LEXIS: LATENT PROXIMAL INTERACTION SIGNATURES FOR 3D HOI FROM AN IMAGE

Anonymous authors
Paper under double-blind review



Figure 1: We present LEXIS-Flow, a flow matching 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.

ABSTRACT

Reconstructing 3D Human–Object Interaction from an RGB image is essential for perceptive systems and it requires capturing the subtle physical coupling between the body and objects. Current methods rely on *sparse, binary contact* cues. These miss the continuous proximity and dense spatial relationships inherent in interactions. We address this with InterFields, which encode *dense, continuous proximity* and the *direction* to the interacting geometry across the entire body and object surfaces. Inferring such fields from a single image is ill-posed. Our intuition is that interaction patterns are structured by the action and object geometry. We capture this structure in LEXIS, a novel discrete manifold of interaction signatures learned via a quantized autoencoder. We then develop LEXIS-Flow, a flow-matching framework that uses LEXIS manifold to estimate human and object meshes alongside their InterFields. These InterFields guide sampling that improves physical plausibility without post-hoc optimization. On Open3DHOI and BEHAVE, LEXIS-Flow outperforms SotA baselines in reconstruction, contact and physical plausibility. It generalizes better and the reconstructions are perceived as more realistic, moving us closer to holistic 3D scene understanding.

1 INTRODUCTION

Our actions are inherently defined and physically constrained by the 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 relationship*.

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

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Both components remain open problems, because we lack a representation that bridges *sparse* 2D image cues and *dense* 3D physical interaction. Past work exploits *binary contact* (Dwivedi et al., 2025; Taheri et al., 2020; Shimada et al., 2022), i.e. classifying body and object surface points as in contact or not, to align body–object geometry, resolving, to some extent, the occlusion and depth ambiguities.

However, contact has fundamental limitations: it is only a *sparse* signal, and only *binary*, encoding small contacting areas while ignoring the rest. Think of working out by lying on a bench and pressing dumbbells, as shown in Fig. 2; contact between the body and objects stays fixed across time, failing to encode the interaction's progression.

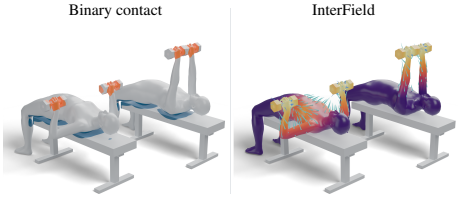

Figure 2: **Toy example.** Fixed contact; but, InterField follows the interaction.

To tackle the above limitations, we need a richer representation. Our key observation is that *distances* and *directions* between body and object change as the interaction progresses. This yields a *dense*, *continuous* signal of surface-to-surface proximity and of where the counterpart surface lies, namely a 3D *Interaction Field*, or InterField for short (see Fig. 1). Contact is only its zero level set, so the InterField does not "replace" contact but "*extends*" it by capturing how the human and object surfaces are spatially related, which helps 3D reconstruction. Yet InterFields have not been explored for full-body interactions or explicitly used at inference time to improve reconstruction.

One reason this remains unexplored is that inferring 3D fields, i.e., distances and directions for all points across two 3D surfaces from just a 2D image, is inherently ill-posed, and the space of possible interactions is vast and highly non-convex. We thus learn a novel *manifold* of the interaction patterns captured by InterFields, called LEXIS (LatEnt proXimal Interaction Signatures). It is a discrete *manifold* encoding *prior knowledge* about interaction signatures. Specifically, to learn LEXIS we train a FSQ autoencoder (Mentzer et al., 2024) that encodes 3D human and object geometry in interaction into latent tokens, and decodes this into body articulation, and body and object InterFields. As training data we use MoCap-based, synthetically-augmented 3D HOI pairs.

We exploit LEXIS to estimate 3D interactions from a single RGB image. Existing methods mostly work in two stages: coarse human and object shape and pose estimation, followed by a post-hoc optimization with contact-based losses. This is suboptimal, because contact provides only *sparse, binary* cues and these are considered only "*too late*," i.e., post-hoc. Thus, we introduce LEXIS-Flow, a Flow-Matching model that exploits *dense, continuous* InterFields "*early on*" to jointly estimate posed 3D human and object meshes along with their InterFields. Crucially, the predicted InterFields steer the sampling, replacing the need for post-hoc optimization.

We evaluate primarily on the Open3DHOI (Wen et al., 2025) dataset due to its focus on *in-the-wild* images and *generalizability*. We additionally measure *physical plausibility* through a proxy stability metric that complements standard geometric metrics. Dense, continuous InterFields prove more helpful than sparse, binary contacts for formulating 3D reconstruction constraints, both in LEXIS-Flow's inference and in a fitting framework (Dwivedi et al., 2025). Benchmarking shows that LEXIS-Flow clearly outperforms SotA methods on 3D reconstruction in the wild and that its estimates are perceived as significantly more realistic, while in-distribution evaluation on the BEHAVE (Bhatnagar et al., 2022) dataset echoes the above. Ablations support key design choices, and show that InterFields at the core yield physically more plausible reconstructions. Moreover, LEXIS-Flow can initialize from 3D estimates of SotA tools, which its guided refinement improves.

In summary, here we make the following contributions:

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

Code and models will be released.

2 Related Work

We organize prior work by how it represents interaction, from binary contact to spatial probability and fields, and by whether it recovers body and object separately or jointly (more in Section S.1).

Binary Contact: Contact is a binary cue (points are in contact or not) that is used to taxonomize grasps (Kamakura et al., 1980; Feix et al., 2016; Brahmbhatt et al., 2020), to synthesize hand and body interaction (Taheri et al., 2020; Jiang et al., 2021; Liu et al., 2023c; Hassan et al., 2021; Diller &Dai, 2024), and to reconstruct interaction from pixels (Zhang et al., 2020; Grady et al., 2021; Xie et al., 2022). For the latter, contact is detected in pixels (Chen et al., 2023), thresholded from distances (Huang et al., 2022), annotated by hand (Cseke et al., 2025), inferred in 3D (Tripathi et al., 2023a; Yang et al., 2024; Dwivedi et al., 2025), or measured as pressure (Tripathi et al., 2023b).

Spatial Probability: CHORUS (Han &Joo, 2023) learns a 3D object occupancy field w.r.t. the body from text-to-image diffusion samples, and ComA (Kim et al., 2024) extends this to an object-to-human distance field with orientation cues. Both learn separate models per object or action class and need heavy sampling and filtering.

Interaction Fields: Beyond binary contact, and spatial probability, ARCTIC (Fan et al., 2023) introduces bidirectional hand-object distances, and CHORE (Xie et al., 2022) thresholds learned distance fields (DF) into binary contacts. DFs also drive interaction synthesis (Taheri et al., 2022; Zhang et al., 2021; Diller &Dai, 2024; Taheri et al., 2024; Xue et al., 2025) and hand-grasp reconstruction (Yang et al., 2021; Qi et al., 2024), but only as training supervision (ARCTIC), in post-hoc optimization (CPF), or as the input of a separate refinement network without a feedback loop (Zhang et al., 2025; Cha et al., 2024). LEXIS-Flow differs in three ways. **Usage:** it predicts full-body InterFields and uses them explicitly *during* inference, to guide its own sampling. **Task:** it does so for full-body HOI reconstruction from an image. **Representation:** besides the distance, our InterFields carry the unit *direction* to the nearest counterpart point, which, to our knowledge, no prior work predicts, and they come from a learned *manifold* of interaction signatures.

Independent Body and Object Estimation: In previous work, the 3D body is recovered by fitting a parametric body model to image cues (Bogo et al., 2016), or by regressing its parameters directly from pixels (Kanazawa et al., 2018; Patel &Black, 2025). For objects, pose is estimated by regression (Gkioxari et al., 2019) or template fitting (Antić et al., 2025), while shape is encoded as voxels (Choy et al., 2016), point clouds (Fan et al., 2017), superquadrics (Paschalidou et al., 2019), meshes (Wang et al., 2018) or neural fields (Park et al., 2019), and is increasingly generated by diffusion (Liu et al., 2023b) or retrieved from a database (Liu et al., 2023a). Recently SAM3D (SAM 3D Team et al., 2026) estimates robust object shape but occlusion- and interaction-unaware pose; we *initialize* LEXIS-Flow with its object shape estimates and *refine the pose* with interaction awareness.

Single-Image HOI Reconstruction: Template-free, HDM (Xie et al., 2024) infers human and object point clouds with a hierarchical diffusion model that predicts a rough joint cloud and then refines each separately. Template-based methods take the object shape as input: PHOSA (Zhang et al., 2020), CHORE (Xie et al., 2022) and HOI-Gaussian (Wen et al., 2025) fit body and object poses to the image under contact constraints, taking contact from annotations, thresholded distance fields, or splat opacities, respectively. CONTHO (Nam et al., 2024) and HOI-TG (Wang et al., 2025b) regress coarse body parameters and object orientation, and refine these with a learned network conditioned on binary contacts. Trained on scarce in-lab captures with pre-scanned objects (Bhatnagar et al., 2022; Huang et al., 2024), these methods see few objects and interactions and generalize poorly to in-the-wild images. Recent in-the-wild methods therefore resort to slow test-time optimization, fitting body and object to contact from a database (PICO) or a vision-language model (InteractVLM, Dwivedi et al., 2025), or to text and appearance alignment (TeHOR, Nam et al., 2026), with object shapes retrieved from Objaverse (Deitke et al., 2023; Liu et al., 2023a) or from an image-to-3D model (Xu et al., 2024).

All of these are *two-stage*: they estimate rough meshes, then refine them with *sparse, binary* contact. Instead, ours is *one-stage*, with interaction at the core. LEXIS-Flow *jointly* infers body, object pose and a LEXIS signature that decodes to *dense, continuous* InterFields. These guide sampling and improve the physical plausibility of estimates.

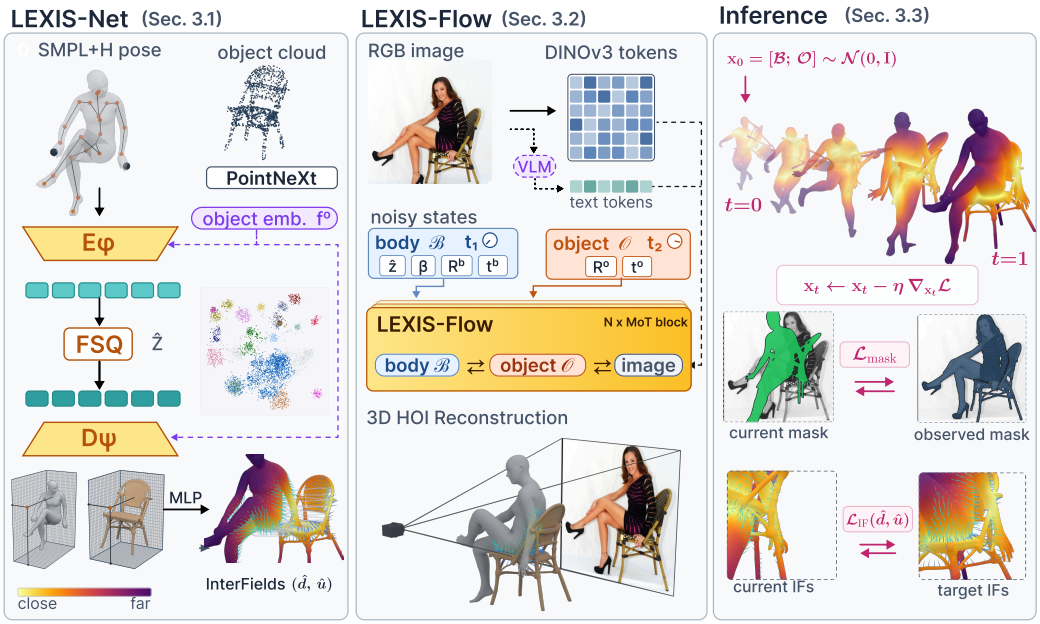

Figure 3: **Method overview.** LEXIS-Net (left, Section 3.1) tokenizes body articulation and object shape into the LEXIS manifold and decodes the tokens into body pose, InterField distance $\hat{d}$ (color) and direction $\hat{u}$ (quills). LEXIS-Flow (center, Section 3.2) denoises body, object and LEXIS tokens jointly from the image. Inference (right, Section 3.3) integrates from noise or from an off-the-shelf estimate (LEXIS-Flow*), steered by the mask and InterField gradients.

3 Methodology

Overview: Given an RGB image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ of a human–object interaction (HOI) and an off-the-shelf 3D object shape, we estimate the 3D human mesh (SMPL+H body (Romero et al., 2017)) and the object mesh, posed so that they interact realistically and match image cues. To this end, we estimate two states: **object state** $\mathcal{O} = \{R^o, t^o\}$ in a human-root-relative frame, with the global rotation $R^o \in \mathbb{R}^6$ (6D form (Zhou et al., 2019)) and translation $t^o \in \mathbb{R}^3$; **body state** $\mathcal{B} = \{R^b, t^b, \boldsymbol{\beta}, \mathcal{Z}\}$ in the camera frame, with rotation, translation, SMPL+H shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and the LEXIS tokens $\mathcal{Z} \in \mathbb{R}^{23\times D}$ (Section 3.1), encoding body pose and interaction.

Representation: Recent work uses InterFields (Fan et al., 2023), $\mathrm{IF}(\mathbf{p}) \in \mathbb{R}_{\geq 0}$, which encodes, for every body/object point $\mathbf{p}$, the distance to the nearest point on the interacting counterpart. We extend this to a distance *and* a direction: alongside $d(\mathbf{p}) \in \mathbb{R}_{\geq 0}$ we predict the unit vector $u(\mathbf{p}) \in \mathbb{S}^2$ pointing from $\mathbf{p}$ toward that nearest point, so $\mathrm{IF}(\mathbf{p}) = \big(d(\mathbf{p}),\, u(\mathbf{p})\big)$. Distance says how far the counterpart is, direction *where*; both are *dense* and *continuous*, but ill-posed to infer from an image.

Approach: We address this in two steps (Fig. 3): LEXIS-Net learns **LEXIS**, a manifold of *compact interaction signatures* (Section 3.1), and LEXIS-Flow estimates the LEXIS code and the object placement from the image, guided in its own sampling by the decoded InterFields (Sections 3.2 and 3.3). Unlike methods that apply contact constraints after inference, the InterFields constrain the sampling itself via gradient-based guidance in a single *one-stage* network.

3.1 LEXIS: Latent Proximal Interaction Signatures

The distances and directions of InterFields follow recurring patterns that are specific to the action and the object. LEXIS-Net, an FSQ autoencoder (Mentzer et al., 2024) (Fig. 3, left), captures these patterns in a structured latent space, which is far easier to infer from an image than the full fields. It encodes the body pose and object's canonical shape (a point cloud $\mathbf{P}^o \in \mathbb{R}^{N\times 3}$) into one discrete token per body joint and per hand. From these tokens the decoder reconstructs body pose and the InterFields of both body and object. The encoder sees the object shape embedding because the

interaction signature depends on the object, whereas the decoder is given that condition so it can lay the fields out on that surface. The LEXIS manifold therefore capture the semantics of the interaction without storing the fine geometry, and the decoder can transfer them to unseen shapes, such as the estimated meshes at test time (e.g. chair with a different shape). Both fields (human-to-object and object-to-human) live in respective canonical frames. The learned LEXIS space is organized by action (see Section S.3) and smooth, as interpolating between two signatures (Figs. S.2 and 4) decodes to intermediate interactions whose body pose, proximity and directions morph gradually.

Object shape enters both networks as $\mathbf{f}^o$, 128 control points derived from $\mathbf{P}^o$ by a frozen PointNeXt (Qian et al., 2022). In detail, an encoder $\mathcal{E}_\phi$ (with network weights ϕ) maps body pose $\boldsymbol{\theta}$ (axis-angle joint rotations) and $\mathbf{f}^o$ to a token sequence $\mathcal{Z} \in \mathbb{R}^{23\times D}$ by *finite scalar quantization*. Each latent dimension is bounded and then rounded onto a small integer grid:

$$\mathcal{Z} = \mathrm{round}\big(\,\mathrm{bound}(\hat{\mathcal{Z}})\big), \qquad \hat{\mathcal{Z}} = \mathcal{E}_\phi(\boldsymbol{\theta}, \mathbf{f}^o), \tag{1}$$

where $\hat{\mathcal{Z}}$ are continuous latents, bound is a tanh squashing and round a straight-through rounding.

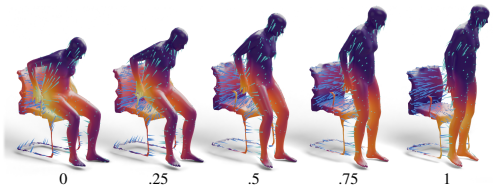

Figure 4: **LEXIS interpolation.** Two unseen BEHAVE interactions linearly interpolated and decoded to a plausible signature.

The decoder D_ψ maps the tokens back to body pose and full InterFields without committing to a mesh topology. It attends to the same object control points $\mathbf{f}^o$ and writes two TriPlane (Chan et al., 2022) feature volumes, $\mathcal{T}^b, \mathcal{T}^o = D_\psi(\mathcal{Z}, \mathbf{f}^o) \in \mathbb{R}^{3\times 24\times 24\times 64}$. Any surface point $\mathbf{p}$ is then projected onto the three planes, its sampled features are summed, and a small MLP reads $\hat{d}(\mathbf{p}), \hat{u}(\mathbf{p})$ from them (Fig. 3, left). In parallel, a light transformer head reads the joint rotations $\hat{\boldsymbol{\theta}}$ back from the tokens. More architecture details are in Section S.2.1.

The LEXIS-Net training objective is:

$$\mathcal{L}_{\text{LexisNet}} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{IF}} \tag{2}$$

$$\mathcal{L}_{\text{IF}} = \mathcal{L}_{\text{d}}^h + \mathcal{L}_{\text{d}}^o + \mathcal{L}_{\text{u}}^h + \mathcal{L}_{\text{u}}^o \tag{3}$$

where $\mathcal{L}_{\text{recon}}$ is a vertex loss between the body meshes posed by $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$, $\mathcal{L}_{\text{d}}$ an L_1 loss on decoded distances and $\mathcal{L}_{\text{u}}$ a cosine loss on directions, each on 1024 surface points of body and object separately. The steering weights λ are set empirically; see Section S.2.

3.2 LEXIS-Flow

With LEXIS-Net frozen, LEXIS-Flow (Fig. 3 middle) generates a LEXIS Code jointly with the body and object states, modeling $P(\mathcal{B}, \mathcal{O} \mid \mathcal{I})$ via Flow Matching (Lipman et al., 2023). With $\mathbf{x} = [\mathcal{B}; \mathcal{O}]$, the flow interpolates between noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at t=0 and a ground-truth sample $\mathbf{x}_1$ at t=1, $\mathbf{x}_t = (1-t)\,\boldsymbol{\epsilon} + t\,\mathbf{x}_1$, and a network $v_\gamma(\mathbf{x}_t, t, \mathcal{I})$ regresses the velocity $\mathbf{x}_1 - \boldsymbol{\epsilon}$. The vector field, v_γ, is parameterized by a multi-stream transformer backbone inspired by DiT (Peebles &Xie, 2023) and Mixture of Transformers (SAM 3D Team et al., 2026). The body and object streams are coupled by cross-attention and run on decoupled noise schedules with independent timesteps t_1, t_2 (Petrov et al., 2025), sampled independently in training. The network thus models the joint $P(\mathcal{B}, \mathcal{O} \mid \mathcal{I})$ as well as both conditionals: each stream must stay informative at any noise level of the other, which prevents one stream from being ignored (Bao et al., 2023). Optionally, LEXIS-Flow takes a coarse text description of the interaction (object, stance, verb, contact height, proximity, contacting parts), built from a fixed template that a vision-language model fills from the image at test time (Section S.2.3). Its tokens, from the frozen text tower of our image backbone, join the image tokens in the cross-attention context. The LEXIS-Flow training objective is:

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

$\mathcal{L}_{\text{FM}}$ is the conditional flow-matching objective (Section S.2); the other terms add geometric and visual supervision: $\mathcal{L}_{\text{IF}}$ is the L_1 discrepancy between decoded $(D_\psi(\hat{\mathcal{Z}}, \mathbf{f}^o))$ and GT InterFields, $\mathcal{L}_{\text{v2v}}$ an L_2 vertex loss on the body mesh, $\mathcal{L}_{\text{2D}}$ a 2D keypoint re-projection loss, and $\mathcal{L}_{\text{obs}}$ compares GT InterFields with those "observed" on the body and object posed by the estimated $\{\hat{R}^o, \hat{t}^o\}$, tying the estimated poses to the fields. Weights λ are in Section S.2.2.

3.3 INTERFIELD-GUIDED REFINEMENT WITHIN LEXIS-FLOW

LEXIS-Flow (Section 3.2) samples from a learned prior, and a prior alone does not guarantee that the sampled body and object agree with each other and with the image. Prior work restores this agreement with post-hoc contact optimization (Xie et al., 2022), which is slow, prone to local minima, and runs 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, each ODE step is followed by steering $\mathbf{x}_t$ along the negative gradient of a guidance loss (Eq. (6)) before integration continues. To this end, we add a guidance loss with the terms discussed in the following.

InterField Guidance ($\mathcal{L}_{\text{IF}}$): 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 and directional relationships between interacting surfaces. The InterField term is the field loss $\mathcal{L}_{\text{IF}}(\hat{\text{IF}}, \text{IF}_t)$ (Eq. (3)) between this decoded field and the "observed" field IF_t, computed from the body and the object posed with the running estimates $\{R_t^b, t_t^b\}, \{R_t^o, t_t^o\}$. The gradient thus steers the body and object toward the interaction that the LEXIS tokens encode, as analysis-by-synthesis inside the sampler. The two fields agree only when pose, placement and tokens describe the same interaction, so the gradient favors configurations the learned prior deems plausible over arbitrary contacts.

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 (Kirillov et al., 2023) model estimates robust 2D masks, and SAM3D (SAM 3D Team et al., 2026) provides good 3D shape estimates, with noisy pose estimates. Thus, we add a mask-based loss for the object.

Object vertices $\mathbf{V}^o$, transformed by running object state estimate (R^o, t^o), are projected onto the image plane using a differentiable renderer (Ravi et al., 2020). Vertices falling outside the observed 2D object mask (Kirillov et al., 2023), M^o, are penalized by:

$$\mathcal{L}_{\text{mask}} = \sum\nolimits_{\mathbf{v} \in \mathbf{V}^o} \| 1 - M^o\big(\pi(R_t^o \mathbf{v} + t_t^o)\big) \|^2. \tag{5}$$

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 single gradient step that combines the InterField and mask guidance terms:

$$\mathbf{x}_t \leftarrow \mathbf{x}_t - \eta \, \nabla_{\mathbf{x}_t} \big[\mathcal{L}_{\text{IF}}(\hat{\text{IF}}, \text{IF}_t) + \lambda_{\text{mask}} \, \mathcal{L}_{\text{mask}} \big], \tag{6}$$

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. The gradient reaches every component of $\mathbf{x}_t$, so tokens, body and object move jointly rather than the object alone.

This formulation can also refine off-the-shelf body- and object-only estimates that lack interaction awareness. We encode them into a latent state $\mathbf{x}_{\text{init}}$, transport it to an intermediate timestep $t_{\text{start}} \in (0, 1)$ by mixing in noise, $\mathbf{x}_{t_{\text{start}}} = (1 - t_{\text{start}})\, \boldsymbol{\epsilon} + t_{\text{start}}\, \mathbf{x}_{\text{init}}$, as in SDEdit (Meng et al., 2022), and run the guided integration from t_{start} to $t{=}1$. In Section 4 we denote this variant LEXIS-Flow*.

3.4 IMPLEMENTATION DETAILS

LEXIS-Net is a per-joint transformer encoder with FSQ tokens, decoded through object-conditioned TriPlanes; it trains in $\sim$5 hours on one A100 (64 GB). LEXIS-Flow is a 16-layer multi-stream transformer over frozen DINOv3 features (Siméoni et al., 2026), trained for 100k steps on four A100 (64 GB) GPUs ($\sim$48 hours). Sampling uses a 25-step Euler solver with guidance from $t_{\text{start}}{=}0.6$ (step 15 of 25); training drops the image condition with probability 0.3 and sampling uses classifier-free guidance (scale 1.5). Loss weights are $\lambda_{\text{IF}}{=}10$, $\lambda_{\text{v2v}}{=}5$, $\lambda_{\text{2D}}{=}1$, $\lambda_{\text{obs}}{=}0.5$ for LEXIS-Flow and 10 (distance) and 1 (direction) for LEXIS-Net; guidance uses $\eta{=}0.25$ and $\lambda_{\text{mask}}{=}0.5$; both networks train with AdamW, LEXIS-Flow with a 5k-iteration warmup (details in Section S.2).

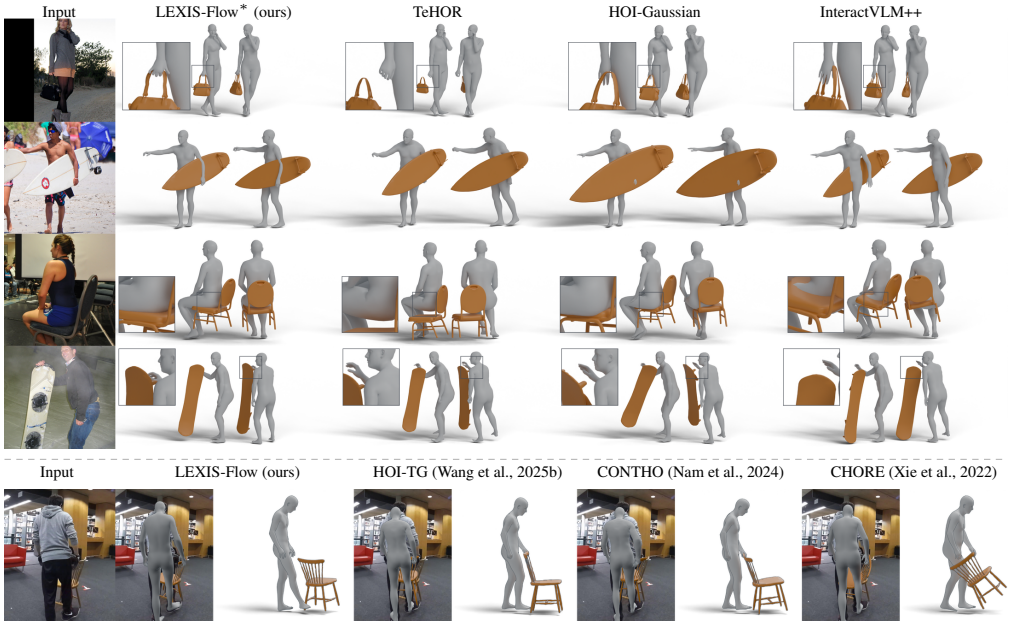

Figure 5: **LEXIS-Flow vs SotA.** Baselines estimate the object floating or penetrating the body; LEXIS-Flow recovers tight coupling through dense InterFields. Rows 1–4 show in-the-wild Open3DHOI; the last row shows in-lab BEHAVE (Section 4.3) against the baselines of Table 3.

4 Experiments

Data and Benchmarks: LEXIS-Net trains on geometry-only HOI data ProciGen (Xie et al., 2024) and the InterAct-refined (Xu et al., 2025) NeuralDome (Zhang et al., 2023), OMOMO (Li et al., 2023) and IMHD (Zhao et al., 2024). LEXIS-Flow is trained on the 3D-paired ProciGen images. We evaluate in-the-wild reconstruction performance on Open3DHOI (Wen et al., 2025) and on in the lab BEHAVE (Bhatnagar et al., 2022) dataset on its official split (Xie et al., 2022).

LEXIS-Flow samples the HOI state from noise given the image and a SAM3D (SAM 3D Team et al., 2026) shape; like HDM it trains on ProciGen. LEXIS-Flow* noises off-the-shelf estimates (CameraHMR (Patel &Black, 2025) body, SAM3D object, aligned by MoGe (Wang et al., 2025a) metric depth, Section S.2.4) to an intermediate timestep (Meng et al., 2022), then integrates with guidance (Section 3.3). InteractVLM++ (Dwivedi et al., 2025), HOI-Gaussian (Wen et al., 2025) and TeHOR (Nam et al., 2026) get the same initialization (details in Section S.2.6). Text descriptions for all methods are produced by the same VLM, Claude Opus 5 (Anthropic, 2026a), which fills respective templates from the image. See Section S.2.3 for sensitivity study under descriptions from other VLMs.

Metrics: Following Nam et al. (2024), we Procrustes-align the body, apply its transform to the object and evaluate Chamfer Distances in cm (CD_{hum}, CD_{obj}), Collision (%) and Contact F1 (@5 cm). These measure accuracy but not physical plausibility, so on BEHAVE we measure stability as its proxy. This measures how often the contacts hold the object under gravity. InterFields accuracy is scored by the L_2 error in cm and the L_1 angle error in degrees. All metrics are defined in Section S.2.5.

4.1 Interaction Representation

We compare dense, continuous InterFields with sparse, binary contacts, obtained by thresholding them, on two tasks (Table 1): fitting by render-and-compare, as in InteractVLM (Dwivedi et al., 2025), and guided generative sampling with LEXIS-Flow. InterFields win on both, improving CD_{obj} and Contact for both fitting and generative sampling. This shows that our InterFields are the richer spatial signal.

Task	Representation	CD_{obj} ↓	Contact ↑
Fitting	Binary Contact	35.4	0.178
	InterField (ours)	**31.2**$^{-11.9\%}$	**0.213**$^{+19.7\%}$
Generative	Binary Contact	49.11	0.152
	InterField (ours)	**32.01**$^{-34.8\%}$	**0.281**$^{+84.9\%}$

Table 1: **Interaction representation.** We compare binary contacts and InterFields on Open3DHOI on: render-and-compare fitting and generative reconstruction.

Method	Text	Open3DHOI			
		CD_{hum} ↓	CD_{obj} ↓	Collision (%) ↓	Contact ↑
From noise					
A. HDM (w. scale align.) (Xie et al., 2024)		13.50	49.38	8.9	0.141
B. LEXIS-Flow (Ours)		**7.75**$^{-42.6\%}$	**32.01**$^{-35.2\%}$	**6.0**$^{-32.6\%}$	**0.281**$^{+99.3\%}$
Refinement					
C. CameraHMR + SAM3D		7.20	37.30	5.1	0.182
D. HOI-Gaussian (Wen et al., 2025)		7.28	32.02	6.1	0.151
E. InteractVLM (Dwivedi et al., 2025)	✓	7.20	38.20	5.4	0.372
F. TeHOR (Nam et al., 2026)	✓	7.05	31.15	4.8	0.411
G. InteractVLM++ (Dwivedi et al., 2025)	✓	7.20	30.11	4.7	0.394
H. LEXIS-Flow* (Ours)		6.79$^{-5.7\%}$	21.26$^{-33.6\%}$	4.1$^{-19.6\%}$	0.451$^{+148\%}$
I. LEXIS-Flow* (Ours)	✓	**6.30**$^{-10.6\%}$	**20.06**$^{-33.4\%}$	**3.3**$^{-29.8\%}$	**0.467**$^{+13.6\%}$

Table 2: **3D HOI in the wild (Section 4.2).** On Open3DHOI (Wen et al., 2025). Rows A–B sample from noise; rows C–I refine the same off-the-shelf initialization C (CameraHMR (Patel &Black, 2025) body, SAM3D (SAM 3D Team et al., 2026) object, MoGe depth), row E its own CameraHMR initialization. ✓: the method receives a text description of the interaction (Section 3.2). LEXIS-Flow* denotes refinement. LEXIS-Flow improves over HDM when sampling from noise, while LEXIS-Flow* performs best across all metrics in the refinement setting. The gains are strongest on interaction-sensitive metrics, highlighting the benefits of dense InterField-guided refinement. **Bold**: best per block; underlined: best competitor.

4.2 In-the-wild 3D HOI Reconstruction

Generative Reconstruction (Table 2): LEXIS-Flow (row B) outperforms HDM (Xie et al., 2024) (row A) on CD_{hum} (7.8 vs 13.5 cm), CD_{obj} (32.0 vs 49.4 cm) and Contact F1 (0.28 vs 0.14); note that HDM produces geometric artifacts and intersecting meshes on real-world images (Fig. S.4). Furthermore, LEXIS-Flow already achieves lower CD_{obj} and higher Contact F1 than combining the off-the-shelf experts CameraHMR + SAM3D (row C). Therefore, jointly sampling body and object through InterFields beats SotA per-entity estimation.

Guided Refinement (Table 2): LEXIS-Flow* achieves the best results on all four metrics, both with and without text. Without text (row H), it outperforms HOI-Gaussian (row D); with a text description (row I), it outperforms TeHOR (F), InteractVLM++ (G), and InteractVLM (E). The largest gains are on interaction-sensitive metrics: against the strongest baselines, CD_{obj} drops from 30.11 to 20.06 cm (-33.4%), while Contact F1 rises from 0.411 to 0.467 ($+13.6\%$). CD_{hum} also improves from 7.05 to 6.30 cm and Collision from 4.7% to 3.3%.

These gains come from refinement rather than a better initialization. All refinement methods start from the same estimates, and the advantage persists even on unseen object classes (Table S.2). Under object-rotation noise of up to $45°$, CD_{obj} changes by less than 0.5 cm for LEXIS-Flow*, compared with up to 5.6 cm for HOI-Gaussian (Table S.3).

Unlike prior methods that optimize final estimates against sparse contacts, LEXIS-Flow* guides the denoising states using dense InterFields from the learned manifold. This refinement takes 0.7 s after the shared $\sim$10 s initialization, compared with more than 360 s for TeHOR.

Qualitative comparison: Figure 5 compares our LEXIS-Flow* method against SotA methods in the wild; for more results see Section S.4.1. LEXIS-Flow* recovers physically-plausible interactions via InterField-guided refinement. Baselines often leave the object floating or penetrating the body, since binary contacts only say which points should touch, and such a sparse signal gives no gradient once the object is a few centimeters off. Conversely, our InterFields say how far and in which direction *every* point is off, so our guidance corrects both.

Perceptual study: We conduct a perceptual study on 60 Open3DHOI images with 62 participants (protocol in Section S.4.2). This shows that LEXIS-Flow* estimations are perceived as more realistic 75.8% of times over HOI-Gaussian. Participants consistently prefer our reconstructions, suggesting that the geometric improvements also translate to more realistic 3D interactions.

	PHOSA (Zhang et al., 2020)	HDM (Xie et al., 2024)	CHORE (Xie et al., 2022)	CONTHO (Nam et al., 2024)	HOI-TG (Wang et al., 2025b)	LEXIS-Flow (ours) no text	LEXIS-Flow (ours) w/ text
CD_{hum} ↓	12.17	11.61	5.65	4.798	<u>4.412</u>	$4.05^{-8.2\%}$	**3.95**$^{-10.5\%}$
CD_{obj} ↓	26.62	11.35	10.64	8.343	<u>7.980</u>	$7.12^{-10.8\%}$	**6.01**$^{-24.7\%}$
Stability (%) ↑	N/A	N/A	72.1%	70.2%	<u>77.8%</u>	$82.2^{+5.7\%}$	**88.5**$^{+13.8\%}$

Table 3: **In-lab 3D HOI:** On BEHAVE (Bhatnagar et al., 2022) against methods trained on it. **Bold**: best; <u>underlined</u>: best competitor.

	distance (cm) ↓		direction (°) ↓	
	h2o	o2h	h2o	o2h
Gauss. VAE	8.71	10.15	25	45
VQ-VAE	<u>4.73</u>	<u>6.12</u>	<u>17</u>	<u>30</u>
FSQ (ours)	**2.57**$^{-45.7\%}$	**4.23**$^{-30.9\%}$	**12.5**$^{-26.5\%}$	**21**$^{-30.0\%}$

Table 4: **Latent space ablation.** LEXIS-Net on BEHAVE. Decoded-InterField distance (cm) and direction (°) errors, human-to-object and object-to-human. **Bold**: best; <u>underlined</u>: second best.

	CD_{hum} ↓	CD_{obj} ↓	Stability↑
none	5.21	9.13	71.2
$\mathcal{L}_{mask}$	5.05	8.85	73.3
$\mathcal{L}_{IF}$ dist.	4.62	8.36	76.3
$\mathcal{L}_{IF}$ dist.+dir.	<u>4.30</u>	<u>8.20</u>	<u>77.5</u>
both (ours)	**3.95**$^{-8.1\%}$	**6.01**$^{-26.7\%}$	**88.5**$^{+14.2\%}$

Table 5: **Guidance ablation.** LEXIS-Flow sampling on BEHAVE with text descriptions, with each guidance term and both; Stability in %. **Bold**: best; <u>underlined</u>: second best.

4.3 In-lab 3D HOI Reconstruction

On BEHAVE official test split (Table 3) we compare against the optimization-based PHOSA (Zhang et al., 2020) and CHORE (Xie et al., 2022) and the learning-based CONTHO (Nam et al., 2024), HOI-TG (Wang et al., 2025b) and HDM (Xie et al., 2024) SotA methods. LEXIS-Flow outperforms state-of-the-art method, HOI-TG (Wang et al., 2025b), by 10.5% on CD_{hum} (3.95 vs 4.41 cm) and 24.7% on CD_{obj} (6.01 vs 7.98 cm) when given a text description, and by 8.2% and 10.8% without it. The largest gain is in stability, the test of whether the reconstructed interaction actually holds, where LEXIS-Flow reaches 88.5% (82.2% without text) against 77.8% for HOI-TG. The ground-truth meshes themselves reach only 88.9%, since BEHAVE misses the receptacle mesh (e.g. table) in certain examples, thus an object placed on it has nothing to be supported by. This stability criterion is directly relevant to downstream applications such as robot imitation and simulation. The last row of Fig. 5 shows BEHAVE example; additional results appear in Figs. S.5 and S.6.

Ablations: We isolate our design choices on BEHAVE (Tables 4 and 5). In the ablation of the LEXIS structure, replacing the discrete latent with a continuous Gaussian latent increases human-to-object InterField error from 2.6 to 8.7 cm and object-to-human error from 4.2 to 10.2 cm; direction errors show the same trend. More importantly, ablating the guidance terms in LEXIS-Flow isolates their contribution to physical plausibility. Mask guidance alone has little effect on the unguided estimate, whereas InterField guidance improves all metrics, with the direction term providing further gains. Combining both yields the most accurate and stable reconstructions. Introducing directions increases stability from 71.2% to 88.5%, showing that distances, directions and 2D cues are complementary, but necessary for physical plausibility.

5 Conclusion

We go beyond sparse, binary contact by leveraging dense, continuous InterFields, which encode both a distance and a direction to the interacting surface, for reconstructing 3D Human-Object Interaction from single images. To make InterFields inference tractable from a single image, we learn LEXIS, a latent manifold of interaction signatures, encoding interaction proximity patterns. Then, we develop LEXIS-Flow, a multi-stream Flow-Matching model that jointly estimates 3D human and object meshes alongside their InterFields, and exploits these for guided refinement of reconstruction during inference, eliminating post-hoc optimization. Experiments show that continuous InterFields outperform binary contacts in both fitting and generative settings. On Open3DHOI and BEHAVE, LEXIS-Flow outperforms all baselines, and its InterField guidance makes reconstructions not only more accurate but also more physically plausible. These results establish dense interaction fields as an effective representation for interaction-aware 3D reconstruction from images. This moves us closer to holistic 3D scene understanding, necessary for real-world and physical AI applications.

AI USE STATEMENT

We used generative AI tools, including LLMs and coding assistants, to assist with parts of the training code, and edit figure-rendering scripts. We did not use them to generate data, develop the method, formulate claims or hypotheses, design experiments, or interpret results. All AI-assisted content was reviewed, tested, and verified by the authors, who take full responsibility for the final work.

REFERENCES

Anthropic. Claude Opus 5 system card, 2026a.

Anthropic. Claude Sonnet 5 system card, 2026b.

Dimitrije Antić, Georgios Paschalidis, Shashank Tripathi, Theo Gevers, Sai Kumar Dwivedi, and Dimitrios Tzionas. 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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. arXiv:2502.13923, 2025.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning (ICML)*, pp. 1692–1717, 2023.

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 *Computer Vision and Pattern Recognition (CVPR)*, pp. 15914–15925, 2022.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, volume 9909, pp. 561–578, 2016.

Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, volume 12358, pp. 361–378, 2020.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3D motion generation for hand-object interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1577–1585, 2024.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 16102–16112, 2022.

Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 17100–17110, 2023.

Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *European Conference on Computer Vision (ECCV)*, volume 9912, pp. 628–644, 2016.

Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1783–1794, 2025.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 13142–13153, 2023.

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3D human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 19888–19901, 2024.

Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. InteractVLM: 3D interaction reasoning from 2D foundational models. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 22605–22615, 2025.

Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3D object reconstruction from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2463–2471, 2017.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954, 2023.

Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The GRASP taxonomy of human grasp types. *Transactions on Human-Machine Systems (THMS)*, 46(1):66–77, 2016.

Carlo Ferrari and John F. Canny. Planning optimal grasps. In *International Conference on Robotics and Automation (ICRA)*, pp. 2290–2295, 1992.

Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. In *International Conference on Computer Vision (ICCV)*, pp. 9784–9794, 2019.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1471–1481, 2021.

Sookwan Han and Hanbyul Joo. CHORUS: Learning canonicalized 3D human-object spatial relations from unbounded synthesized images. In *International Conference on Computer Vision (ICCV)*, pp. 15789–15800, 2023.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 14708–14718, 2021.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *International Conference on Learning Representations (ICLR)*, 2024.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 13274–13285, 2022.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 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.

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, pp. 11087–11096, 2021.

Noriko Kamakura, Michiko Matsuo, Harumi Ishii, Fumiko Mitsuboshi, and Yoriko Miura. Patterns of static prehension in normal hands. *American Journal of Occupational Therapy*, 34(7):437–445, 1980.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 7122–7131, 2018.

Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3D objects from pre-trained 2D diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 400–419, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *Transactions on Graphics (TOG)*, 42(6):197:1–197:11, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 44860–44879, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *International Conference on Computer Vision (ICCV)*, pp. 9264–9275, 2023b.

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. ContactGen: Generative contact modeling for grasp generation. In *International Conference on Computer Vision (ICCV)*, pp. 20552–20563, 2023c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pp. 5441–5450, 2019.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations (ICLR)*, 2024.

Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC Press, 1994. ISBN 978-0-8493-7981-9.

Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3D human and object via contact-based refinement transformer. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 10218–10227, 2024.

Hyeongjin Nam, Daniel Sungho Jung, and Kyoung Mu Lee. TeHOR: Text-guided 3D human and object reconstruction with textures. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 7100–7110, 2026.

Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 165–174, 2019.

Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3D shape parsing beyond cuboids. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 10344–10353, 2019.

Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *International Conference on 3D Vision (3DV)*, pp. 1562–1571, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023.

Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. TriDi: Trilateral diffusion of 3D humans, objects, and interactions. In *International Conference on Computer Vision (ICCV)*, pp. 5523–5535, 2025.

Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. HOISDF: Constraining 3D hand-object pose estimation with global signed distance fields. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 10392–10402, 2024.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: Revisiting pointnet++ with improved training and scaling strategies. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. arXiv:2007.08501, 2020.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3Dfy anything in images. In *Computer Vision and Pattern Recognition (CVPR)*, 2026.

Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. HULC: 3D human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision (ECCV)*, volume 13682, pp. 516–533, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research (TMLR)*, 2026.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, volume 12349, pp. 581–600, 2020.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 13263–13273, 2022.

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

Tencent Hunyuan3D Team. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

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 *International Conference on Computer Vision (ICCV)*, pp. 7967–7979, 2023a.

Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 4713–4725, 2023b.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *European Conference on Computer Vision (ECCV)*, volume 11215, pp. 55–71, 2018.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 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 2025a.

Zhenrong Wang, Qi Zheng, Sihan Ma, Maosheng Ye, Yibing Zhan, and Dongjiang Li. End-to-end HOI reconstruction transformer with graph-based encoding. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 27706–27715, 2025b.

Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong, Yulong Chen, Lizhuang Ma, and Yong-Lu Li. Reconstructing in-the-wild open-vocabulary human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 17426–17436, 2025.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879, 2024.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D shapenets: A deep representation for volumetric shapes. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1912–1920, 2015.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, Human and Object REconstruction from a single RGB image. In *European Conference on Computer Vision (ECCV)*, pp. 125–145, 2022.

Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 10003–10015, 2024.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models. arXiv:2404.07191, 2024.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3D human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 7048–7060, 2025.

Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 22714–22723, 2025.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *International Conference on Computer Vision (ICCV)*, pp. 11097–11106, 2021.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. LEMON: Learning 3D human-object interaction relation from 2D images. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 16284–16295, 2024.

He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. *Transactions on Graphics (TOG)*, 40(4):121:1–121:14, 2021.

Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, pp. 34–51, 2020.

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 8834–8845, 2023.

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. BimArt: A unified approach for the synthesis of 3D bimanual interaction with articulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 27694–27705, 2025.

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'M HOI: Inertia-aware monocular capture of 3D human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 729–741, 2024.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 5745–5753, 2019.

Xinghao Zhu, Zixi Liu, Shalin Jain, Chenran Li, Milad Noori, Michael Andres Lin, Huihua Zhao, John Welsh, Mrinal Verghese, Wei Liu, et al. Learning dexterous manipulation using contact wrench guidance from human demonstration. arXiv:2607.00033, 2026.

APPENDIX

S.1 EXTENDED RELATED WORK

We expand on the works closest to ours, complementing Section 2.

Interaction fields: CHORUS (Han &Joo, 2023) learns a 3D occupancy field of the object around the body from text-to-image diffusion samples, and ComA (Kim et al., 2024) extends it to an object-to-human distance field with orientation cues. Both train one model per object or action class, need heavy sampling and filtering of the generated images, and use the field to synthesize interactions rather than to reconstruct them from a photo. ARCTIC (Fan et al., 2023) defines bidirectional hand-object distances but uses them only as training supervision, CHORE (Xie et al., 2022) thresholds learned distance fields into binary contacts, and CPF (Yang et al., 2021) and HOISDF (Qi et al., 2024) fit hand grasps to distance fields by post-hoc optimization. For synthesis, distance fields enter the features or losses of whole-body grasping (Taheri et al., 2022), hand manipulation (Zhang et al., 2021; Xue et al., 2025), body-object motion (Diller &Dai, 2024; Taheri et al., 2024) and text-driven generation (Cha et al., 2024; Zhang et al., 2025), where a separate network consumes them without a feedback loop. None predicts the direction to the counterpart surface, and none samples the field jointly with the pose it constrains.

Single-image HOI reconstruction: PHOSA (Zhang et al., 2020) fits body and object to the image under hand-written contact and interpenetration priors with known contact regions. CHORE (Xie et al., 2022) predicts distance fields and pose maps from the image and fits templates to them, and HOI-Gaussian (Wen et al., 2025) fits Gaussian splats whose opacities yield contacts. CONTHO (Nam et al., 2024) regresses the body mesh and object orientation directly, and HOI-TG (Wang et al., 2025b) regresses both with a graph transformer over contact. HDM (Xie et al., 2024) diffuses a joint human-object point cloud and then refines the two by cross-attentive diffusion, without a template. For in-the-wild images, PICO (Cseke et al., 2025) predicts noisy body-only contact and retrieves clean labels from its database, InteractVLM (Dwivedi et al., 2025) predicts contact through a vision-language model, and TeHOR (Nam et al., 2026) aligns text and appearance, and all three then optimize the object against the contact. These are two-stage methods with binary contact as the constraint, whereas LEXIS-Flow samples pose and dense InterFields jointly and lets the fields guide the sampling.

S.2 IMPLEMENTATION DETAILS

S.2.1 LEXIS-NET

Data and schedule: We train LEXIS-Net in two stages. For the first 20k iterations it sees only body poses from AMASS (Mahmood et al., 2019), without interaction supervision, so that the tokens first cover the space of body poses and the FSQ grid is populated before any interaction is seen. These samples have no object, so the cross-attention to $\mathbf{f}^o$ is masked out in both the encoder and the decoder, and only the pose reconstruction loss is active. We then mix in the interaction data, ProciGen (Xie et al., 2024) and the InterAct-refined (Xu et al., 2025) NeuralDome (Zhang et al., 2023), OMOMO (Li et al., 2023) and IMHD (Zhao et al., 2024), and train for $60{,}000$ iterations in total with AdamW (Loshchilov &Hutter, 2019) (lr=10^{-4}, batch size 1024).

Encoder: The encoder is a per-joint transformer (256-D, 6 layers, 8 heads) over the joint rotations that cross-attends to the object control points $\mathbf{f}^o$, so that the tokens describe the pose relative to the object it interacts with, and emits 23 tokens (21 body joints $+\,2$ hand tokens, one per hand) of $D{=}8$ dimensions, each quantized to 8 levels and lifted to 128-D for the decoder. LEXIS is thus the implicit grid of $8^8{\approx}1.7{\times}10^7$ entries with no learned parameters. Every code is reachable, so there is no codebook collapse and no commitment or EMA loss (Mentzer et al., 2024), and the space is bounded and fully covered, which is what lets LEXIS-Flow sample it directly (Section 3.2).

Decoder: D_ψ is an LRM-style (Hong et al., 2024) transformer that writes the triplanes, with 1728 learned tokens per region (3×24^2, width 256) refined by 6 blocks of 8-head attention. Each block cross-attends to the 23 LEXIS tokens and to the 128 control points $\mathbf{f}^o$. Object control points are derived from PointNeXt pretrained on ModelNet40 (Wu et al., 2015) shape classification. The decoder is conditioned on the object because LEXIS tokens are meant to store the semantics of the

interaction, and not the geometry of that particular object, so the decoder needs the actual shape to predict the triplanes correctly. This is also what makes the tokens transferable at test time, since a code decoded onto the control points of a slightly different chair morphs the planes to that chair instead of recalling the training one. The InterField readout MLP has two layers ($[128, 64]$ hidden units) with one head per field, and the body pose head is a 4-layer transformer with one learned query per joint.

Losses: In Eq. (3), $\mathcal{L}_{\text{recon}} = \left\|\mathcal{M}(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta}) - \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})\right\|_2$, where $\mathcal{M}$ is the SMPL+H (Romero et al., 2017) mesh for the estimated or GT pose and the GT shape $\boldsymbol{\beta}$. $\mathcal{L}_{\text{d}} = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{p}\in\mathcal{S}} |\hat{d}(\mathbf{p}) - d(\mathbf{p})|$ and $\mathcal{L}_{\text{u}} = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{p}\in\mathcal{S}} \left(1 - \hat{u}(\mathbf{p})^\top u(\mathbf{p})\right)$ over 1024 points $\mathcal{S}$ sampled uniformly on the body and, separately, on the object.

S.2.2 LEXIS-FLOW

LEXIS-Flow is a 16-layer mixture-of-transformers (MoT) backbone (SAM 3D Team et al., 2026) (512 hidden units). Image and text are encoded by frozen DINOv3 (ViT-L/16) (Siméoni et al., 2026) and the text tower released with it. The conditional flow-matching loss, $\mathcal{L}_{\text{FM}}$, is the squared error between the predicted velocity and the target, in expectation over the two timesteps and the noise:

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

where $\mathbf{x}_{t_1,t_2}$ is the noised state with body components interpolated at t_1 and object components at t_2, $\mathbf{u}_{\text{target}}$ is the conditional vector field target, and the expectation is approximated by mini-batch averaging. 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 (Loshchilov &Hutter, 2019) (batch size 1024). The learning rate is 5×10^{-4} after a 5k-iteration linear warmup, with no further schedule; we use a 0.3 dropout rate for classifier-free guidance, with guidance-scale 1.5 at inference. LEXIS-Flow trains on the 3D-paired images of ProciGen only. Training takes $\sim$48 hours on four A100 (64 GB) GPUs. We use Euler ODE solver (25 steps) for sampling and perform guidance/refinement (Eq. (6)) starting at t_{start}=0.6 (step 15), λ_{mask}=0.5.

S.2.3 TEXT CONDITIONING

Template: Every description is one sentence built from a fixed template with following slots that a vision-language model can read off an image, namely the object name, the stance (standing, sitting, lying, kneeling, unknown), a coarse verb (holding, carrying, sitting on, lying on, pushing or pulling, touching, not touching), the contact height (below the hips, torso, above the shoulders), the proximity (against the body, within reach). For example, "A person is sitting on the chair with both thighs, low down, against the body."

Training captions: Training descriptions are derived from the datasets' annotations (contact labels and poses), not from a language model. Each training frame carries eight variants: one clean, and seven with random slot dropout (probability 0.30) and slot noise (0.10), so the model learns to tolerate missing or wrong slots. For classifier-free guidance the text is dropped with probability 0.30, and image and text jointly with probability 0.10. When text is dropped it is replaced by a learned null token.

Description	$\Delta\text{CD}_{\text{hum}}$	$\Delta\text{CD}_{\text{obj}}$
None (null token)	+0.05	+0.52
Qwen2.5-VL 7B (Bai et al., 2025)	+0.19	+1.08
Claude Sonnet 5 (Anthropic, 2026b)	+0.13	+0.57
Claude Opus 5 (Anthropic, 2026a)	+0.01	+0.21
Annotation-derived (oracle)	0	0

Table S.1: **Text conditioning.** Change (cm) of CD_{hum} and CD_{obj} on BEHAVE relative to the annotation-derived description, same model, keys and noise.

Encoding: The text embeddings are L_2-normalized per token, linearly projected to the width of the image tokens and appended to them in the cross-attention context of the flow transformer. The no-text rows of Tables 2 and 3 come from the same model given the null token instead of a description.

Test time: Given the image, a vision-language model fills the six slots as a structured record, and its turned into the sentence, so test-time descriptions are in distribution by construction. Records

that fail validation or leave every slot empty fall back to the null token. For one BEHAVE frame of a person picking up a backpack, the descriptions read "A person is standing near the backpack, low down, within reach" (Claude Opus), "A person is standing and holding the backpack with a hand, low down, within reach" (Claude Sonnet) and "A person is standing and carrying the backpack with both hands, at torso height, against the body" (Qwen2.5-VL 7B).

Effect: Table S.1 compares descriptions from Claude Opus, Claude Sonnet and the open Qwen2.5-VL (Bai et al., 2025) 7B, and no description at all. All runs use the same model, keys and noise. The reference is the annotation-derived description, an oracle since it is filled from the 3D annotations of the test frame. Claude Opus stays within the noise of the oracle on both metrics, so it is the model we use for all text-conditioned results, whereas the smaller models drift further and no description costs the most.

S.2.4 MoGe-Based Object Initialization

For each refinement method (rows D and F–I) in Table 2, we apply the same initialization, as follows. We initialize the object pose R^o, t^o from SAM3D (SAM 3D Team et al., 2026). The rotation estimate is empirically accurate in many cases, so we keep it unchanged. However, the translation and scale are noisy. We align these via MoGe-estimated (Wang et al., 2025a) metric depth. Let the median MoGe depths under body and object masks (Kirillov et al., 2023) 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 (Patel &Black, 2025) 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.2.5 Evaluation Metrics

For the human body, as in Nam et al. (2024); Wang et al. (2025b), we apply Procrustes alignment on the estimated SMPL+H mesh (to align it to the GT SMPL-X (Pavlakos et al., 2019) 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):

$$\text{CD}(\widehat{\mathbf{V}}, \mathbf{V}) = \frac{1}{|\mathbf{V}|} \sum_{\mathbf{v} \in \mathbf{V}} \min_{\hat{\mathbf{v}} \in \widehat{\mathbf{V}}} \|\mathbf{v} - \hat{\mathbf{v}}\| + \frac{1}{|\widehat{\mathbf{V}}|} \sum_{\hat{\mathbf{v}} \in \widehat{\mathbf{V}}} \min_{\mathbf{v} \in \mathbf{V}} \|\mathbf{v} - \hat{\mathbf{v}}\|, \tag{S.2}$$

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:

$$P = {}^{TP}/_{|\hat{M}|}, \quad R = {}^{TP}/_{|M|}, \quad F1 = {}^{2PR}/_{P+R}, \tag{S.3}$$

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:

$$\text{Collision}(\mathbf{V}^b, \mathbf{V}^o) = \frac{1}{|\mathbf{V}^b|} \sum_{\mathbf{v}^b \in \mathbf{V}^b} \mathbb{I}(\mathbf{v}^b \in \mathbf{V}^o), \tag{S.4}$$

where $\mathbf{V}^b, \mathbf{V}^o$ are the set of body and object vertices, respectively, $\mathbb{I}(\mathbf{v}^b \in \mathbf{V}^o)$ 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 (Jiang et al., 2021). Lower is better (↓).



Figure S.1: **Stability test.** Contacts (orange) and the gravity wrench at the object centroid on a BEHAVE frame. The arrow is red when the contacts cannot balance gravity and green when they can. CONTHO (left) holds the box with one hand while the other arm passes through it, so the contacts cannot balance gravity, whereas LEXIS-Flow (right) holds it with both hands.

Physical Plausibility (Stability): Accuracy metrics do not tell whether a reconstructed interaction would actually hold, so we test static equilibrium under gravity for each test frame (Fig. S.1). Object surface points within 5 cm of the body count as human contacts and points within 3 cm of the floor plane fitted to the BEHAVE background scan count as ground contacts, each with a Coulomb friction cone (coefficient 0.5) discretized into 8 edge rays (Murray et al., 1994). The object must resist unit-mass gravity at its centroid, and we measure how far this required wrench lies from the cone of wrenches the contacts can generate, normalized by the wrench norm and with torques scaled by the bounding-ball radius, in the spirit of grasp-quality measures in wrench space (Ferrari &Canny, 1992; Zhu et al., 2026). A residual of 0 means the contacts hold the object and 1 means nothing supports it, so we count a frame as supported when the residual is below 0.1, with contact-free frames counted as unsupported, and report the percentage of supported test frames. Objects resting on tables cannot be supported under this test because BEHAVE has no table geometry, which is why even the ground-truth meshes reach only 88.9%.

S.2.6 Evaluation Protocol

Open3DHOI: HDM runs in its object-specific version for the classes it was trained on and in its general version for all others. The refinement baselines HOI-Gaussian, InteractVLM++ and TeHOR receive the same CameraHMR + SAM3D initialization, aligned with MoGe depth, as LEXIS-Flow* (Section S.2.4), so that any difference comes from the refinement itself. InteractVLM originally initializes the object pose by ICP, which fails in difficult scenes, so we report this original setting as "InteractVLM" and the shared initialization as "InteractVLM++."

BEHAVE: The baselines are trained on BEHAVE, whereas LEXIS-Net and LEXIS-Flow are not (Section S.2), so the comparison in Table 3 favors the baselines. We evaluate on the official test split (Xie et al., 2022) in the generative setting, where LEXIS-Flow samples the HOI state from noise given the image, the object shape and, when used, the text description. All methods are evaluated under one protocol, with Chamfer distances computed on 10k surface samples.

S.3 Additional Analysis

Latent structure: We embed the BEHAVE test set, whose interactions are unseen by LEXIS-Net, into LEXIS tokens and classify the action of each code with a k-NN classifier (k=10, majority

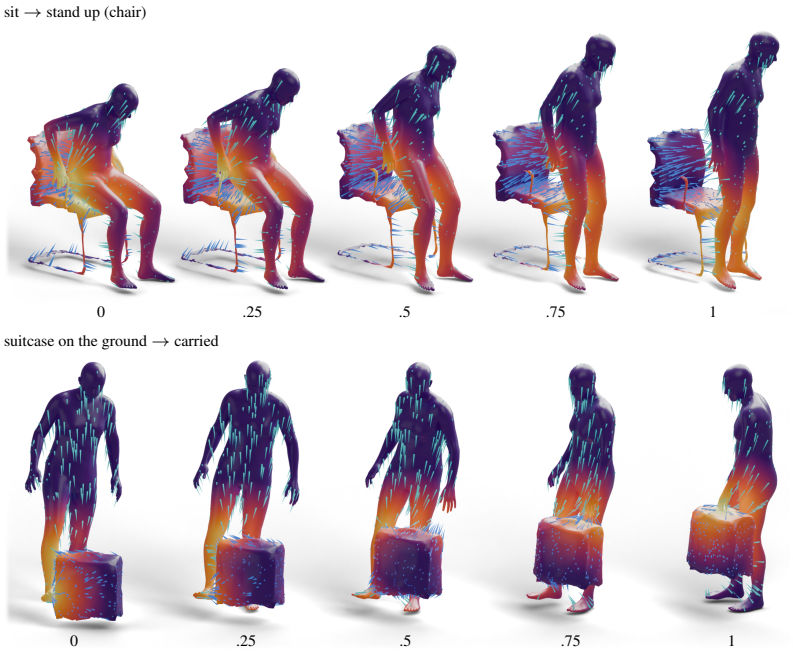

Figure S.2: **LEXIS interpolations.** Each row decodes five states of a linear interpolation of $\hat{\mathcal{Z}}$ between two BEHAVE captures of the same physical object, decoded by LEXIS-Net, which never saw BEHAVE. Object pose is interpolated between the two captures, since $\hat{\mathcal{Z}}$ says nothing about the object position.

vote). The accuracy is 87.3% for LEXIS and 59% for a Gaussian-VAE with the same architecture, so nearby tokens describe the same/similar interactions. The space is also smooth, as Fig. S.2 shows for two pairs of signatures at full width, extending Fig. 4.

Unseen interaction classes: LEXIS-Flow trains on the synthetic ProciGen only, so all Open3DHOI images are out of distribution, but some object classes overlap. We split the Open3DHOI classes into seen and unseen during training; 61% are unseen for both LEXIS-Flow and HDM. Table S.2 reports the relative CD_{obj} reduction over each baseline on the two subsets. The margin does not shrink on the unseen classes; it stays at 36% over HDM and grows from 24% to 29% over HOI-Gaussian, from 35% to 41% over InteractVLM and from 34% to 38% over TeHOR, so the gain is not explained by objects seen in training.

CD_{obj} reduction	Seen	Unseen
LEXIS-Flow vs. HDM	36%	36%
LEXIS-Flow* vs. HOI-Gaussian	24%	29%
LEXIS-Flow* vs. InteractVLM	35%	41%
LEXIS-Flow* vs. TeHOR	34%	38%

Table S.2: **Seen vs. unseen classes.** Relative CD_{obj} reduction of our two settings (Table 2) over each baseline on Open3DHOI.

	$\Delta\text{CD}_{\text{obj}}$, rotation noise σ_R		$\Delta\text{CD}_{\text{hum}}$, pose noise σ_θ	
σ (°)	LEXIS-Flow*	HOI-Gaussian	LEXIS-Flow*	HOI-Gaussian
5	**+0.1**	+0.3	+1.2	**+1.0**
10	**0.0**	+0.7	**+1.8**	+3.5
20	**+0.1**	+2.2	**+2.2**	+8.3
45	**+0.4**	+5.6	–	–

Table S.3: **Sensitivity to the initialization.** Delta (cm) against the unperturbed initialization (rows D and H of Table 2) when its object rotation or per-joint body pose is perturbed with Gaussian noise.

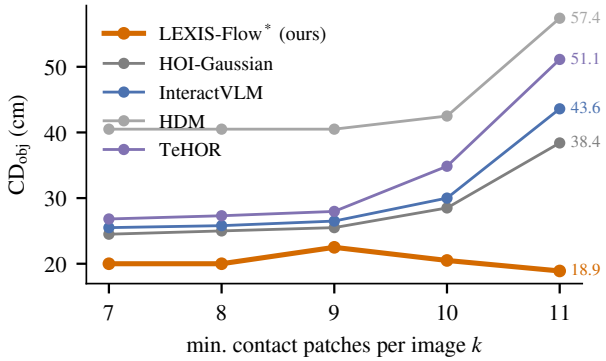

Figure S.3: **Gain vs. contact frequency.** CD_{obj} on the Open3DHOI images with at least k contact patches (connected components of the ground-truth contact map). The margin of LEXIS-Flow* over the baselines grows with k, so the InterField guidance helps most for high-frequency contacts.

Sensitivity to the initialization and to the object prior: Table S.3 perturbs the initialization given to all methods: LEXIS-Flow* is stable up to $45°$ of object-rotation noise, while HOI-Gaussian degrades steadily. Replacing the SAM3D shape with InstantMesh (Xu et al., 2024) or Hunyuan3D (Tencent Hunyuan3D Team, 2025), posed by FoundationPose (Wen et al., 2024), raises CD_{obj} of LEXIS-Flow* by 2.9 and 1.4 cm, against 4.1 and 3.4 cm for HOI-Gaussian. Thus, the gains do not depend on one object prior.

Contact frequency: We measure whether dense InterFields help more when the interaction has more, smaller contacts. We count contact patches (connected components of the ground-truth contact map) per Open3DHOI image and evaluate on the images with at least k patches. Figure S.3 shows that the baselines degrade as k grows while LEXIS-Flow* stays flat, so its margin is largest for high-frequency contacts, where sparse binary contact carries the least information.

Runtime and post-hoc optimization: All methods share the SAM3D initialization (about 10 s). Refinement then takes 0.7 s for LEXIS-Flow*, 50 s for InteractVLM, 120 s for HOI-Gaussian and more than 360 s for TeHOR, so in total LEXIS-Flow* is $4.7\times$, $11\times$ and over $30\times$ faster. Running a post-hoc contact optimization after LEXIS-Flow* multiplies the runtime by 40 and improves CD_{obj} by 0.9 cm, so we do not use it.

S.4 Qualitative Evaluation

S.4.1 Additional Results

Additional comparisons: Figure S.4 compares LEXIS-Flow* with the baselines of Table 2 on further in-the-wild images with diverse, complex interactions, and the supplementary video shows these results from rotating viewpoints. Figure S.5 does the same on BEHAVE against the baselines of Table 3, extending the last row of Fig. 5. The final row is a failure case, discussed below.

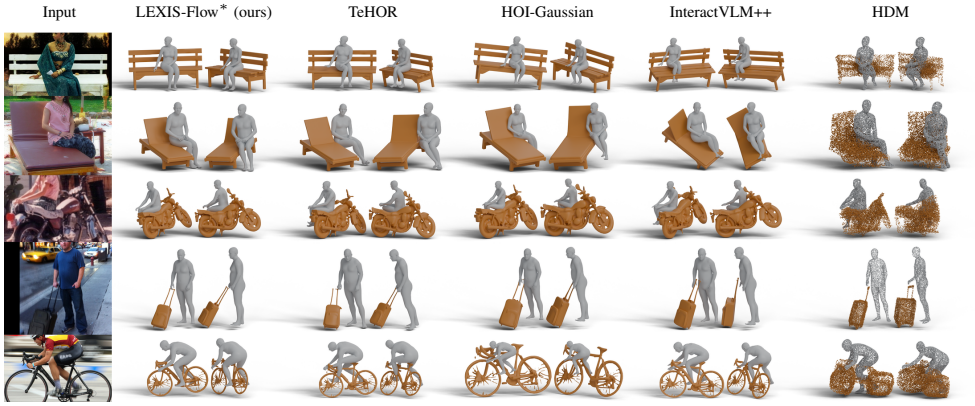

Figure S.4: **LEXIS-Flow vs SotA methods.** Five in-the-wild Open3DHOI images (Wen et al., 2025), each reconstructed by LEXIS-Flow* (ours), TeHOR, HOI-Gaussian, InteractVLM++ and the template-free HDM (Xie et al., 2024) (a point cloud) and shown from the input view and a rotated view. The refinement baselines start from identical initialization, so a row compares the refinement itself. Across a wide variety of objects, LEXIS-Flow estimates more physically-plausible 3D HOI, capturing better physical coupling while reducing floating artifacts and interpenetrations. For more results shown with a rotating viewpoint, see our supplementary video. Here: 🔍 **Zoom in** to see 3D details.



Figure S.5: **Additional BEHAVE comparisons.** For each example we show the input image, then every method twice: overlaid on the input viewpoint and from a side view that exposes the depth arrangement. The first five rows are examples on which LEXIS-Flow attains the lowest human *and* lowest object error against all three baselines. The last row, set apart below the rule, is a representative failure: on `tablesquare` our object orientation is off by roughly a quarter turn, while its placement stays as accurate as in the rows above.

Failure Cases: Performance can degrade when poor off-the-shelf initialization exceeds the recovery capacity of the LEXIS-Flow, or when the MoGe alignment fails due to bad depth predictions, or for highly atypical interactions underrepresented in existing datasets (Xie et al., 2024; Li et al., 2023; Zhang et al., 2023; Zhao et al., 2024). In such Out-Of-Distribution (OOD) cases, the LEXIS manifold (and the decoded InterFields) may misguide LEXIS-Flow, causing floating or penetrating artifacts. A second, more specific mode is visible in the last row of Fig. S.5: the object is placed at the right location but rotated by about a quarter turn, an orientation error we observe consistently on that object category rather than in isolated frames.



Figure S.6: **Failure cases.** Errors arise from inaccurate body pose and depth estimation. This can lead to misplaced human–object configurations. Each example is the input photo next to our reconstruction in the input view and rotated. Reading left to right, top to bottom: the body penetrates the couch; the elephant is reconstructed at the wrong scale; the person makes no contact with the bench; the body pose on the chair is wrong; the skateboard is translated away from the feet; and the surfboard penetrates the torso.

S.4.2 Perceptual Study

To evaluate perceived realism, we conduct a perceptual study. We randomly sample 60 images from Open3DHOI (Wen et al., 2025) and reconstruct each using our LEXIS-Flow* and the SotA HOI-Gaussian (Wen et al., 2025) 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.7.

Introduction: We compare two methods that reconstruct 3D human–object interactions from a single image. Each method takes a single RGB image as input and reconstructs a 3D human body (shown in blue) and the object of interaction (shown in brown).

Criteria: The 3D body should have a natural pose, and the 3D object should have a plausible shape. Moreover, the interaction between these entities should resemble the interaction depicted in the target image.

Your task: The study comprises 60 examples. For each example, you will see two rotating videos side by side, and you will have to choose between Option 1 and Option 2 by answering the following question:

Which reconstruction (Option 1 or Option 2) better matches the target image and is more physically plausible?

You may pause the videos to observe details more carefully.

Your participation is completely anonymous. The study should take about 10 minutes to complete.

Figure S.7: **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 (Wen et al., 2025) method. Then, for each image they select the reconstruction that better matches the image and appears more physically plausible.