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. 1. We report further implementation details of TriDi, description of text labels annotation, insights on symmetry augmentation, and training losses in Sec. 2. In Sec. 3, 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 an quantitative comparison with the baselines. In Sec. 4, we include a discussion on the broader impacts of our work. Details on all four datasets used in the experiments are summarized in Sec. 5. Sec. 6 introduces an optional post-processing refinement procedure that increases the realism of the generated interactions. Finally, in Sec. 7, 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.

1. Background and Notation

Background. We follow the formulation of Denoising Diffusion Probabilistic Model (DDPM) [7] 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 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$:

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

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

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathcal{D}_\psi(\mathbf{z}_t; c, t) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon, \tag{2}$$

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

Notation. Tab. S1 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 S1. **Notation Table**. The main notation used in our paper.

2. 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 [10]. 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 [11] framework. Following [20] 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 [6] for visualization.

We implement diffusion reconstruction guidance within the 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*. We study the performance of the contact encoding model in relation to a set of text templates used for train-

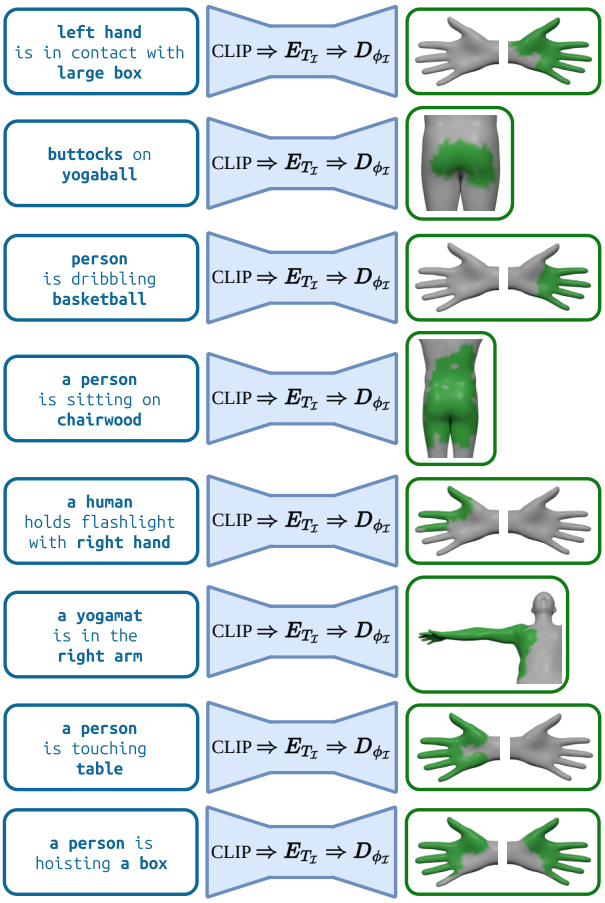

Figure S1. **Contact maps**. Examples of contact maps decoded from text queries.

ing. The model trained using only one generic template (i.e. "<body parts> <is / are? in contact with <object class>") has a significantly lower recall 63.5 compared to 75.0 of a model trained with the full set of templates. Recall is important because the GT contact maps contain mostly zeros with only a few body points in contact with the object. Moreover, the model trained with the full set of templates exhibits generalization to unseen text inputs (e.g. last row in Fig. S1).

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. S2 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.

```
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.

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

$$
\begin{aligned}
L_n^{\mathcal{H}} &= \|\theta_\mathcal{H} - \hat{\theta}_\mathcal{H}\|_1 + \|\beta_\mathcal{H} - \hat{\beta}_\mathcal{H}\|_1 + \|\mathbf{g}_\mathcal{H} - \hat{\mathbf{g}}_\mathcal{H}\|_1 \\
L_n^{\mathcal{O}} &= \|\mathbf{g}_\mathcal{O} - \hat{\mathbf{g}}_\mathcal{H}\|_1 \\
L_n^{\mathcal{I}} &= \|\mathbf{z}_\mathcal{I} - \hat{\mathbf{z}}_\mathcal{I}\|_2 \\
L_v^{\mathcal{H}} &= \|\mathbf{V}_\mathcal{H} - \hat{\mathbf{V}}_\mathcal{H}\|_2 \\
L_v^{\mathcal{O}} &= \|\mathbf{V}_\mathcal{O} - \hat{\mathbf{V}}_\mathcal{O}\|_2 \\
L_v^{\mathcal{I}} &= \|\mathbf{d} - \hat{\mathbf{d}}\|_2
\end{aligned}
\tag{3}
$$

The resulting loss function is:

$$
\begin{aligned}
L_{TriDi} = &\lambda_n^{\mathcal{H}} L_n^{\mathcal{H}} + \lambda_n^{\mathcal{O}} L_n^{\mathcal{O}} + \lambda_n^{\mathcal{I}} L_n^{\mathcal{I}} + \\
&\lambda_v^{\mathcal{H}} L_v^{\mathcal{H}} + \lambda_v^{\mathcal{O}} L_v^{\mathcal{O}} + \lambda_v^{\mathcal{I}} L_v^{\mathcal{I}}
\end{aligned}
\tag{4}
$$

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

3. Additional Evaluation

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

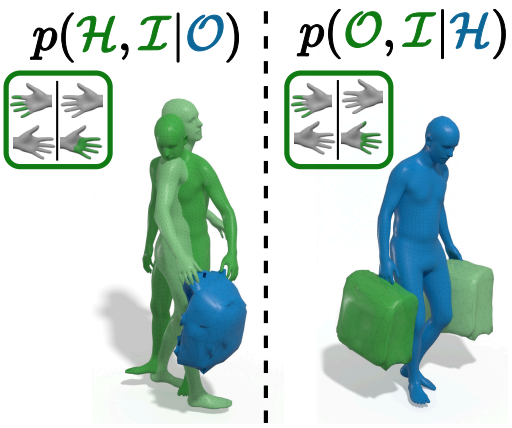

Figure S2. **Qualitative examples**. Results demonstrating the effectiveness of the symmetry augmentation. TriDi generates left- and right-handed interactions given the same condition.

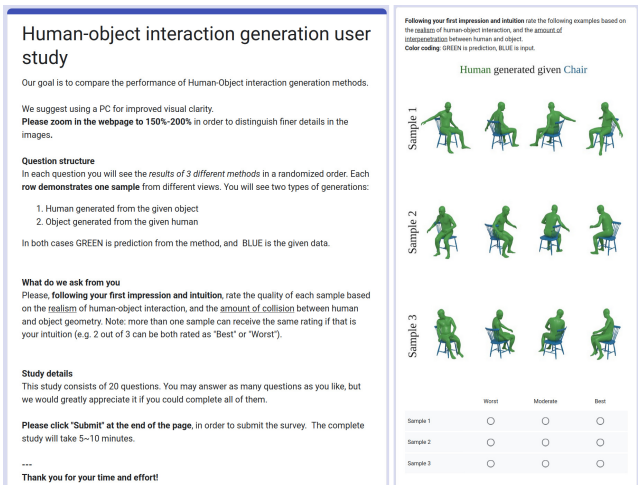

Figure S3. **User study**. The interface of the user study.

to rate the quality of each sample based on the realism of human-object interaction, and the amount of interpenetration between human and object. Each sample is rendered from the same 4 orthogonal views to allow comprehensive assessment . 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. S3. As a result, we have collected 40 responses. We summarize the results in the Tab. S2, comparing the ratings assigned to the samples by users. On average, results of TriDi were preferred to the baselines in 89.0% of the cases and preferred to the ground-truth examples in 52.0% of the cases. This suggests that the results of TriDi are more appreciable than the baselines and produce a realism comparable to captured data.

Mode	**Rating comparison**	**Result in %**
$p(\mathcal{H},\mathcal{I}\|\mathcal{O})$	TriDi $>$ GNet	87.75%
	TriDi $>$ GT data	47.75%
$p(\mathcal{O},\mathcal{I}\|\mathcal{H})$	TriDi $>$ ObjPOP+cVAE	90.25%
	TriDi $>$ GT data	56.25%

Table S2. **User study**. Summary of the user study results.

Evaluation of $\mathcal{H}|\mathcal{O},\mathcal{I}$. We compare TriDi with COINS [19] on the task of human generation given object and text in Table S3. We observe that while COINS is able to generate sitting poses it struggles to generate realistic and diverse interactions with other objects.

Method	**BEHAVE,** $\mathcal{H}\|\mathcal{O},\mathcal{I}$					
	1-NNA ($\rightarrow 50$)	COV↑	MMD↓	MPJPE↓	MPJPE-PA↓	Acc_{cont} ↑
COINS	$96.4^{\pm 0.1}$	$19.6^{\pm 0.1}$	$3.02^{\pm 0.008}$	43.4	15.9	93.9/NA
TriDi	$\mathbf{66.1}^{\pm \mathbf{0.4}}$	$\mathbf{50.8}^{\pm \mathbf{0.1}}$	$\mathbf{1.30}^{\pm \mathbf{0.010}}$	**16.9**	**10.6**	**96.7/99.5**

Table S3. **Comparison with COINS**.

Diversity and multimodality. We follow Action2Motion [5] and compute diversity (Div) and multimodality (MMod) for GT data and TriDi to demonstrate that the generated distributions in all *seven* cases are non-trivial. Additionally, we evaluate the quality of the generated contacts to prove that the generated HOI is plausible. We compute contact accuracy (Acc_c) for cases where GT contacts are available and contact presence ($Presence_c$) that reflects the percentage of generated samples with at least one vertex in contact for other cases. The contact metrics are averaged across three generated samples. The results are presented in Table S4. The variance of the distribution generated by TriDi is on par with the variance of the GT data, which means that the generated samples are non-trivial. At the same time, high contact accuracy (96.3 on average) and contact presence (98.4 on average) hint that generated interactions are plausible. The formulas for Div and MMod are provided in Section 7.

Evaluation of $\mathcal{H},\mathcal{O}|\mathcal{I}$. We compare the performance of TriDi with a model s-TriDi-HO that has the same architecture but is trained specifically on $\mathcal{H},\mathcal{O}|\mathcal{I}$ task (similar to s-TriDi-OI and s-TriDi-HI in Tables 1 and 2 of the main paper). We evaluate the methods in two modes: sampling conditioned on contact maps (CM) and conditioned on text query (Text). The results are summarized in Table S5. TriDi benefits from joint training on all the tasks together, generating a more diverse and higher quality distribution compared to the model trained specifically on one task. Results also demonstrate that text provides weaker conditioning, the resulting distribution exhibits slightly less diversity compared to the distribution of generations from contact maps.

We choose s-TriDi-HO as a baseline because, to the best of our knowledge, there are no existing methods that

Method	BEHAVE $\mathcal{H}\|\mathcal{O},\mathcal{I}$			$\mathcal{O}\|\mathcal{H},\mathcal{I}$			$\mathcal{I}\|\mathcal{H},\mathcal{O}$			$\mathcal{H},\mathcal{O}\|\mathcal{I}$		
	DIV →	MMod→	Acc_c ↑	DIV →	MMod→	Acc_c ↑	DIV →	MMod→	Acc_c ↑	DIV →	MMod→	Acc_c ↑
GT	4.32	4.15	-	2.32	2.20	-	6.68	6.16	-	4.99	4.75	-
TriDi	4.43	4.23	94.5 ± 1.2	2.34	2.21	94.8 ± 1.2	5.29	4.75	96.2 ± 0.1	5.25	4.98	94.9 ± 1.5

Method	BEHAVE $\mathcal{H},\mathcal{I}\|\mathcal{O}$			$\mathcal{O},\mathcal{I}\|\mathcal{H}$			$\mathcal{H},\mathcal{O},\mathcal{I}$		
	DIV →	MMod→	$Presence_c$ ↑	DIV →	MMod→	$Presence_c$ ↑	DIV →	MMod→	$Presence_c$ ↑
GT	8.09	7.55	-	7.15	6.62	-	8.47	7.92	-
TriDi	8.86	8.16	98.8 ± 0.1	7.89	7.15	99.3 ± 0.1	9.28	8.73	96.1 ± 2.2

Method	GRAB $\mathcal{H}\|\mathcal{O},\mathcal{I}$			$\mathcal{O}\|\mathcal{H},\mathcal{I}$			$\mathcal{I}\|\mathcal{H},\mathcal{O}$			$\mathcal{H},\mathcal{O}\|\mathcal{I}$		
	DIV →	MMod→	Acc_c ↑	DIV →	MMod→	Acc_c ↑	DIV →	MMod→	Acc_c ↑	DIV →	MMod→	Acc_c ↑
GT	5.95	5.18	-	2.33	1.53	-	4.33	3.77	-	6.45	5.46	-
TriDi	6.79	5.90	96.7 ± 0.8	2.26	1.53	97.5 ± 0.7	3.52	3.07	98.2 ± 0.1	7.39	6.88	97.7 ± 0.8

Method	GRAB $\mathcal{H},\mathcal{I}\|\mathcal{O}$			$\mathcal{O},\mathcal{I}\|\mathcal{H}$			$\mathcal{H},\mathcal{O},\mathcal{I}$		
	DIV →	MMod→	$Presence_c$ ↑	DIV →	MMod→	$Presence_c$ ↑	DIV →	MMod→	$Presence_c$ ↑
GT	7.43	6.60	-	5.04	4.17	-	7.85	6.81	-
TriDi	8.32	7.68	99.7 ± 0.1	5.08	4.39	99.3 ± 0.1	9.08	8.61	97.3 ± 1.9

Table S4. **Evaluation of diversity and multi-modality for all sampling modes.** The variance of the distribution generated by TriDi is on par with the variance of the GT data, which means that the generated samples are non-trivial. At the same time high contact accuracy (96.3 on average) and contact presence (98.4 on average) hint that generated interactions are plausible.

Method	BEHAVE $\mathcal{H},\mathcal{O}\|\mathcal{I}$		
	1-NNA (→ 50)	COV↑	MMD↓
s-TriDi-HO **(Ours)** (CM)	$71.75^{\pm 0.3}$	$47.81^{\pm 0.5}$	$3.15^{\pm 0.01}$
s-TriDi-HO **(Ours)** (Text)	$74.18^{\pm 0.2}$	46.33 ± 0.2	$3.21^{\pm 0.02}$
TriDi **(Ours)** (CM)	$\mathbf{70.03^{\pm 0.1}}$	$\mathbf{48.47^{\pm 0.1}}$	$\mathbf{3.07^{\pm 0.02}}$
TriDi **(Ours)** (Text)	$\underline{70.14^{\pm 0.3}}$	$\underline{48.10^{\pm 0.4}}$	$\underline{3.10^{\pm 0.02}}$

Method	GRAB $\mathcal{H},\mathcal{O}\|\mathcal{I}$		
	1-NNA (→ 50)	COV↑	MMD↓
s-TriDi-HO **(Ours)** (CM)	$88.81^{\pm 0.6}$	$36.29^{\pm 0.6}$	$3.10^{\pm 0.08}$
s-TriDi-HO **(Ours)** (Text)	$89.61^{\pm 0.3}$	34.81 ± 0.2	$3.28^{\pm 0.03}$
TriDi **(Ours)** (CM)	$\mathbf{87.53^{\pm 0.4}}$	$\underline{37.71^{\pm 0.1}}$	$\mathbf{3.05^{\pm 0.01}}$
TriDi **(Ours)** (Text)	$\underline{88.56^{\pm 0.3}}$	$\mathbf{37.42^{\pm 0.1}}$	$\underline{3.19^{\pm 0.02}}$

Table S5. **Quality of Generated Distribution for $\mathcal{H},\mathcal{O}|\mathcal{I}$.** TriDi outperforms s-TriDi-HO in both sampling from contact maps and text queries. Text provides weaker conditioning than contact maps, thus the resulting distribution exhibits slightly less diversity.

are able to generate static human-object interaction from text. We attempted to adapt CG-HOI [4] to consider static samples instead of motion. However, we observed that the model failed to converge after being adapted to our setting (training on static examples from GRAB and BEHAVE). Our hypothesis is that CG-HOI is designed to work with motion and is initially trained on a relatively small scale dataset (e.g., 500 short motion sequences for BEHAVE), thus generalization to significantly larger data (e.g., 130k static samples for GRAB and BEHAVE) might be too challenging for this model.

Generalization to unseen data. We provide qualitative examples of TriDi on eight unseen objects in two sampling modes in Fig. S4. 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 S5.

Ablations Here, we report the quantitative evaluations of our ablations described in the main paper. Table S6 covers the quality of the generated distributions, while Table S7 covers geometrical consistency of the generation.

Evaluation of penetration. We compute SDF-based penetration metrics in Tab.S8: average min. dist. between H and O (Min. D. $[cm.]$), contact percentage and penetration score ($C_{\%}, P_{sc}[cm.]$), CHOIS [19]), penetration depth ($PD[cm.]$, DiffH$_2$O [3]). TriDi's results are close to values computed for ground-truth data, outperforming the baselines that generate floaters and more penetrations.

Qualitative results This section includes additional qualitative results on BEHAVE (Figure S9) and GRAB (Figure

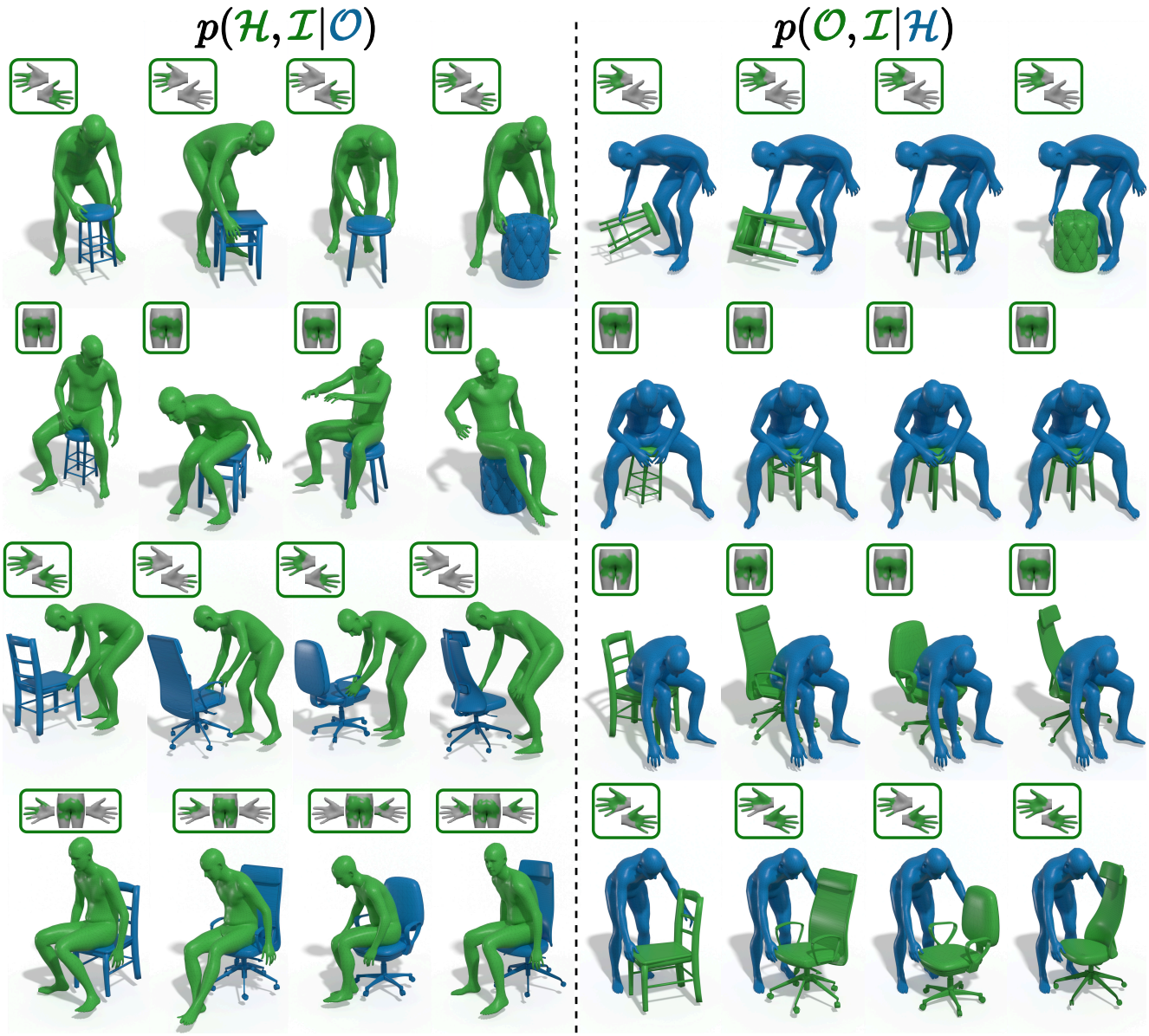

Figure S4. **Generalization to unseen geometry**. TriDi samples from $p(\mathcal{H}, \mathcal{I}|\mathcal{O})$ and $p(\mathcal{O}, \mathcal{I}|\mathcal{H})$ with unseen objects.

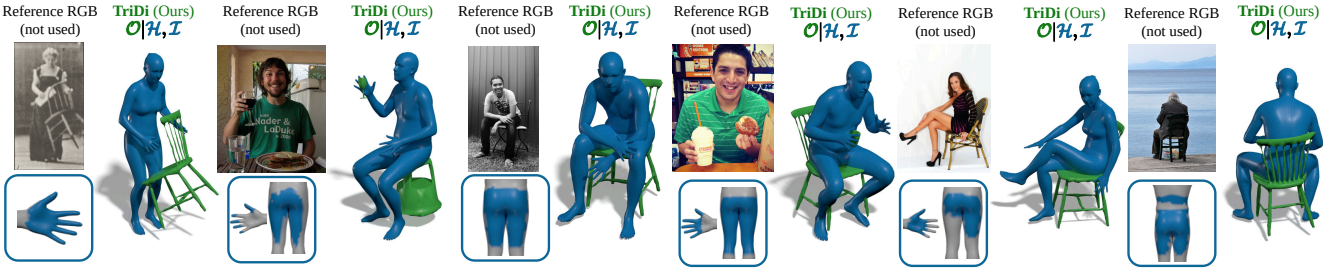

Figure S5. **Interaction reconstruction**. DECO [14] annotates human $\mathcal{H}$ and contact $\mathcal{I}$ for the RGB image, while our TriDi recovers the object $\mathcal{O}$, showing generalization on unseen data distributions.

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 S6. **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 S7. **Ablation - Geometrical Consistency of Generation**. Impact of augmentation, $\mathcal{I}$ diffusion, and guidance.

Method	BEHAVE $\mathcal{H},\mathcal{I}\|\mathcal{O}$				BEHAVE $\mathcal{O},\mathcal{I}\|\mathcal{H}$			
	Min. D.↓	$C_{\%}$ ↑	PD ↓	P_{sc} ↓	Min. D.↓	$C_{\%}$ ↑	PD ↓	P_{sc} ↓
Data	**0.65**	**99.6**	**1.73**	**0.03**	0.65	**99.6**	**1.73**	**0.03**
ObjPOP + cVAE	-	-	-	-	1.25	92.8	2.8	0.07
GNet	1.50	91.9	3.6	0.08	-	-	-	
TriDi **(Ours)**	0.74	96.5	2.6	0.07	**0.56**	97.7	2.38	0.05

Method	GRAB $\mathcal{H},\mathcal{I}\|\mathcal{O}$				GRAB $\mathcal{O},\mathcal{I}\|\mathcal{H}$			
	Min. D.↓	$C_{\%}$ ↑	PD ↓	P_{sc} ↓	Min. D.↓	$C_{\%}$ ↑	PD ↓	P_{sc} ↓
Data	**0.08**	**100.0**	0.46	0.0012	**0.08**	**100.0**	0.46	0.0012
ObjPOP + cVAE	-	-	-	-	4.98	60.1	**0.17**	**0.0005**
GNet	5.99	58.9	**0.18**	**0.0002**	-	-	-	
TriDi **(Ours)**	0.26	99.3	0.79	0.0052	0.29	98.8	0.81	0.0048

Table S8. **Penetration analysis**.

S10), and introduces examples from InterCap (Figure S7) and OMOMO (Figure S8).

Comparison with baselines In Fig. S11 we provide an extended comparison with baselines, showing 3 generated samples per same input.

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

5. Datasets

BEHAVE. BEHAVE [2] 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 [13] introduced in [12]. 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 [17]: 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 $10fps$.

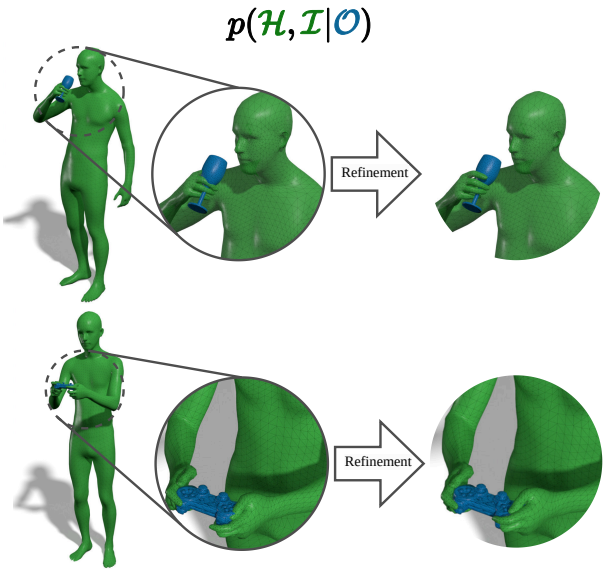

Figure S6. **Post-processing refinement result**. Example results demonstrating the effectiveness of the post-processing refinement. Optionally, TriDi results can be refined using an optimization procedure that improves fine hand details.

6. 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 examples of post-processing refinement in Fig. S6 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 [16] to design an optimization procedure refining the generated hands. The original refinement minimizes the error term:

$$E_{fc} + w_{dis}E_{dis} + w_{pen}E_{pen} + w_{spen}E_{spen} + w_{prior}E_{prior} \tag{5}$$

where E_{fc} is a force closure term proposed in [9] 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 [16] 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:

$$E_{reg} = \|\hat{\theta_{\mathcal{H}}} - \tilde{\theta_{\mathcal{H}}}\|_2 \tag{6}$$

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 [8, 15] that detects the collision between hand and object meshes, penalizing the quantity:

$$E_{isect} = \sum_{(\mathbf{f}_{\mathcal{H}}, \mathbf{f}_{\mathcal{O}}) \in C} \left[\sum_{\mathbf{v}_{\mathcal{H}} \in \mathbf{f}_{\mathcal{H}}} \| - \Psi_{\mathbf{f}_{\mathcal{O}}}(\mathbf{v}_{\mathcal{H}})\|^2 + \sum_{\mathbf{v}_{\mathcal{O}} \in \mathbf{f}_{\mathcal{O}}} \| - \Psi_{\mathbf{f}_{\mathcal{H}}}(\mathbf{v}_{\mathcal{O}})\|^2 \right] \tag{7}$$

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, $\Psi_{\mathbf{f}} : \mathbb{R}^3 \to \mathbb{R}_+$ is a cone distance field from the face $\mho$ (full definition can be found in [15]).

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:

$$\begin{aligned} E_{stage\_1} =& w_{dis}E_{dis} + w_{pen}E_{pen} + \\ & w_{reg}E_{reg} + w_{isect}E_{isect} \\ E_{stage\_2} =& E_{fc} + w_{dis}E_{dis} + w_{pen}E_{pen} + \\ & w_{spen}E_{spen} + w_{prior}E_{prior} + \\ & w_{reg}E_{reg} + w_{isect}E_{isect} \end{aligned} \tag{8}$$

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 [16] with 1000 iterations for the first stage and 2000 iterations for the second stage.

7. Error Metrics

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

$$COV(S_g, S_r) = \frac{|\{\arg\min_{r \in S_r} D(g, r) | g \in S_g\}|}{|S_r|}, \tag{9}$$

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]:

$$MMD(S_g, S_r) = \frac{1}{|S_r|} \sum_{r \in S_r} \min_{g \in S_g} D(g, r) \tag{10}$$

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

1-Nearest Neighbor Accuracy (1-NNA) [18]. 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:

$$\text{1-}NNA(S_g, S_r) = \frac{\sum_{X \in S_g} \mathbb{1}[N_X \in S_g] + \sum_{Y \in S_r} \mathbb{1}[N_Y \in S_r]}{|S_g| + |S_r|}, \tag{11}$$

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 $D(\cdot, \cdot)$.

Diversity (Div) [5]. Diversity measures the variance of the generated samples. Two subsets $S_1 = \{v_1, ..., v_{|S|}\}$ and $S_2 = \{v'_1, ..., v'_{|S|}\}$ of the same size $|S| = 200$ are drawn from either S_g or S_r (depending on whether we want to evaluate the metric for the method or the GT data). The diversity then is computed as follows:

$$Div(S_1, S_2) = \frac{1}{|S|} \sum_{i=1}^{|S|} \|v_i - v'_i\|_2, \tag{12}$$

Multimodality (MMod) [5]. Multimodality measures the variance of the generated samples within the same object category. For every object class $c \in 1, ...C$ two subsets $S_1^c = \{v_{c,1}, ..., v_{c,|S|}\}$ and $S_2^c = \{v'_{c,1}, ..., v'_{c,|S|}\}$ of the same size $|S| = 200$ are drawn from either S_g or S_r. The multimodality is then computed as follows ($S_1 = \{S_1^1, ..., S_1^C\}, S_2 = \{S_2^1, ..., S_2^C\}$):

$$MMod(S_1, S_2) = \frac{1}{C * |S|} \sum_{c=1}^{C} \sum_{i=1}^{|S|} \|v_{c,i} - v'_{c,i}\|_2, \tag{13}$$

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:

$$E_{v2v} = \frac{1}{|\mathbf{V}_\mathcal{O}|} \sum_{i \in |\mathbf{V}_\mathcal{O}|} \|\mathbf{V}^i_\mathcal{O} - \hat{\mathbf{V}}^i_\mathcal{O}\|_2 \tag{14}$$

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

$$E_c = \left\| \frac{1}{|\mathbf{V}_\mathcal{O}|} \sum_{i \in |\mathbf{V}_\mathcal{O}|} \mathbf{V}^i_\mathcal{O} - \frac{1}{|\hat{\mathbf{V}}_\mathcal{O}|} \sum_{i \in |\hat{\mathbf{V}}_\mathcal{O}|} \hat{\mathbf{V}}^i_\mathcal{O} \right\|_2. \tag{15}$$

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

$$Acc_{cont} = \frac{1}{|\mathbf{V}_\mathcal{H}|} \sum_{i \in |\mathbf{V}_\mathcal{H}|} \mathbb{1}[\hat{\phi}^i_\mathcal{I} = \phi^i_\mathcal{I}], \tag{16}$$

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

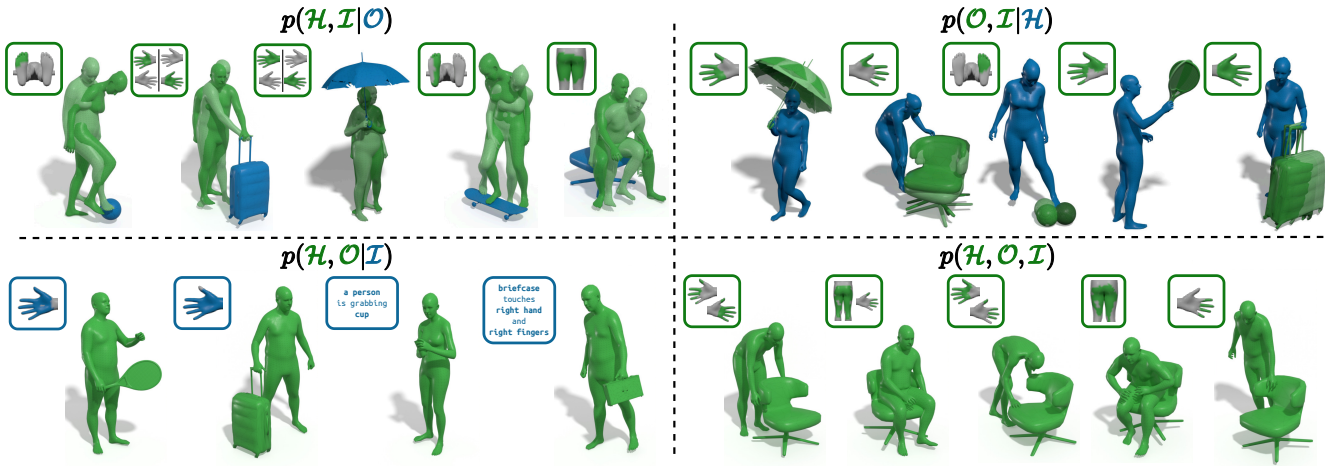

Figure S7. **Qualitative results of TriDi on InterCap**.

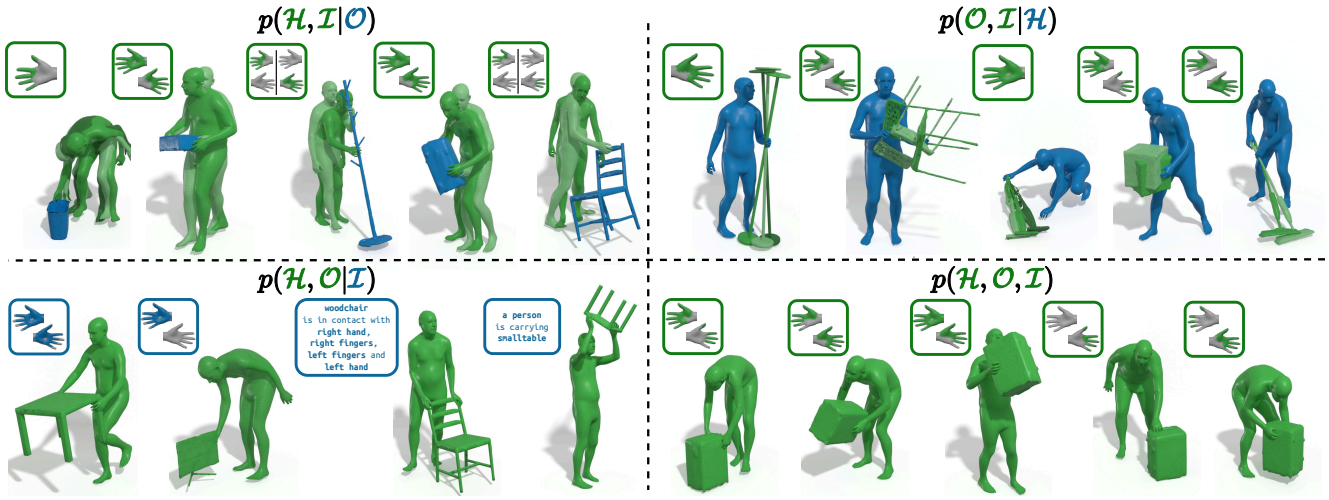

Figure S8. **Qualitative results of TriDi on OMOMO**.

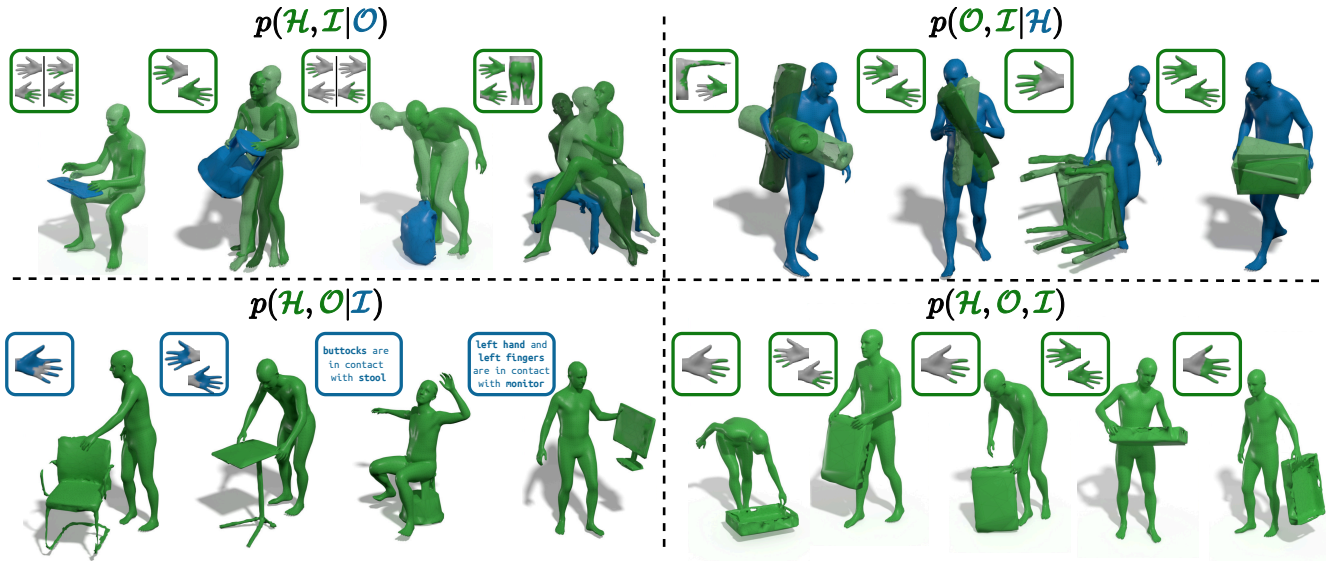

Figure S9. **Qualitative results of TriDi on BEHAVE**.

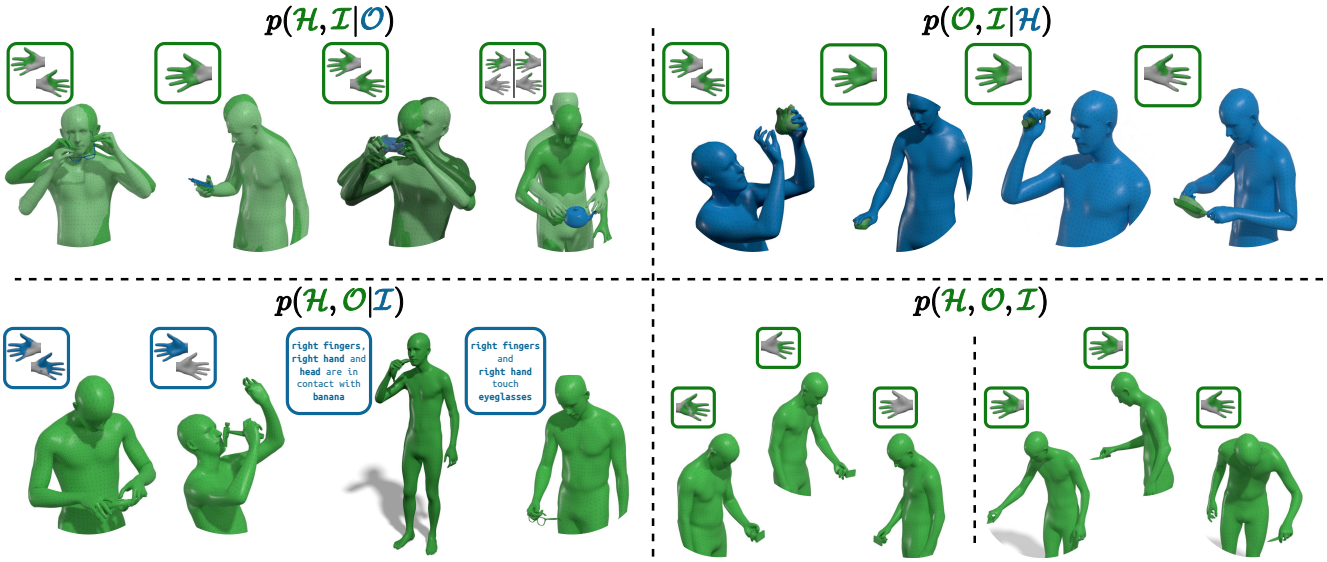

Figure S10. **Qualitative results of TriDi on GRAB**.

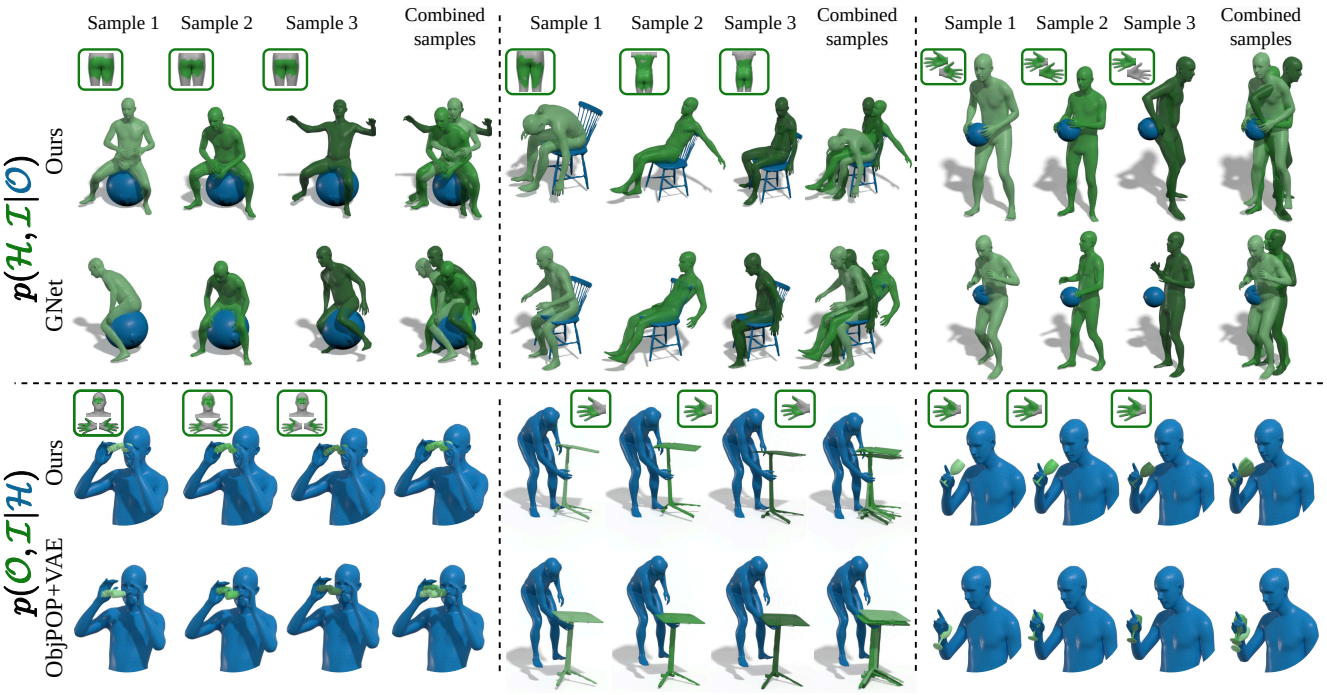

Figure S11. **Comparison with baselines**. In each group we show three samples (colored in different shades of green) for the same input, as well as one image with the same samples combined. The conditioning is taken from BEHAVE and GRAB test sets.

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. 7, 8

[2] 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. 6

[3] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 4

[4] 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. 4

[5] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 3, 8

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

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

[8] 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. 7

[9] 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. 7

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

[11] 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

[12] 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. 6

[13] 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

[14] 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. 5

[15] 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. 7

[16] 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. 7

[17] 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. 6

[18] 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. 8

[19] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, , and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European conference on computer vision (ECCV)*, 2022. 3, 4

[20] 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