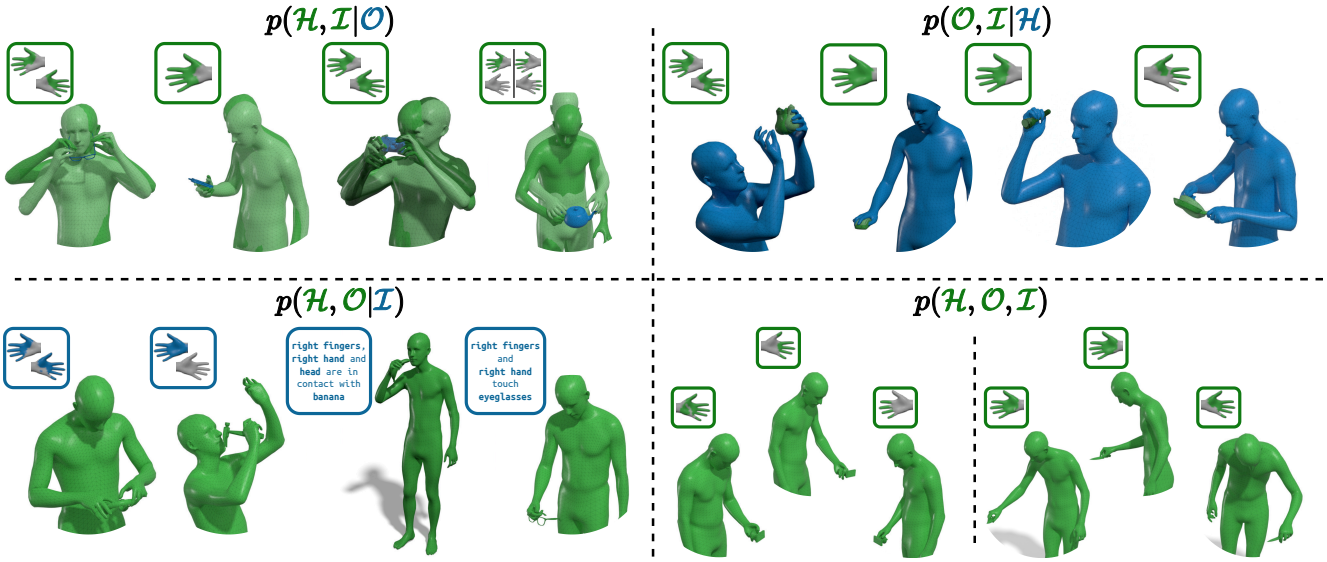

Figure 17. **Qualitative results of TriDi on GRAB**.

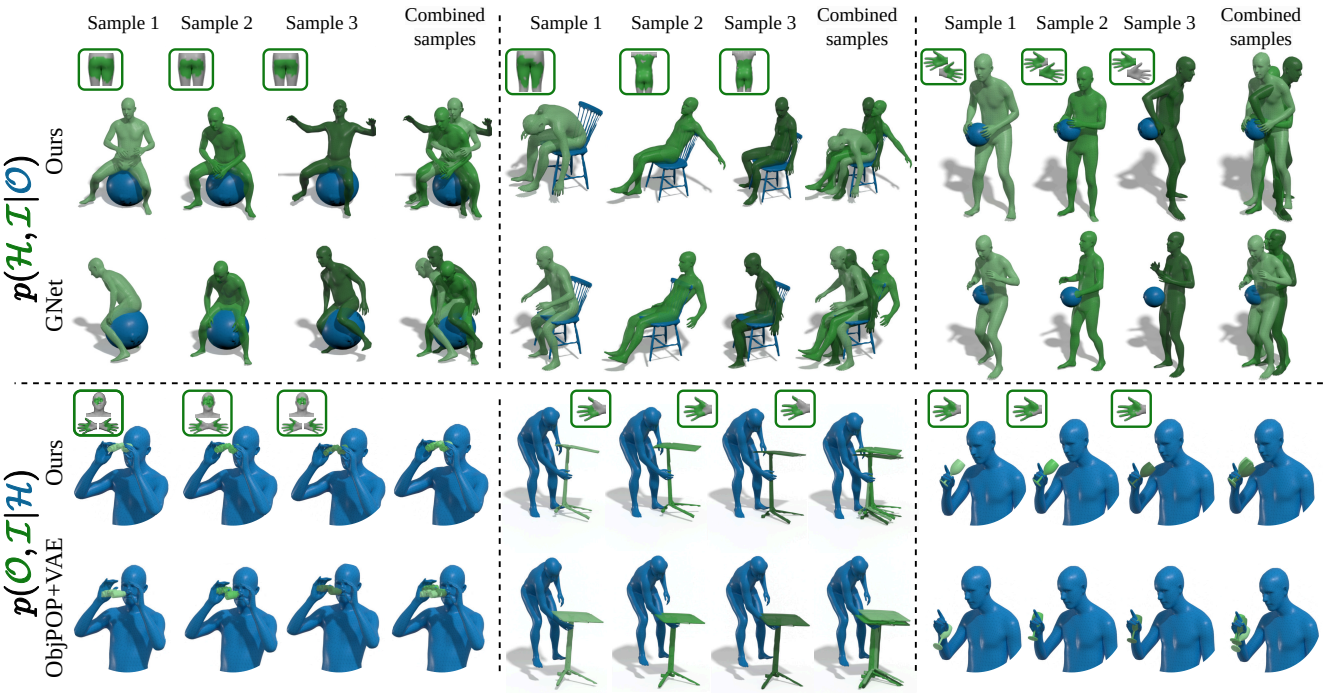

Figure 18. **Comparison with baselines**. In each group we show three samples (colored in different shades of green) for the same input, as well as one image with the same samples combined. The conditioning is taken from BEHAVE and GRAB test sets.