LEXIS: LatEnt proXimal Interaction Signatures for 3D HOI from an Image

Dimitrije Antić[1] Alvaro Budria[*1] George Paschalidis[*1]
Sai Kumar Dwivedi[2] Dimitrios Tzionas[1,3]

[*] Equal contribution
[1]University of Amsterdam, The Netherlands
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]Aristotle University of Thessaloniki, Greece

Fig. 1: We present LEXIS-Flow, a framework for 3D Human-Object Interaction (HOI) reconstruction from single images. We go beyond sparse, binary contact by learning LEXIS, a latent manifold of dense, continuous interaction signatures. LEXIS-guided sampling enables recovering physically-plausible HOI without post-hoc optimization.

Abstract. Reconstructing 3D Human–Object Interaction from an RGB image is essential for perceptive systems. Yet, this remains challenging as it requires capturing the subtle physical coupling between the body and objects. While current methods rely on *sparse, binary contact* cues, these fail to model the continuous proximity and dense spatial relationships that characterize natural interactions. We address this limitation via InterFields, a representation that encodes *dense, continuous proximity* across the entire body and object surfaces. However, inferring these fields from single images is inherently ill-posed. To tackle this, our intuition is that interaction patterns are characteristically structured by the action and object geometry. We capture this structure in LEXIS, a novel discrete manifold of interaction signatures learned via a VQ-VAE. We then develop LEXIS-Flow, a diffusion framework that leverages LEXIS signatures to estimate human and object meshes alongside their InterFields. Notably, these InterFields help in a guided refinement that ensures physically-plausible, proximity-aware reconstructions without requiring post-hoc optimization. Evaluation on Open3DHOI and BEHAVE shows that LEXIS-Flow significantly outperforms existing SotA baselines in reconstruction, contact, and proximity quality. Our approach not only improves generalization but also yields reconstructions perceived as more realistic, moving us closer to holistic 3D scene understanding. Code & models will be public at `https://anticdimi.github.io/lexis`.

1 Introduction

Our actions are inherently defined and physically constrained bythe objects around us. Accurately recovering 3D human–object interaction (HOI) from a single RGB image is essential for virtual and robotic assistants, mixed reality, animation, and analyzing interactions from internet-scale images. Technically, this task involves estimating not only the 3D pose and shape of the human and the object, but also their relative 3D *spatial configuration* and subtle *physical coupling*.

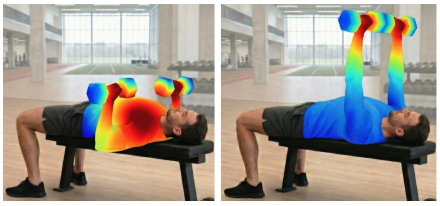

Fig. 2: Toy example. Contact between the body and bench, and between hands and dumbbells, remains fixed. Instead, *distances* from each body point to the objects, v.v., form a rich, geometry- and proximity-aware *interaction signature*.

This is challenging due to depth/scale ambiguities, self-occlusions, and the mutual occlusions between bodies and objects in RGB images. These ambiguities pose significant hurdles for computational methods. Yet, humans navigate these challenges effortlessly via prior experience. Capturing this human-like capability requires two components: a *prior model* that captures the interaction "geometry," and a mechanism to *integrate* this prior into the *reconstruction process*.

Despite their importance, both components involve open problems because we currently lack a representation that can bridge the gap between sparse 2D image cues and dense 3D physical interaction. To tackle this, past work exploits *contact* [6, 16, 47, 67, 69, 88]; body and object surface points are classified as in contact or not. Contact points act as anchors that align body–object geometry, accounting for the occlusions and depth ambiguities of images.

However, contact has fundamental limitations. First, it is only a *sparse* signal; non-contacting points are ignored. Second, it is only *binary*; it encodes areas of zero mesh distance, ignoring non-zero ones. Think of a person working out (see Fig. 2); contact between the body and objects stays fixed across time, failing to encode the interaction's progression.

To tackle the above limitations, we need a richer representation. Our key observation is that *distances* between body points and object points change as the interaction progresses. This yields a *dense*, *continuous* signal of surface-to-surface proximity, namely a 3D *Interaction Field*, or InterField for short (see Fig. 1). Note that contact is only a sparse subset (zero level set) of the broader interaction field; that is, the InterField does not "replace" contact but instead "*extends*" it. Intuitively, the InterField representation captures geometry- and proximity-aware *interaction signatures* that help 3D reconstruction.

Yet, this potential remains underexplored. Existing work estimates hand-only InterFields [20] from an image, without using them in any task. Other work exploits hand-only InterFields [68, 89] to synthesize grasps, but only "implicitly" for training losses. No existing method uses InterFields to guide *neural in-model refinement* at *inference* time. We fill this gap by inferring both body and object InterFields from an image, and exploiting these for 3D reconstruction.

However, inferring 3D fields, i.e., distances for all points across two 3D surfaces from just a 2D image is inherently ill-posed. Moreover, the space of possible interactions is vast and highly non-convex. To tackle these challenges, we empirically observe that InterFields contain interaction- and object-specific patterns, or "*interaction signatures*." Thus, we learn a novel *lexicon* of these patterns called LEXIS (LatEnt proXimal Interaction Signatures), namely a discrete *manifold* of interaction signatures, encoding *prior knowledge* about these. Specifically, to learn LEXIS we train a VQ-VAE [18, 22] that takes as input 3D human and object geometry in interaction, encodes this into a latent code, and decodes this into body and object InterFields. As training data we use MoCap-based, synthetically-augmented 3D geometry of interactions.

We exploit LEXIS to estimate 3D interactions from a single RGB image. Existing methods mostly follow a two-stage pipeline: they first estimate coarse human and object shape/pose (often missing contacts), and then refine them in post-hoc optimization with contact-based losses to improve interaction plausibility. This is suboptimal for two reasons; first, contact provides only *sparse, binary* cues; second, these are considered only "*too late*," i.e., post-hoc. Instead, we leverage *dense, continuous* InterFields and consider them "*early on*" in our model. More specifically, we develop LEXIS-Flow, a Flow-Matching model that jointly estimates posed 3D human and object meshes, along with their InterFields. Crucially, the predicted InterFields help in a guided refinement of meshes, eliminating the need for post-hoc optimization in a separate stage.

We evaluate primarily on the Open3DHOI [78] dataset (unless otherwise specified) due to its focus on *in-the-wild* images and *generalizability*. Evaluation shows that dense, continuous InterFields are more helpful than sparse, binary contacts for formulating 3D reconstruction constraints; we test this both in LEXIS-Flow's guided refinement and in a fitting framework [16]. Benchmarking shows that LEXIS-Flow clearly outperforms SotA methods on 3D reconstruction in the wild and its estimates are perceived as significantly more realistic. In-distribution evaluation on the BEHAVE [4] dataset echoes the above. Ablation studies support our key design choices. Last, evaluation shows that LEXIS-Flow can be initialized with 3D estimates provided by SotA tools, and LEXIS-Flow's guided refinement clearly improves these.

In summary, here we make the following contributions:

1. We go beyond sparse, binary contact by inferring *dense, continuous* InterFields that encode *proximity* cues across entire body and object surfaces.
2. We learn the novel LEXIS *dictionary* that encodes action- and object-aware *latent interaction signatures*, which are learned from 3D interaction datasets.
3. We develop LEXIS-Flow, a model that exploits LEXIS to estimate a 3D human, object, and their InterFields, from an image. Notably, LEXIS-based InterFields guide the Flow-Matching generative sampling to refine estimates and improve the 3D HOI plausibility, without post-hoc optimization.

Code and models will be available at `https://anticdimi.github.io/lexis`.

2 Related Work

2.1 HOI Representations

Binary Contact: Contact is a binary cue (points are in contact or not) that is used to taxonomize grasps [3,21,36], to synthesize hand [34,47,69,92] or body [15, 30, 38, 59] interaction, and reconstruct hand [26, 27, 86] or body [52, 82, 90] interaction from pixels. For the latter, existing work: (i) detects contact pixels [8], (ii) thresholds distances to compute 3D contact [31,69], (iii) exploits manual 3D contact labels [13,90], (iv) infers 3D contact [16,52,72] or (v) pressure [25,73].

Spatial Probability: Recent work uses text-to-image diffusion to generate 3D training data [28,38]. CHORUS [28] learns a 3D object occupancy field w.r.t. the body. ComA [38] extends this by learning an object-to-human 3D distance field with orientation cues. These works learn separate models per object/action class, and need heavy sampling/filtering, but their zero-shot nature is promising.

Interaction Fields (InterFields): Going beyond sparse, binary contact, InterFields encode dense, continuous surface-to-surface distances. CHORE [82] thresholds learned Distance Fields (DF) to extract binary contacts. Other work uses DFs to synthesize body [15] or hand interaction [68,70,85,89,92] or to reconstruct grasps [20,60,86], but mostly implicitly in features or losses. LEXIS-Flow differs from these in two ways: **(1) Usage:** LEXIS-Flow is the first method to use full-body InterFields explicitly in *inference*. **(2) Task:** LEXIS-Flow does the above for *reconstruction* (with InterField-guided refinement), while CG-HOI [15] for synthesis. Uniquely, LEXIS-Flow exploits a novel *dictionary* of InterFields.

2.2 3D HOI Reconstruction from a Single Image

Human-only Reconstruction: Optimization methods fit a parametric 3D body model [35,49,83] to image cues [5,35]. Regression-based methods estimate parameters of such models [17,37,40,56,87] or a non-parametric mesh [12,42,65].

Object-only Reconstruction: Pose is estimated via regression [23, 80] or by fitting a template [24, 48, 74] or implicit SDF [1] with optimization. Shape is encoded as voxels [9,11], point clouds [19,19], superquadrics [55], meshes [23,75], or neural fields [10,54,94]. Recent work uses diffusion [45,46,62] or database retrieval [33,44]. Recently SAM3D [71] estimates robust object shape, but its pose estimation lacks interaction awareness w.r.t. humans; we *initialize* LEXIS-Flow with SAM3D object estimates, and *refine pose* by adding interaction awareness.

HOI Datasets: Images paired with 3D GT are scarce. Most datasets are captured in constrained in-lab [20, 91], indoor [4, 29, 32, 66] or outdoor [31] settings. Sometimes these [4, 32] are synthetically augmented [81]. To recover better 3D pseudo-GT, the BEHAVE [4] and InterCap [32] datasets are built using multi-view RGB-D cameras, while other datasets are built with multi-view RGB cameras [31, 91, 93] along with IMU sensors [93]. Recently 3D contact labels are crowd-sourced [13, 72, 78, 88] to help 3D reconstruction for in-the-wild images.

Template-free HOI methods: HDM [81] infers human and object point clouds via a hierarchical diffusion framework; this first predicts rough a point cloud for each, and then refines these separately via cross-attentive diffusion.

Template-based HOI methods: Object shape is often assumed known. Optimization [78, 82, 90] fits 3D human/object poses to images while satisfying contact constraints. PHOSA [90] assumes known 3D contacts. HOI-Gaussian [78] derives contacts from Gaussian splat opacities. CHORE [82] infers distance fields around the human/object and extracts from these binary contacts. Instead of optimizing, CONTHO [52] directly infers a human mesh and object orientation.

Most methods work for in-distribution images, captured in constrained environments with pre-scanned objects, so they struggle generalizing. To tackle this, recent work annotates *in-the-wild* images with 3D contact labels (at a part- [78] or vertex-level [13, 72] granularity), used to constrain optimization [13, 78]. For unlabeled images, PICO [13] first infers (noisy) body-only 3D contact from pixels [72], and uses this as query to retrieve clean contact labels from its PICO-DB database. InteractVLM [16] learns contact from scarce 3D labels via vision-language models. PICO and InteractVLM retrieve shape from a large 3D object database [14] via a joint image-geometry latent space [44], but databases have finite sizes. Instead, we estimate shape via the SAM3D [71] foundational model. Most work [13, 16, 52, 77, 82, 90] follows a *two-stage* approach; it first estimates rough meshes, and then refines these via *sparse, binary* contact in *post-hoc* optimization [82, 90] or in a transformer [52]. Instead, we follow a *one-stage* approach; we use diffusion to *jointly* infer a 3D body, object, and respective *dense, continuous* InterFields, which guide sampling to refine estimates.

3 Methodology

Overview: Given an RGB image depicting a human–object interaction (HOI), $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$, and a 3D object shape estimation by an off-the-shelf model, the goal is to estimate the HOI in 3D, namely to estimate a 3D human mesh (SMPL+H[1] body [64]) and object mesh, shaped and posed such that they interact realistically and match image cues. Technically, this estimates two states: **(1) Object state** $\mathcal{O} = \{R^o, t^o\}$ in a **human-root-relative frame**, where $R^o \in \mathbb{R}^6$ is the global rotation (in 6D form [95]), and $t^o \in \mathbb{R}^3$ the global translation; **(2) Body state** $\mathcal{B} = \{R^b, t^b, \boldsymbol{\beta}, \mathcal{Z}\}$ in the **camera frame**, where $R^b \in \mathbb{R}^6$ is the global rotation (in 6D form [95]), $t^b \in \mathbb{R}^3$ is the global translation, $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are SMPL+H shape parameters, and $\mathcal{Z} \in \mathbb{R}^{21 \times D}$ are tokens of a *novel* lexicon, called LEXIS (see Sec. 3.1), encoding both pose and "*interaction signatures*."

Representation (InterField): To encode interaction relationships, most work uses *sparse, binary contact*; points are either in contact or not. Recent work uses InterFields, $\text{IF}(\mathbf{p}) \in \mathbb{R}_{\geq 0}$, which, for every[2] 3D body/object point, $\mathbf{p}$, encodes the distance to the nearest point on the interacting counterpart. This is a *dense*, *continuous* signal, that models both contact and *proximity*. Our approach builds on this richer representation. However, estimating 3D InterFields from a 2D image is a highly ill-posed, high-dimensional problem. We tackle this below.

[1] Our method works directly for SMPL+H [64] & SMPL [49].
[2] InterFields are **high-dimensional**; to store a distance per SMPL+H vertex: $\mathbb{R}^{6890}$.

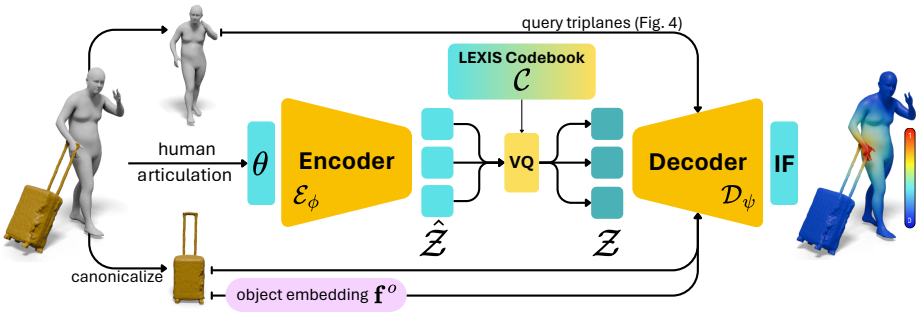

Fig. 3: LEXIS (Sec. 3.1). To estimate 3D HOI from images we need a prior model of interactions. We train LEXIS-Net, a VQ-VAE that learns a compact dictionary of proximal interaction signatures, termed LEXIS. The encoder $\mathcal{E}_\phi$ maps 3D body pose to continuous latents $\hat{\mathcal{Z}}$, quantized via a learned codebook $\mathcal{C}$ into discrete tokens $\mathcal{Z}$, and decoded via $\mathcal{D}_\psi$ into 3D body pose and body/object InterFields (shown color-coded).

Approach: We develop a framework that learns compact "*interaction signatures*" and *explicitly* exploits these for 3D HOI reconstruction from an image: **Interaction Signatures (Sec. 3.1):** We employ a VQ-VAE and learn **LEXIS**, a novel lexicon of *compact interaction signatures* that encode interaction- and object-specific patterns observed in InterFields. Note that LEXIS codes can be decoded into full 3D InterFields to form constraints for downstream applications. **Neural Reconstruction (Secs. 3.2 and 3.3):** We build LEXIS-Flow (Sec. 3.2), a dual-stream Flow-Matching transformer, that takes an image (and object shape estimation) and estimates a 3D human and object in interaction, along with a LEXIS code describing this. This LEXIS code is decoded into InterFields to guide sampling to refine (Sec. 3.3) body and object pose, along with updated InterFields, in a form of neural analysis-by-synthesis. This makes LEXIS-Flow a *one-stage end-to-end* model, which does not require post-hoc optimization.

3.1 LEXIS: Learned Interaction Signatures

InterFields store a distance value for every point across a 3D surface (to the closest point on the interacting counterpart). Thus, they are high-dimensional ($\mathbb{R}^{6890}$ for SMPL+H), so they are challenging to infer. Moreover, estimating 3D InterFields from only a single 2D image is highly ill-posed. Our key observation is that InterFields encode action- and object-specific patterns, i.e., compact "interaction signatures". Thus, these can be used to infer low-dimensional interaction signatures, which can be decoded into full InterFields to form geometric losses.

To this end, we employ a VQ-VAE, called **LEXIS-Net**, to learn a novel compact *lexicon of interaction signatures*, called **LEXIS**; see Fig. 3 for an overview. LEXIS-Net builds on the pose tokenization of TokenHMR [18] and extends it so that, conditioned on object shape in canonical pose (denoted as point cloud $\mathbf{P}^o \in \mathbb{R}^{N\times 3}$), a decoder reconstructs both 3D body pose and 3D (body and object) InterFields. Thus, the learned tokens encode not only the body configuration, but, crucially, also the proximal human-object relationships and physical coupling that describes the interaction computationally.

In detail, an encoder $\mathcal{E}_\phi$ (with network weights ϕ) maps body pose, $\boldsymbol{\theta} \in \mathbb{R}^{21\times 3}$ (axis-angle rotations for 21 joints up to the wrist), to a discrete token sequence (one token per joint), $\mathcal{Z} \in \mathbb{R}^{21\times D}$, via a learned codebook, $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K}$, $\mathbf{c}_k \in \mathbb{R}^{21\times D}$, called **LEXIS** (for LatEnt proXimal Interaction Signatures):

$$\mathcal{Z} = \arg\min_{\mathbf{c}_k \in \mathcal{C}} \|\hat{\mathcal{Z}} - \mathbf{c}_k\|_2, \qquad \hat{\mathcal{Z}} = E_\phi(\boldsymbol{\theta}), \tag{1}$$

where $\hat{\mathcal{Z}}$ is a sequence of continuous latents, and $\mathcal{Z}$ of discrete latents, the quantization operator (arg min) maps continuous latents ($\hat{\mathcal{Z}}$) to their nearest discrete codebook entry ($\mathbf{c}_k$), and D is the token length and K the codebook size; for the values of the latter two, see "Implementation Details" in Sec. 3.4.

The decoder D_ψ (with network weights ψ) maps latents $\mathcal{Z}$ to full InterFields. Naively doing so for a specific mesh is straightforward, but ties the decoder, D_ψ, to a specific topology; note that different body models have a different number of vertices, and objects vary widely. To tackle this, D_ψ projects quantized tokens $\mathcal{Z}$ into two TriPlane [7] feature volumes:

$$\mathcal{T}^b, \mathcal{T}^o = D_\psi(\mathcal{Z}, \mathbf{f}^o), \tag{2}$$

where $\mathcal{T}^b, \mathcal{T}^o \in \mathbb{R}^{3\times T_r \times T_r \times T_c}$ are TriPlane feature volumes for the body and object respectively, and $\mathbf{f}^o$ is an object shape embedding extracted by a pre-trained frozen PointNeXt [61] encoder applied on the object point cloud, $\mathbf{P}^o$. To query the decoded InterField at any 3D surface point, $\mathbf{p} \in \mathbb{R}^3$, LEXIS-Net projects $\mathbf{p}$ onto the three axis-aligned planes of the respective TriPlane (see Fig. 4), bilinearly samples and sums the features, and infers an InterField value through a lightweight MLP.

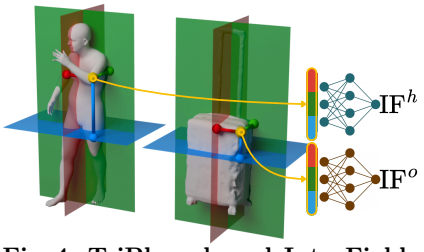

Fig. 4: TriPlane-based InterFields. A 3D surface point $\mathbf{p}$ (see yellow point on human/object surface) is orthogonally projected onto the three feature planes (see red, blue, and green points) to sample features, which are aggregated and passed to a small MLP to infer the InterField value for point $\mathbf{p}$.

The LEXIS-Net training objective is:

$$\mathcal{L}_{\text{LexisNet}} = \mathcal{L}_{\text{recon}} + \lambda^h_{\text{IF}}\,\mathcal{L}^h_{IF} + \lambda^o_{\text{IF}}\,\mathcal{L}^o_{IF} + \lambda_{\text{VQ}}\,\mathcal{L}_{\text{VQ}} + \lambda_{\text{commit}}\,\mathcal{L}_{\text{commit}}, \tag{3}$$

where $\mathcal{L}_{\text{recon}}$ is a reconstruction loss defined as $\mathcal{L}_{\text{recon}} = \left\|\mathcal{M}(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta}) - \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})\right\|_2$, $\mathcal{M}(\cdot,\cdot)$ denotes the SMPL+H [64] body mesh produced for given pose and shape parameters, $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ are the estimated and GT body poses, respectively, $\boldsymbol{\beta}$ is GT SMPL+H shape parameters, and $\mathcal{L}_{IF}$ is a loss that penalizes the discrepancy between estimated and GT InterFields, $\hat{\text{IF}}(\mathbf{p})$ and $\text{IF}(\mathbf{p})$, respectively, as $\mathcal{L}_{IF} = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{p}\in\mathcal{S}} |\hat{\text{IF}}(\mathbf{p}) - \text{IF}(\mathbf{p})|$, where $\mathbf{p}$ are 1024 points uniformly sampled on a surface $\mathcal{S}$; we compute $\mathcal{L}_{IF}$ separately for the body ($\mathcal{L}^b_{IF}$) and object ($\mathcal{L}^o_{IF}$). The codebook losses $\mathcal{L}_{\text{VQ}}$ and $\mathcal{L}_{\text{commit}}$ follow standard VQ-VAE training [18]. The steering weights λ are set empirically; see "Implementation Details" in Sec. 3.4.

In Secs. 3.2 and 3.3 we refer to the codebook $\mathcal{C}$ as "**LEXIS**", and to the *continuous* $\hat{\mathcal{Z}}$ as "**LEXIS Code**" as diffusion operates on codes $\hat{\mathcal{Z}}$. In practice, estimated $\hat{\text{IF}}$=$D_\psi(\hat{\mathcal{Z}}, \mathbf{f}^o)$ is mapped via $\exp(-\omega\cdot\hat{\text{IF}})$ to emphasize close proximity.

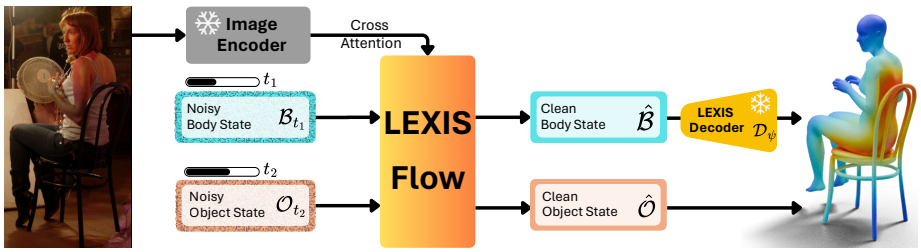

Fig. 5: LEXIS-Flow (Sec. 3.2). We develop a dual-stream Flow-Matching model that takes as input a single image, and estimates 3D body and object meshes in interaction along with LEXIS-based InterField proximal relationships (Sec. 3.1; shown with heatmap color-coding on the 3D meshes). Guiding sampling via LEXIS-based InterFields refines estimates to improves the physical plausibility of 3D interaction.

3.2 LEXIS-Flow

After pretraining a LEXIS codebook of interaction signatures (Sec. 3.1), we exploit this to reconstruct 3D HOI from an image. To this end, we build LEXIS-Flow (see Fig. 5 for an overview), a model that captures the conditional distribution $P(\mathcal{B}, \mathcal{O} \mid \mathcal{I})$ via Flow Matching [43]. The generative process is formulated as an Ordinary Differential Equation (ODE) that transports samples from a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ at t=0 to the target HOI distribution at t=1.

Let $\mathbf{x} = [\mathcal{B}; \mathcal{O}]$ denote the HOI state. Then, the flow interpolates between Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at t=0 and a ground-truth sample $\mathbf{x}_1$ at t=1 via $\mathbf{x}_t = (1-t)\,\boldsymbol{\epsilon} + t\,\mathbf{x}_1$. To approximate the vector field that generates this flow, we train a neural network, $v_\gamma(\mathbf{x}_t, t, \mathcal{I})$; then the target flow velocity is $^{d\mathbf{x}_t}/_{dt} = \mathbf{x}_1 - \boldsymbol{\epsilon}$.

The vector field, v_γ, is parameterized by a multi-stream transformer backbone inspired by DiT [58] and Mixture of Transformers [71]. The body stream (for state $\mathcal{B}$), and object stream (for state $\mathcal{O}$), operate with decoupled noise schedules, conditioned on independent timesteps t_1 and t_2, respectively, as in TriDi [59]. Cross-attention between the two streams conditions each modality on the other one, with sinusoidal timestep embeddings informing the network of each stream's noise level. During training, t_1 and t_2 are sampled independently [2]. This lets the network model both the joint distribution $P(\mathcal{B}, \mathcal{O} \mid \mathcal{I})$ and the conditional ones, $P(\mathcal{B} \mid \mathcal{O}, \mathcal{I})$ and $P(\mathcal{O} \mid \mathcal{B}, \mathcal{I})$, preventing modality collapse (where one stream might get ignored) by requiring each stream to be informative at all noise levels.

The LEXIS-Flow training objective is:

$$\mathcal{L}_{\text{LEXIS-Flow}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{IF}}\,\mathcal{L}_{IF} + \lambda_{\text{v2v}}\,\mathcal{L}_{\text{v2v}} + \lambda_{\text{2D}}\,\mathcal{L}_{\text{2D}} + \lambda_{\text{obs}}\,\mathcal{L}_{\text{obs}}, \tag{4}$$

where $\mathcal{L}_{\text{FM}}$ is the conditional flow-matching loss:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t_1,t_2\sim\mathcal{U}(0,1),\ \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left\| v_\gamma(\mathbf{x}_{(t_1,t_2)},\, t_1,\, t_2,\, \mathcal{I}) - \mathbf{u}_{\text{target}} \right\|_2^2, \tag{5}$$

where $\mathbf{x}_{t_1,t_2}$ is the noised state with body components interpolated at t_1 and object components at t_2, and $\mathbf{u}_{\text{target}}$ is the conditional vector field target; note that the expectation is approximated by mini-batch averaging. The rest of the loss

terms provide explicit geometric and visual supervision; $\mathcal{L}_{IF}$ penalizes the L_1 discrepancy between the estimated ($D_\psi(\hat{\mathcal{Z}}, \mathbf{f}^o)$) and GT InterFields, $\mathcal{L}_{\text{v2v}}$ is a L_2 vertex-to-vertex loss between estimated ($\mathcal{M}(\hat{\mathcal{B}})$) and GT ($\mathcal{M}(\mathcal{B})$) body meshes, $\mathcal{L}_{\text{2D}}$ is a 2D keypoint re-projection loss encouraging alignment with image cues, and $\mathcal{L}_{\text{obs}}$ is a loss penalizing deviation between "observed" and GT InterFields, encouraging consistency between the estimated poses and InterFields. To this end, it first poses the object relative to the body using the estimated transformations $\{\hat{R}^o, \hat{t^o}\}$, computes the "observed" InterFields between the posed body and object meshes, and penalizes their deviation from the GT InterField. The steering weights λ are set empirically; see "Implementation Details" in Sec. 3.4.

3.3 InterField-Guided Refinement within LEXIS-Flow

Our LEXIS-Flow (Sec. 3.2) is a flow-based generative model, so it imposes a prior that pulls estimations towards the training-data distribution. However, for rare/unseen images and interactions it can naturally make errors. Typically, optimization performs corrections via contact constraints [82], but is slow, prone to local minima, and takes place post-hoc, i.e., in a *separate* stage *after* inference.

We tackle this limitation with a *guided refinement*; LEXIS-Flow *itself* performs the refinement in inference time (*without* additional stages) by adding gradient-based guidance steps in the ODE sampling loop. Specifically, at intermediate timesteps t, the ODE solver pauses, and the current state $\mathbf{x}_t$ is updated via a guidance gradient before the next iteration step.

To this end, we add a guidance loss with the terms discussed in the following:

$$\mathcal{L}_{\text{guide}} = \mathcal{L}_{\text{pose+IF}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}. \tag{6}$$

InterField Guidance ($\mathcal{L}_{IF}$): The LEXIS-Net decoder, D_ψ, maps the running estimations of latent tokens, $\hat{\mathcal{Z}}$, to a predicted InterField, $\hat{\text{IF}} = D_\psi(\hat{\mathcal{Z}}, \mathbf{f}^o)$, which encodes the proximal relationships between interacting surfaces. Moreover, an "observed" InterField is computed based on the running estimations for the body and object meshes. Then, the novel loss term $\mathcal{L}_{\text{pose+IF}}$ penalizes the discrepancy between the "observed" and estimated InterFields:

$$\mathcal{L}_{\text{pose+IF}} = \|\hat{\text{IF}} - \text{IF}_t\|^2, \tag{7}$$

where IF_t denotes the computed InterField by posing body (decoded from running $\hat{\mathcal{Z}}$) and object with running estimates $\{R_t^b, t_t^b\}, \{R_t^o, t_t^o\}$, respectively. This acts as a field that pulls the object and body into a 3D configuration informed by the LEXIS interaction tokens and their corresponding InterFields, in a form of LEXIS-/InterField-based analysis-by-synthesis.

Mask-pixel Guidance ($\mathcal{L}_{\text{mask}}$): To align 3D estimations with image cues, a common way is to use mask-based losses. Doing this for the human is challenging, as people have hair, accessories, and loose clothing that "contaminate" masks and cannot be explained by parametric body models such as SMPL+H. However, doing this for the object is reliable, because the recent SAM [39] model estimates

robust 2D masks, and SAM3D [71] provides good 3D shape estimates (though pose estimates are noisy). Thus, we add a mask-based loss for the object.

Object vertices $\mathbf{V}^o$, transformed by running state estimate (R^o, t^o), are projected onto the image plane using a differentiable renderer [63]. Vertices falling outside the observed 2D object mask [39], M^o, are penalized by:

$$\mathcal{L}_{\text{mask}} = \sum\nolimits_{\mathbf{v}\in\mathbf{V}^o} \|1 - M^o\big(\pi(R_t^o\mathbf{v} + t_t^o)\big)\|^2, \tag{8}$$

providing a gradient signal that aligns the 3D shape with the 2D image cues.

Guided Flow Trajectory: The state $\mathbf{x}_t$ is updated by a gradient step:

$$\mathbf{x}_t \leftarrow \mathbf{x}_t - \eta \cdot \nabla_{\mathbf{x}_t}(\mathcal{L}_{\text{pose+IF}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}). \tag{9}$$

Correcting $\mathbf{x}_t$ during sampling leads to a *guided flow trajectory*, dynamically balancing interaction awareness (LEXIS-based InterFields) and image cues (mask). This update performs guided refinement, *without* additional, separate stages.

Importantly, this formulation also lets LEXIS-Flow exploit off-the-shelf SotA 3D body-only and object-only estimators that are *not* interaction-aware, so that it has a better initialization for adding the missing interaction awareness. That is, given rough initial 3D body and object estimations from a baseline model, LEXIS-Flow encodes these into a unified latent state $\mathbf{x}_{\text{init}} = \{\mathcal{B}_{\text{init}}, \mathcal{O}_{\text{init}}\}$. Then, rather than generating a 3D HOI estimation from pure noise, a controlled amount of noise is injected to transport this estimate to an *intermediate* flow timestep $t_{\text{start}} \in (0, 1)$: $\mathbf{x}_{t_{\text{start}}} = (1 - t_{\text{start}})\,\boldsymbol{\epsilon} + t_{\text{start}}\,\mathbf{x}_{\text{init}}$, as in SDEdit [51]. A partial ODE integration then proceeds from t_{start} to t=1 with the guided sampling. In the "Experiments"section (Sec. 4), this variant is denoted as LEXIS-Flow*.

3.4 Implementation Details

LEXIS: We initialize LEXIS-Net VQ-VAE with TokenHMR [18] tokenizer and further train it to InterFields (evaluated at 1024 uniform surface points). We use ω=5 for exponential mapping of InterFields. The LEXIS codebook contains K=2048 entries (D=128). A frozen PointNeXt [61] (pretrained on ModelNet40 [79]) encodes object geometry into a 1024-D global descriptor. This conditions the $\mathcal{T}^b$ and $\mathcal{T}^o$ TriPlanes (T_r=64 resolution, T_c=64 channels), which feed a 2-layer MLP ([128, 64] hidden units). We optimize Eq. (3) (λ^h_{IF}=10, λ^o_{IF}=10, λ_{VQ}=λ_{commit}=0.5) for 60,000 iterations using AdamW [50] (lr=10^{-4}, batch size 1024). Training takes ∼5 hours on a single RTX-6000Ada.

LEXIS-Flow: LEXIS-Flow uses a 18-layer multi-stream transformer backbone [71] (512 hidden units). As image encoder we use frozen DINOv2 (ViT/L-14) [53]. Optimizing the training loss of Eq. (4) (λ_{IF}=10, λ_{v2v}=5, λ_{2D}=1, λ_{obs}=0.5) runs for 100k steps via AdamW [50] (batch size 1024). We use a warmup cosine plateau scheduler (learning rate 10^{-6} to $5{\times}10^{-4}$) and a 0.3 dropout rate for classifier-free guidance, with guidance-scale 1.5. Training takes ∼48 hours on 4 RTX-6000Ada GPUs. We use Euler ODE solver (25 steps) for sampling and perform guidance/refinement (Eq. (9)) starting at t_{start}=15, λ_{mask}=0.5.

4 Experiments

4.1 Experimental Setup

Training Data: We train LEXIS-Net on geometry-only HOI datasets, combining the synthetic ProciGen [81] with InterAct-refined [84] MoCap datasets: NeuralDome [91], OMOMO [41], and IMHD [93]. We train LEXIS-Flow on ProciGen 3D-paired images for image-conditioned generation, and include 3D-only InterAct-refined data for classifier-free guidance.

Benchmarks: We use the in-the-wild Open3DHOI [78] (Sec. 4.3) and the in-lab BEHAVE [4] dataset (Sec. 4.4). For the former we have 2 tasks; see below, where our generative model (sampling from noise) is denoted as LEXIS-Flow and a refinement variant (initialized by SotA models via SDEdit [51]) as LEXIS-Flow*.

Generative Reconstruction (Open3DHOI): LEXIS-Flow estimates 3D *directly from noise*, conditioned on image features, without any external initialization for body or object pose. For fair comparison, we train LEXIS-Flow only on ProciGen, as the HDM [81] baseline. For object shape, we provide the SAM3D shape estimate as input to LEXIS-Flow. For HDM, we use its object-specific version for classes it was trained on, and its general version for all others.

Guided Refinement (Open3DHOI): LEXIS-Flow* starts from off-the-shelf estimates rather than noise; it uses CameraHMR [56] for the body and SAM3D [71] for the object, aligned to a MoGe [76] depth estimate; see alignment details in Sec. S.1.1. This initialization is encoded into the latent state and transported to an intermediate timestep t_{start}=15 (out of total 25) via SDEdit [51], after which InterField-guided ODE integration refines estimates (Sec. 3.3). LEXIS-Flow* trains on a combination of the ProciGen and InterAct-refined datasets. We compare against the optimization-based InteractVLM [16] and HOI-Gaussian [78] methods, initialized with CameraHMR + SAM3D + MoGe as our method for fair comparison. We refer to this version of InteractVLM as "InteractVLM++." Originally, InteractVLM uses ICP-based initialization of object pose, which struggles in difficult scenarios; for completeness, we also report this original setting, denoted as "InteractVLM."

In-lab Evaluation (BEHAVE): For fair comparison to baselines that train on BEHAVE, we train dataset-specific variants of LEXIS-Net and LEXIS-Flow on this dataset, using the official train/test split [82]. We evaluate under the generative setting only; for this, LEXIS-Flow generates the HOI state from noise without any external expert initialization. We compare against CHORE [82], CONTHO [52], HOI-TG [77], and HDM [81], all trained on BEHAVE.

Evaluation Metrics: We apply Procrustes alignment to the predicted body meshes and apply the same transformation to the object as in [52, 77], before computing all metrics. To tackle the different topology of SMPL-X [57] (Open3DHOI) and SMPL+H [64] (ours), we align their common vertices and evaluate on uniformly-sampled surface points. We report Chamfer Distances (CD_{hum}, CD_{obj}; cm, lower is better) for geometry, Collision (% of penetrating vertices [78]; lower is better) for physical plausibility, and Contact F1 (@5cm [52]; higher is better) for interaction fidelity. For metric formulas, see Sec. S.1.2.

Method	CD_{hum} ↓	CD_{obj} ↓	Collision ↓	Contact ↑
A. HDM (w. scale align.) [81]	13.50 (34.4%)	49.38 (29.1%)	0.089 (32.6%)	0.141 (49.6%)
B. LEXIS-Flow (Ours)	**8.85**	**35.01**	0.060	0.211
C. InteractVLM [16]	7.20 (2.1%)	38.20 (39.9%)	0.054 (24.1%)	0.372 (21.2%)
D. CameraHMR [56] + SAM3D [71]	7.20 (2.1%)	37.30 (38.4%)	0.051 (19.6%)	0.182 (147%)
E. HOI-Gaussian [78]	7.28 (3.2%)	32.02 (28.3%)	0.061 (32.8%)	0.151 (198%)
F. InteractVLM++ [16]	7.20 (2.1%)	30.11 (23.7%)	0.047 (12.8%)	0.394 (14.5%)
G. LEXIS-Flow* (Ours)	**7.05**	**22.96**	**0.041**	**0.451**

Table 1: 3D HOI in the wild (Sec. 4.3). We evaluate our LEXIS-Flow against SotA refinement-based methods on the Open3DHOI [78] benchmark. Rows A–B estimate from scratch; row C initializes with CameraHMR, and rows E–G initialize with D.

4.2 Interaction Representation

To evaluate the effect of the interaction representation, we compare the dense, continuous InterFields with sparse, binary contacts; for fairness, contacts are extracted by thresholding InterFields to compute a level set. We evaluate on two tasks: (1) The render-and-compare fitting of InteractVLM [16]. (2) The guided generative sampling of LEXIS-Flow.

Task	Representation	CD_{obj} ↓	Contact ↑
Fitting	Binary Contact	35.4	17.8
Fitting	InterField (ours)	**34.6**	**19.8**
Generative	Binary Contact	49.11	0.152
Generative	InterField (ours)	**41.01**	**0.290**

Table 2: Representation effect (Sec. 4.2). Contacts and InterFields are compared on Open3DHOI [78] on two tasks: render-and-compare fitting and generative reconstruction.

The evaluation results are shown in Tab. 2. For both tasks, the dense, continuous InterFields outperform the sparse, binary contacts. Specifically, for the "fitting" task, CD_{obj} improves by 2.26% and Contact by 12.24%; for "generative" task the improvement is 16.49% and 90.79% respectively. This shows that the dense, continuous InterField representation encodes richer spatial signal than sparse, binary contacts, helping 3D reconstruction and human-object alignment.

4.3 In-the-wild 3D HOI Reconstruction

We evaluate 3D HOI estimation on the in-the-wild images of Open3DHOI [78]; we evaluate two LEXIS-Flow versions, for "generative reconstruction," and for "guided refinement." The results are shown in Tab. 1 and Fig. 6.

Generative Reconstruction (Tab. 1): LEXIS-Flow (row B) outperforms HDM [81] (row A) on CD_{hum} (8.8 vs 13.5 cm), CD_{obj} (35.0 vs 49.3 cm) and Contact F1 (0.21 vs 0.14); note in Fig. 6 that HDM produces geometric artifacts and intersecting meshes on real-world images. Notably, LEXIS-Flow (row B), with its direct generative estimation, already achieves lower CD_{obj} and higher Contact F1 than a baseline that combines off-the-shelf CameraHMR + SAM3D (row D) estimations (aligned using MoGe [76] depth). This suggests that jointly estimating humans and objects by exploiting interaction signatures (InterFields) (row B) produces better spatial configurations than independently combining per-entity experts (row D). When we initialize our method with these off-the-shelf estimates (row D), denoted as LEXIS-Flow*(row G), it improves further; see below.

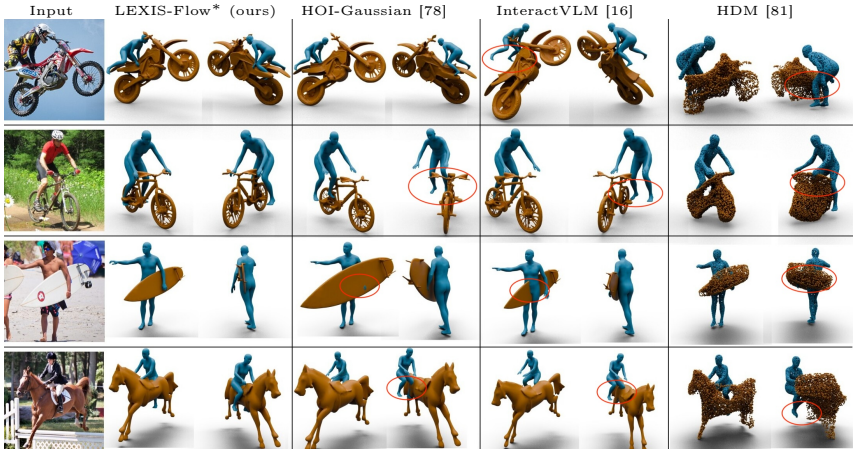

Fig. 6: LEXIS-Flow vs SotA. Existing methods often fail to capture tight physical coupling, yielding floating objects (HOI-Gaussian, InteractVLM) or penetrations (HDM). LEXIS-Flow tackles this via dense InterFields for proximity-aware estimation.

Guided Refinement (Tab. 1): With identical initialization (row D–G) LEXIS-Flow* (row G) achieves the best results across all four metrics, outperforming both InteractVLM (row C) and HOI-Gaussian (row E). LEXIS-Flow* also outperforms InteractVLM++ (row F) (initialized with CameraHMR + SAM3D + MoGe; see Sec. 4.1). This shows that the gains are not solely due to better initialization; even when InteractVLM benefits from the same starting point, the InterField-guided refinement of LEXIS-Flow* yields stronger spatial alignment. The key difference is that the optimization baselines refine in output space with sparse contact losses, while LEXIS-Flow* corrects intermediate latent states during generation using dense InterField signals from the learned LEXIS.

Qualitative comparison: Figure 6 compares our LEXIS-Flow* method against SotA methods in the wild; for more results see Sec. S.2.1. Optimization-based methods often produce floating objects or penetration, because sparse contacts lack the spatial awareness to correct these errors. LEXIS-Flow* recovers physically-plausible interactions via InterField-guided refinement. Additional results in Fig. 7 show our in-the-wild generalization; for more results see Sec. S.2.1.

Perceptual study: We conduct a perceptual study on 60 Open3DHOI images with 62 participants (protocol in Sec. S.2.2). This shows that LEXIS-Flow* estimations are perceived as more realistic 75.8% of times over HOI-Gaussian.

4.4 In-lab 3D HOI Reconstruction

We evaluate in-lab HOI reconstruction on the BEHAVE benchmark (Tab. 3). LEXIS-Flow resets the SotA performance on the BEHAVE benchmark by a significant margin of 15.0% and 5.0% on CD_{hum} and CD_{obj} respectively relative to the closest competitor, HOI-TG [77].

	BEHAVE	
	CD_{hum} ↓	CD_{obj} ↓
PHOSA [90]	12.17 (68.0%)	26.62 (71.5%)
HDM [81]	11.61 (66.4%)	11.35 (33.0%)
CHORE [82]	5.58 (30.1%)	10.66 (28.7%)
CONTHO [52]	4.99 (21.8%)	8.42 (9.7%)
HOI-TG [77]	4.59 (15.0%)	8.00 (5.0%)
LEXIS-Flow	**3.90**	**7.60**

Table 3: In-lab 3D HOI: Evaluation on the in-lab BEHAVE [4] dataset.

Fig. 7: Qualitative results. LEXIS-Flow* recovers physically-plausible interactions from diverse in-the-wild images. By leveraging dense InterField signals, our model produces accurate spatial configurations and realistic articulations even under occlusion.

4.5 Ablation Study

Table 4 evaluates our design choices on the Open3DHOI [78] dataset.

Architecture (unguided): When replacing the discrete LEXIS codebook with a continuous VAE latent (*GaussFlow*) degrades CD_{hum} by 39% and CD_{obj} by 19%. This shows that discrete tokenization preserves pose expressivity and interaction structure.

	Variants	$\text{CD}_{hum} \downarrow$	$\text{CD}_{obj} \downarrow$
Architect. (Unguided)	GaussFlow Baseline	13.45	57.25
	FlowObject Baseline	N/A	65.05
	LEXIS-Flow Baseline	**9.68**	**48.01**
Guidance	Unguided	9.68	48.01
	$\mathcal{L}_{\text{mask}}$	9.46	43.51
	$\mathcal{L}_{\text{pose+IF}}$	9.05	41.01
	$\mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{pose+IF}}$	**8.85**	**35.01**

Table 4: Ablations: We evaluate design choices for our architecture (unguided) and for guided refinement on Open3DHOI [78].

After removing LEXIS, we keep the human fixed (initialized with the SotA CameraHMR method) and denoise object pose (*FlowObject*). This worsens performance, showing that jointly denoising the body and object with LEXIS is essential; note that CD_{hum} is not reported as body parameters are fixed.

Guidance signal (Sec. 3.3): Adding mask guidance ($\mathcal{L}_{\text{mask}}$) reduces CD_{obj} to 43.51. InterField guidance ($\mathcal{L}_{\text{pose+IF}}$) achieves CD_{obj} of 41.01. Combining both yields the best CD_{obj} (35.01), a 27% reduction over unguided sampling. This shows that 2D masks and 3D InterField proximity are complementary signals.

5 Conclusion

We go beyond sparse, binary contact by leveraging dense, continuous InterFields for reconstructing 3D Human-Object Interaction from single images. To make InterField inference tractable from a single image, we learn LEXIS, a latent manifold of interaction signatures, encoding interaction- and object-specific proximity patterns. Then, we develop LEXIS-Flow, a dual-stream Flow-Matching model that jointly estimates 3D human and object meshes alongside their InterFields, and exploits these for guided refinement, eliminating post-hoc optimization. Experiments show that continuous InterFields outperform binary contacts in both fitting and generative settings. On the Open3DHOI benchmark, LEXIS-Flow* achieves the lowest CD_{obj} (22.96) and highest Contact F1 (0.451), outperforming all baselines. This moves us closer to holistic 3D scene understanding.

Acknowledgments

We thank Božidar Antić and Ilya Petrov for valuable insights and discussions. SKD is supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS). We acknowledge HPC support by the EuroHPC Joint Undertaking that awarded access to the EuroHPC supercomputers LEONARDO (project ID EHPC-AI-2024A06-077), hosted by CINECA in Italy, and JUPITER (project ID e-reg-2025r02-393), hosted by JSC in Germany, and by the Dutch national e-infrastructure through the SURF Cooperative grant no. EINF-12852. We also acknowledge support through a research gift from Google, and the NVIDIA Academic Grant Program. This work is supported by the European Research Council (ERC) through the Starting Grant (project STRIPES, Grant agreement ID: 101165317, DOI: 10.3030/101165317, PI: D. Tzionas).

References

1. Antić, D., Paschalidis, G., Tripathi, S., Gevers, T., Dwivedi, S.K., Tzionas, D.: SDFit: 3D object pose and shape by fitting a morphable SDF to a single image. In: International Conference on Computer Vision (ICCV). pp. 9616–9626 (2025)
2. Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., Zhu, J.: One transformer fits all distributions in multi-modal diffusion at scale. In: International Conference on Machine Learning (ICML). pp. 1692–1717 (2023)
3. Bernardin, K., Ogawara, K., Ikeuchi, K., Dillmann, R.: A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. Transactions on Robotics (T-RO) **21**(1), 47–57 (2005)
4. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: BEHAVE: Dataset and method for tracking human object interactions. In: Computer Vision and Pattern Recognition (CVPR). pp. 15914–15925 (2022)
5. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European Conference on Computer Vision (ECCV). vol. 9909, pp. 561–578 (2016)
6. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: ContactPose: A dataset of grasps with object contact and hand pose. In: European Conference on Computer Vision (ECCV). vol. 12358, pp. 361–378 (2020)
7. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: Computer Vision and Pattern Recognition (CVPR). pp. 16102–16112 (2022)
8. Chen, Y., Dwivedi, S.K., Black, M.J., Tzionas, D.: Detecting human-object contact in images. In: Computer Vision and Pattern Recognition (CVPR). pp. 17100–17110 (2023)
9. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: Computer Vision and Pattern Recognition (CVPR). pp. 5939–5948 (2019)
10. Cheng, Y.C., Lee, H.Y., Tuyakov, S., Schwing, A., Gui, L.: SDFusion: Multimodal 3D shape completion, reconstruction, and generation. In: Computer Vision and Pattern Recognition (CVPR). pp. 4456–4465 (2023)
11. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In: European Conference on Computer Vision (ECCV). vol. 9912, pp. 628–644 (2016)
12. Corona, E., Pons-Moll, G., Alenyà, G., Moreno-Noguer, F.: Learned vertex descent: A new direction for 3D human model fitting. In: European Conference on Computer Vision (ECCV). pp. 146–165 (2022)
13. Cseke, A., Tripathi, S., Dwivedi, S.K., Lakshmipathy, A., Chatterjee, A., Black, M.J., Tzionas, D.: PICO: Reconstructing 3D people in contact with objects. In: Computer Vision and Pattern Recognition (CVPR). pp. 1783–1794 (2025)
14. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3D objects. In: Computer Vision and Pattern Recognition (CVPR). pp. 13142–13153 (2023)
15. Diller, C., Dai, A.: CG-HOI: Contact-guided 3D human-object interaction generation. In: Computer Vision and Pattern Recognition (CVPR). pp. 19888–19901 (2024)
16. Dwivedi, S.K., Antić, D., Tripathi, S., Taheri, O., Schmid, C., Black, M.J., Tzionas, D.: InteractVLM: 3D interaction reasoning from 2D foundational models. In: Computer Vision and Pattern Recognition (CVPR). pp. 22605–22615 (2025)

17. Dwivedi, S.K., Schmid, C., Yi, H., Black, M.J., Tzionas, D.: POCO: 3D pose and shape estimation using confidence. In: International Conference on 3D Vision (3DV). pp. 85–95 (2024)
18. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In: Computer Vision and Pattern Recognition (CVPR). pp. 1323–1333 (2024)
19. Fan, H., Su, H., Guibas, L.: A point set generation network for 3D object reconstruction from a single image. In: Computer Vision and Pattern Recognition (CVPR). pp. 2463–2471 (2017)
20. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Computer Vision and Pattern Recognition (CVPR). pp. 12943–12954 (2023)
21. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The GRASP taxonomy of human grasp types. Transactions on Human-Machine Systems (THMS) **46**(1), 66–77 (2016)
22. Fiche, G., Leglaive, S., Alameda-Pineda, X., Agudo, A., Moreno-Noguer, F.: VQ-HPS: Human pose and shape estimation in a vector-quantized latent space. In: European Conference on Computer Vision (ECCV). vol. 15110, pp. 471–490 (2024)
23. Gkioxari, G., Malik, J., Johnson, J.: Mesh R-CNN. In: International Conference on Computer Vision (ICCV). pp. 9784–9794 (2019)
24. Goodwin, W., Vaze, S., Havoutis, I., Posner, I.: Zero-shot category-level object pose estimation. In: European Conference on Computer Vision (ECCV). vol. 13699, pp. 516–532 (2022)
25. Grady, P., Tang, C., Brahmbhatt, S., Twigg, C.D., Wan, C., Hays, J., Kemp, C.C.: PressureVision: Estimating hand pressure from a single RGB image. In: European Conference on Computer Vision (ECCV). vol. 13666, pp. 328–345 (2022)
26. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: ContactOpt: Optimizing contact to improve grasps. In: Computer Vision and Pattern Recognition (CVPR). pp. 1471–1481 (2021)
27. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: HOnnotate: A method for 3D annotation of hand and object poses. In: Computer Vision and Pattern Recognition (CVPR). pp. 3196–3206 (2020)
28. Han, S., Joo, H.: CHORUS: Learning canonicalized 3D human-object spatial relations from unbounded synthesized images. In: International Conference on Computer Vision (ICCV). pp. 15789–15800 (2023)
29. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: International Conference on Computer Vision (ICCV). pp. 2282–2292 (2019)
30. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D scenes by learning human-scene interaction. In: Computer Vision and Pattern Recognition (CVPR). pp. 14708–14718 (2021)
31. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: Computer Vision and Pattern Recognition (CVPR). pp. 13274–13285 (2022)
32. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. International Journal of Computer Vision (IJCV) **132**(7), 2551–2566 (2024)
33. Huang, Z., Jampani, V., Thai, A., Li, Y., Stojanov, S., Rehg, J.M.: ShapeClipper: Scalable 3D shape learning from single-view images via geometric and CLIP-based

consistency. In: Computer Vision and Pattern Recognition (CVPR). pp. 12912–12922 (2023)
34. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: International Conference on Computer Vision (ICCV). pp. 11087–11096 (2021)
35. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3D deformation model for tracking faces, hands, and bodies. In: Computer Vision and Pattern Recognition (CVPR). pp. 8320–8329 (2018)
36. Kamakura, N., Matsuo, M., Ishii, H., Mitsuboshi, F., Miura, Y.: Patterns of static prehension in normal hands. American Journal of Occupational Therapy **34**(7), 437–445 (1980)
37. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)
38. Kim, H., Han, S., Kwon, P., Joo, H.: Beyond the contact: Discovering comprehensive affordance for 3D objects from pre-trained 2D diffusion models. In: European Conference on Computer Vision (ECCV). pp. 400–419 (2024)
39. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: International Conference on Computer Vision (ICCV). pp. 4015–4026 (2023)
40. Kocabas, M., Athanasiou, N., Black, M.: VIBE: Video inference for human body pose and shape estimation. In: Computer Vision and Pattern Recognition (CVPR). pp. 5252–5262 (2020)
41. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. Transactions on Graphics (TOG) **42**(6), 197:1–197:11 (2023)
42. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: International Conference on Computer Vision (ICCV). pp. 12919–12928 (2021)
43. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (ICLR) (2023)
44. Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., Su, H.: OpenShape: Scaling up 3D shape representation towards open-world understanding. In: Conference on Neural Information Processing Systems (NeurIPS) (2023)
45. Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. In: Conference on Neural Information Processing Systems (NeurIPS) (2024)
46. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3D object. In: International Conference on Computer Vision (ICCV). pp. 9264–9275 (2023)
47. Liu, S., Zhou, Y., Yang, J., Gupta, S., Wang, S.: ContactGen: Generative contact modeling for grasp generation. In: International Conference on Computer Vision (ICCV). pp. 20552–20563 (2023)
48. Liu, Z., Zhou, D., Lu, F., Fang, J., Zhang, L.: AutoShape: Real-time shape-aware monocular 3D object detection. In: International Conference on Computer Vision (ICCV). pp. 15621–15630 (2021)
49. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. Transactions on Graphics (TOG) **34**(6), 248:1–248:16 (2015)
50. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)

51. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2022)
52. Nam, H., Jung, D.S., Moon, G., Lee, K.M.: Joint reconstruction of 3D human and object via contact-based refinement transformer. In: Computer Vision and Pattern Recognition (CVPR). pp. 10218–10227 (2024)
53. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR) (2024)
54. Park, J.J., Florence, P.R., Straub, J., Newcombe, R.A., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: Computer Vision and Pattern Recognition (CVPR). pp. 165–174 (2019)
55. Paschalidou, D., Ulusoy, A.O., Geiger, A.: Superquadrics revisited: Learning 3D shape parsing beyond cuboids. In: Computer Vision and Pattern Recognition (CVPR). pp. 10344–10353 (2019)
56. Patel, P., Black, M.: CameraHMR: Aligning people with perspective. In: International Conference on 3D Vision (3DV). pp. 1562–1571 (2025)
57. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
58. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: International Conference on Computer Vision (ICCV). pp. 4172–4182 (2023)
59. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: TriDi: Trilateral diffusion of 3D humans, objects, and interactions. In: International Conference on Computer Vision (ICCV). pp. 5523–5535 (2025)
60. Qi, H., Zhao, C., Salzmann, M., Mathis, A.: HOISDF: Constraining 3D hand-object pose estimation with global signed distance fields. In: Computer Vision and Pattern Recognition (CVPR). pp. 10392–10402 (2024)
61. Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B.: PointNeXt: Revisiting pointnet++ with improved training and scaling strategies. In: Conference on Neural Information Processing Systems (NeurIPS) (2022)
62. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., Ghanem, B.: Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. In: International Conference on Learning Representations (ICLR) (2024)
63. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3D deep learning with PyTorch3D. arXiv:2007.08501 (2020)
64. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. Transactions on Graphics (TOG) **36**(6), 245:1–245:17 (2017)
65. Sárándi, I., Pons-Moll, G.: Neural localizer fields for continuous 3D human pose and shape estimation. In: Conference on Neural Information Processing Systems (NeurIPS) (2024)
66. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: PiGraphs: Learning interaction snapshots from observations. Transactions on Graphics (TOG) **35**(4), 139 (2016)

67. Shimada, S., Golyanik, V., Li, Z., Pérez, P., Xu, W., Theobalt, C.: HULC: 3D human motion capture with pose manifold sampling and dense contact guidance. In: European Conference on Computer Vision (ECCV). vol. 13682, pp. 516–533 (2022)
68. Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: GOAL: Generating 4D whole-body motion for hand-object grasping. In: Computer Vision and Pattern Recognition (CVPR). pp. 13263–13273 (2022)
69. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV). vol. 12349, pp. 581–600 (2020)
70. Taheri, O., Zhou, Y., Tzionas, D., Zhou, Y., Ceylan, D., Pirk, S., Black, M.J.: GRIP: Generating interaction poses using spatial cues and latent consistency. In: International Conference on 3D Vision (3DV). pp. 933–943 (2024)
71. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: SAM 3D: 3Dfy anything in images. arXiv:2511.16624 (2025)
72. Tripathi, S., Chatterjee, A., Passy, J., Yi, H., Tzionas, D., Black, M.J.: DECO: Dense estimation of 3D human-scene contact in the wild. In: International Conference on Computer Vision (ICCV). pp. 7967–7979 (2023)
73. Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: Computer Vision and Pattern Recognition (CVPR). pp. 4713–4725 (2023)
74. Wang, H., Sridhar, S., Huang, J., Valentin, J.P.C., Song, S., Guibas, L.: Normalized object coordinate space for category-level 6D object pose and size estimation. In: Computer Vision and Pattern Recognition (CVPR). pp. 2642–2651 (2019)
75. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2Mesh: Generating 3D mesh models from single RGB images. In: European Conference on Computer Vision (ECCV). vol. 11215, pp. 55–71 (2018)
76. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Computer Vision and Pattern Recognition (CVPR). pp. 5261–5271 (June 2025)
77. Wang, Z., Zheng, Q., Ma, S., Ye, M., Zhan, Y., Li, D.: End-to-end HOI reconstruction transformer with graph-based encoding. In: Computer Vision and Pattern Recognition (CVPR). pp. 27706–27715 (2025)
78. Wen, B., Huang, D., Zhang, Z., Zhou, J., Deng, J., Gong, J., Chen, Y., Ma, L., Li, Y.: Reconstructing in-the-wild open-vocabulary human-object interactions. In: Computer Vision and Pattern Recognition (CVPR). pp. 17426–17436 (2025)
79. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D shapenets: A deep representation for volumetric shapes. In: Computer Vision and Pattern Recognition (CVPR). pp. 1912–1920 (2015)
80. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In: Proceedings of Robotics: Science and Systems. Pittsburgh, Pennsylvania (June 2018)
81. Xie, X., Bhatnagar, B.L., Lenssen, J.E., Pons-Moll, G.: Template free reconstruction of human-object interaction with procedural interaction generation. In: Computer Vision and Pattern Recognition (CVPR). pp. 10003–10015 (2024)
82. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: CHORE: Contact, Human and Object REconstruction from a single RGB image. In: European Conference on Computer Vision (ECCV). pp. 125–145 (2022)

83. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: Generative 3D human shape and articulated pose models. In: Computer Vision and Pattern Recognition (CVPR). pp. 6183–6192 (2020)
84. Xu, S., Li, D., Zhang, Y., Xu, X., Long, Q., Wang, Z., Lu, Y., Dong, S., Jiang, H., Gupta, A., Wang, Y.X., Gui, L.Y.: InterAct: Advancing large-scale versatile 3D human-object interaction generation. In: Computer Vision and Pattern Recognition (CVPR). pp. 7048–7060 (2025)
85. Xue, M., Liu, Y., Guo, L., Huang, S., Ding, C.: Guiding human-object interactions with rich geometry and relations. In: Computer Vision and Pattern Recognition (CVPR). pp. 22714–22723 (2025)
86. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: CPF: Learning a contact potential field to model the hand-object interaction. In: International Conference on Computer Vision (ICCV). pp. 11097–11106 (2021)
87. Yang, X., Kukreja, D., Pinkus, D., Sagar, A., Fan, T., Park, J., Shin, S., Cao, J., Liu, J., Ugrinovic, N., Feiszli, M., Malik, J., Dollar, P., Kitani, K.: SAM 3D body: Robust full-body human mesh recovery. arXiv:2602.15989 (2025)
88. Yang, Y., Zhai, W., Luo, H., Cao, Y., Zha, Z.: LEMON: Learning 3D human-object interaction relation from 2D images. In: Computer Vision and Pattern Recognition (CVPR). pp. 16284–16295 (2024)
89. Zhang, H., Ye, Y., Shiratori, T., Komura, T.: ManipNet: Neural manipulation synthesis with a hand-object spatial representation. Transactions on Graphics (TOG) **40**(4), 121:1–121:14 (2021)
90. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3D human-object spatial arrangements from a single image in the wild. In: European Conference on Computer Vision (ECCV). pp. 34–51 (2020)
91. Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., Yu, J., Xu, L., Wang, J.: NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In: Computer Vision and Pattern Recognition (CVPR). pp. 8834–8845 (2023)
92. Zhang, W., Dabral, R., Golyanik, V., Choutas, V., Alvarado, E., Beeler, T., Habermann, M., Theobalt, C.: BimArt: A unified approach for the synthesis of 3D bimanual interaction with articulated objects. In: Computer Vision and Pattern Recognition (CVPR). pp. 27694–27705 (2024)
93. Zhao, C., Zhang, J., Du, J., Shan, Z., Wang, J., Yu, J., Wang, J., Xu, L.: I'M HOI: Inertia-aware monocular capture of 3D human-object interactions. In: Computer Vision and Pattern Recognition (CVPR). pp. 729–741 (2024)
94. Zheng, Z., Yu, T., Dai, Q., Liu, Y.: Deep implicit templates for 3D shape representation. In: Computer Vision and Pattern Recognition (CVPR). pp. 1429–1439 (2021)
95. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Computer Vision and Pattern Recognition (CVPR). pp. 5745–5753 (2019)

Supplementary Material

LEXIS: LatEnt proXimal Interaction Signatures for 3D HOI from an Image

S.1 Implementation Details

S.1.1 MoGe-Based Object Initialization

For each method (D–G) in Tab. 1 (in main), we apply the same initialization, as follows. We initialize the object pose R^o, t^o from SAM3D [71]. The rotation estimate is empirically accurate, so we keep it unchanged. However, the translation and scale are noisy. We refine these via MoGe-estimated [76] metric depth. Let the median MoGe depths under body and object masks [39] be $\bar{z}^b$, $\bar{z}^o$, respectively. Their ratio $\rho = \bar{z}^o/\bar{z}^b$ is scale-invariant. Anchoring to the metric body depth z^b from CameraHMR [56] gives the depth-correction factor $\delta = \rho\, z^b/t_z^o$, where t_z^o is the z-component of t^o. We resolve the scale ambiguity by anchoring the object's metric size to the body dimensions, obtaining $\sigma_{\text{moge}} = \sigma_{smpl} L^o/L^h$, where σ_{smpl} is the height of the template SMPL mesh, L^o, L^b are bounding-box heights of the object and body, respectively, as extracted from MoGe depth map. Then, we blend σ_{moge} with SAM3D-estimated scale σ_{sam3d}: $\sigma = (\delta\, \sigma_{\text{sam3d}})^{0.8} \cdot \sigma_{\text{moge}}^{0.2}$.

S.1.2 Evaluation Metrics

For the human body, as in [52, 77], we apply Procrustes alignment on the estimated SMPL+H mesh (to align it to the GT SMPL-X mesh), via a similarity transform $\{s, \mathbf{R}, \mathbf{t}\}$ computed on vertices that are common between SMPL+H and SMPL-X. For the object, we apply only the rigid transform $\{\mathbf{R}, \mathbf{t}\}$ (from above) to preserve and evaluate the estimated scale, s. After the above, we compute all metrics on 10,000 points uniformly sampled across the aligned surfaces.

Chamfer Distance (CD): A bidirectional surface-to-surface distance (cm):

$$\text{CD}(\widehat{\mathbf{V}}, \mathbf{V}) = \frac{1}{|\mathbf{V}|} \sum_{\mathbf{v} \in \mathbf{V}} \min_{\hat{\mathbf{v}} \in \widehat{\mathbf{V}}} \|\mathbf{v} - \hat{\mathbf{v}}\| + \frac{1}{|\widehat{\mathbf{V}}|} \sum_{\hat{\mathbf{v}} \in \widehat{\mathbf{V}}} \min_{\mathbf{v} \in \mathbf{V}} \|\mathbf{v} - \hat{\mathbf{v}}\|, \tag{S.1}$$

capturing the 3D geometric error between an estimated and GT mesh with vertex sets $\widehat{\mathbf{V}}$ and $\mathbf{V}$, respectively. Lower is better (↓).

Contact F1: This captures the agreement between the estimated ($\hat{M}$) and ground-truth (M) binary contact masks, $M \in \{0,1\}^N$. For a human body surface point $p^b \in \mathbb{R}^3$, binary contact is defined as $\mathbb{I}(\min_{p^o \in \mathcal{O}} \|p^b - p^o\| \leq 5\text{cm})$ by finding the closest point p^o on the object surface, $\mathcal{O}$, and thresholding the respective distance. Moreover, $\mathbb{I}(\cdot)$ is the indicator function, with $\mathbb{I}(a < b) = 1$ iff $a < b$. Let $TP = |\hat{M} \cap M|$ be the number of true positive contact points. We compute:

$$P = {}^{TP}/_{|\hat{M}|}, \quad R = {}^{TP}/_{|M|}, \quad F1 = {}^{2PR}/_{P+R}, \tag{S.2}$$

where P is Precision and R is Recall. Higher is better (↑).

Collision Score: This quantifies the fraction of body vertices that penetrate into the object mesh:

$$\text{Collision}(\mathbf{V}^b, \mathbf{V}^o) = \frac{1}{|\mathbf{V}^b|} \sum_{\mathbf{v}^b \in \mathbf{V}^b} \mathbb{I}(\mathbf{v}^b \text{ in } \mathbf{V}^o), \tag{S.3}$$

where $\mathbf{V}^b, \mathbf{V}^o$ are the set of body and object vertices, respectively, $\mathbb{I}(\mathbf{v}^b \text{ in } \mathbf{V}^o)$ is a binary indicator of whether a body vertex, $\mathbf{v}^b$, is inside the object volume $\mathbf{V}^o$, computed via ray-parity testing [34]. Lower is better (↓).

S.2 Qualitative Evaluation (Extending Sec. 4.3 of Main)

S.2.1 Additional Results

Comparisons to SotA: The new extensive comparisons in Figs. S.2 and S.3 (which extend Fig. 6 of the main) demonstrate the robustness of LEXIS-Flow* across diverse, complex interactions shown in images taken in the wild. For more results shown with a rotating viewpoint, see our website video.

Failure Cases: As with all methods, performance can degrade when poor off-the-shelf initialization exceeds the recovery capacity of the Flow, or for highly atypical interactions underrepresented in existing datasets [41, 81, 91, 93]. In such Out-Of-Distribution (OOD) cases, the LEXIS manifold (and the decoded InterFields) may misguide LEXIS-Flow, causing floating or penetrating artifacts.

S.2.2 Perceptual Study

To evaluate perceived realism, we conduct a perceptual study. We randomly sample 60 images from Open3DHOI [78] and reconstruct each using our LEXIS-Flow*

Fig. S.1: Failure cases. Errors arise from inaccurate body pose and depth estimation. This can lead to misplaced human–object configurations, e.g., hovering over a bench or skateboard instead of contacting it.

Fig. S.2: LEXIS-Flow vs SotA methods. Across many in-the-wild images [78] and for a wide variety of objects, our LEXIS-Flow method estimates more physically-plausible 3D HOI estimates than existing SotA approaches, capturing better cphysical coupling while reducing floating artifacts and interpenetrations. For more results shown with a rotating viewpoint, see our website video. Here: 🔍 **Zoom in** to see 3D details.

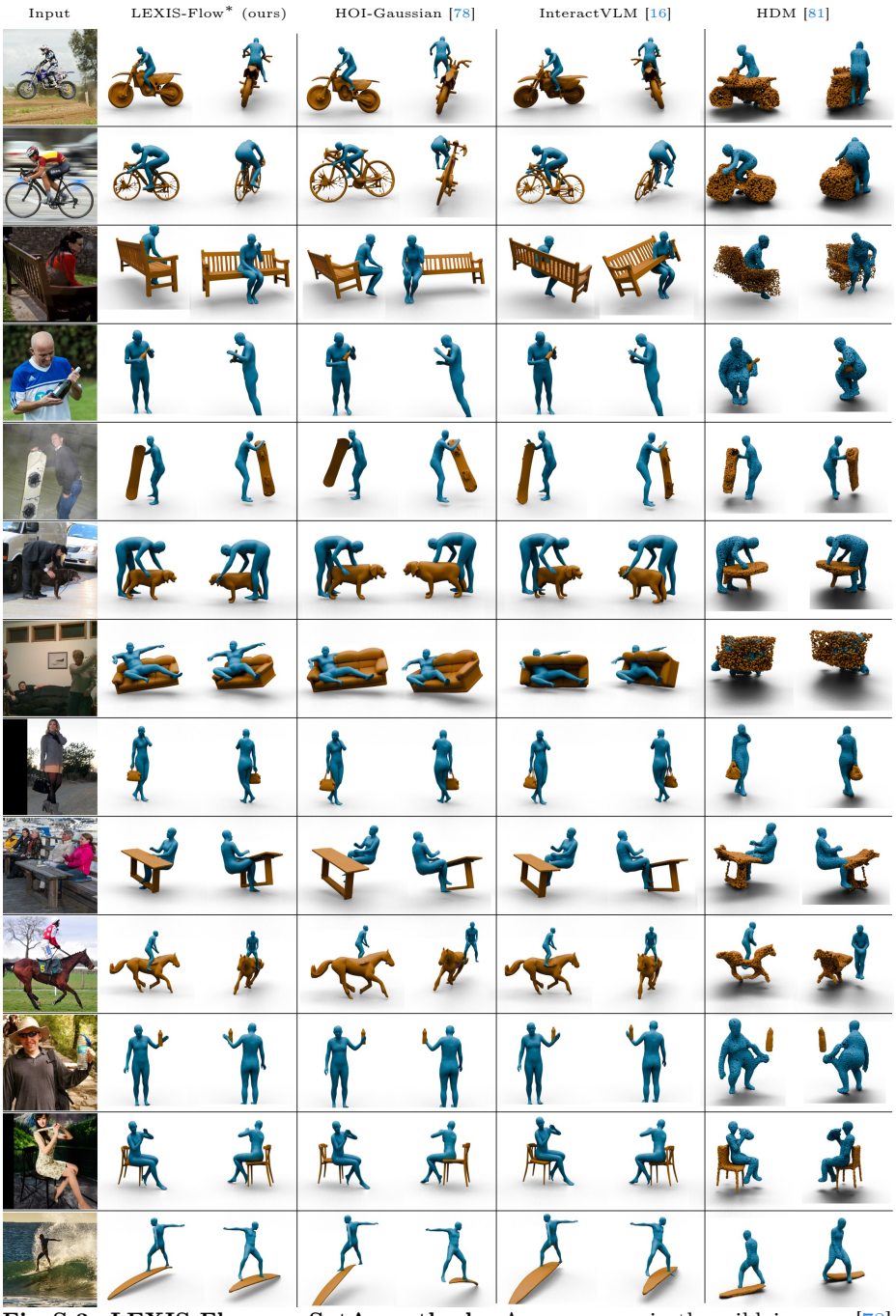

Fig. S.3: LEXIS-Flow vs SotA methods. Across many in-the-wild images [78] and for a wide variety of objects, our LEXIS-Flow method estimates more physically-plausible 3D HOI estimates than existing SotA approaches, capturing better cphysical coupling while reducing floating artifacts and interpenetrations. For more results shown with a rotating viewpoint, see our website video. Here: 🔍 **Zoom in** to see 3D details.

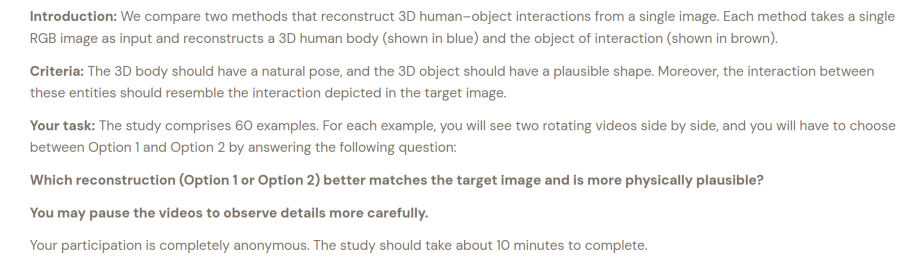

Fig. S.4: Perceptual study protocol – Instructions to participants. We ask 62 participants to view 60 images and respective 3D reconstructions by our LEXIS-Flow* method and the SotA HOI-Gaussian [78] method. Then, for each image they select the reconstruction that better matches the image and appears more physically plausible.

and the SotA HOI-Gaussian [78] method. We randomize both the presentation order and left/right placement. Each image is shown to 62 participants, who select the reconstruction that better matches the input image and appears more physically plausible. We present 60 samples, of which 4 are catch trials to assess participant reliability (e.g., to detect participants that do not understand the task); this filters out 1 participant, leaving out a total of 61 valid participants. LEXIS-Flow* is preferred in 75.8% of comparisons over HOI-Gaussian. For participant instructions, see Fig. S.4.