GRAFT: Geometric Refinement and Fitting Transformer for Human Scene Reconstruction

Pradyumna YM[1,2], Yuxuan Xue[1,2†], Yue Chen[3], Nikita Kister[1,2]
István Sárándi[1,2], and Gerard Pons-Moll[1,2,4]

[1]University of Tübingen [2]Tübingen AI Center [3]Westlake University
[4]Max Planck Institute for Informatics

Abstract. Reconstructing physically plausible 3D human-scene interactions (HSI) from a single image currently presents a trade-off: optimization based methods offer accurate contact but are slow (∼20 s), while feed-forward approaches are fast yet lack explicit interaction reasoning, producing floating and interpenetration artifacts. Our key insight is that geometry-based human–scene fitting can be amortized into fast feed-forward inference. We present **GRAFT** (**G**eometric **R**efinement **A**nd **F**itting **T**ransformer), a learned HSI prior that predicts *Interaction Gradients*: corrective parameter updates that iteratively refine human meshes by reasoning about their 3D relationship to the surrounding scene. GRAFT encodes the interaction state into compact body-anchored tokens, each grounded in the scene geometry via *Geometric Probes* that capture spatial relationships with nearby surfaces. A lightweight transformer recurrently updates human meshes and re-probes the scene, ensuring the final pose aligns with both learned priors and observed geometry. GRAFT operates either as an end-to-end reconstructor using image features, or with geometry alone as a transferable plug-and-play HSI prior that improves feed-forward methods without retraining. Experiments show GRAFT improves interaction quality by up to 113% over state-of-the-art feed-forward methods and matches optimization-based interaction quality at ∼50× lower runtime, while generalizing seamlessly to in-the-wild multi-person scenes and being preferred in 64.8% of three-way user study. Project page: https://pradyumnaym.github.io/graft.

Keywords: Human-Scene Interaction · 3D Reconstruction · 3D Holistic Understanding

1 Introduction

Holistic 3D understanding of humans and their surrounding environments is essential for emerging technologies such as embodied AI, sports analytics, and augmented reality. Such applications require more than just the correct spatial localization of subjects; they demand explicit *interaction reasoning*—the ability to understand how a human relates to the 3D physical constraints of a scene to

† Corresponding Author

Fig. 1: Fast, geometry-grounded human–scene reconstruction. *Left:* From a single RGB image, GRAFT jointly reconstructs the 3D human and scene with physically coherent interactions. *Right:* GRAFT breaks the traditional speed–accuracy tradeoff, achieving high interaction accuracy at near feed-forward inference speeds, while existing methods typically gain one at the expense of the other.

reconstruct coherent contact and reason about affordances. While monocular human mesh recovery (HMR) has seen significant advances, estimating physically plausible human-scene alignment remains a formidable challenge: depth-scale ambiguity compounds with incompatible learned biases across separate human and scene models, so their outputs are rarely metrically or geometrically consistent with one another.

Current state-of-the-art HSI reconstruction methods [14, 15, 32] tackle these challenges using off-the-shelf human-scene initialization, and then optimizing them against energy functions that encourage realistic HSI. However, this optimization is expensive (∼20 s per instance [32]), susceptible to local minima due to its reliance on analytical gradients, and unable to learn generalizable interaction priors, as geometry enters only as a hard test-time constraint.

To improve inference speed, recent feed-forward methods [2, 10] build on foundational pose and depth models to jointly regress humans and scenes in a single pass. However, they perform no explicit interaction reasoning: scene geometry is never queried during pose decoding, so the model relies entirely on learned priors—producing meshes that frequently float above or penetrate the very scene it reconstructed. In practice, a costly test-time optimization such as PhySIC [32] is still required to obtain accurate HSI reconstructions.

What is missing is a *learned HSI prior*: a model that understands how humans physically relate to 3D scenes and can resolve floating and interpenetration in a single feed-forward pass—without costly test-time optimization. We therefore introduce **GRAFT** (**G**eometric **R**efinement **A**nd **F**itting **T**ransformer). Our key insight is to *amortize* geometry-based optimization into a feed-forward transformer: instead of minimizing handcrafted energy functions at test time, GRAFT directly regresses the corrective trajectory—the *Interaction Gradient*—from an implausible human-scene state to a physically grounded configuration. Operating on explicit 3D inputs, GRAFT supports two complementary modes: conditioned on image features it serves as a standalone HSI reconstruction system, and on geometry alone it acts as a universal plug-in prior that improves any existing method without retraining.

Concretely, GRAFT encodes the interaction state into a set of *HSI Tokens*—one per body joint, one per hand, and one for the full body—each grounded in the scene via *Geometric Probes*: nearest-neighbor queries that capture metric distances, directions, and surface normals from the surrounding scene point cloud. A lightweight transformer refines these tokens via self-attention to model inter-limb dependencies (*e.g.*, support-contact coupling between feet and torso) and *Geometry-Aware Cross-Attention* to fuse geometric cues with image features, then decodes corrective parameter updates. Crucially, this is *recurrent*: probes are recomputed after each update, giving the network direct physical feedback that progressively resolves penetrations and establishes contact.

Experimentally, GRAFT improves contact F1 by up to 113% over feed-forward baselines and matches optimization-based PhySIC at ∼50× lower runtime (0.38 s vs. 20 s). In geometry-only mode it boosts Human3R's contact F1 by up to 44% without retraining. Our method generalizes to diverse in-the-wild images, including multi-person scenes and complex interactions, and is preferred in 64.8% of trials in a three-way user study. Our contributions are as follows:

- **GRAFT**: a learned HSI prior that amortizes geometry-based optimization into a feed-forward transformer, predicting corrective *Interaction Gradients* through a recurrent refinement loop, explicitly reasoning about 3D HSI.
- **HSI Tokenization**: *Geometric Probes* and *Geometry-Aware Cross-Attention* anchor body-part queries directly to the scene point cloud and fuse metric 3D cues with image features, encoding the full interaction state in only 24 compact tokens.
- **Dual-mode architecture**: conditioned on image features, GRAFT serves as a standalone HSI reconstruction system; on geometry alone, it acts as a universal plug-in prior that improves any existing method without retraining, boosting Human3R's contact F1 by up to 44%.

2 Related Works

2.1 3D Human-Scene Reconstruction

Despite strong progress in monocular human pose estimation [16, 20], physically plausible human–scene interaction (HSI) reconstruction remains difficult. Optimization-based methods such as PROX [4] model contact and interpenetration explicitly, but depend on static scene scans and expensive per-instance fitting. Follow-up methods (e.g., HolisticMesh [30, 33] and related RGB approaches) reduce static scene assumptions, while recent systems such as PhySIC [14, 15, 32] combine optimization with stronger depth/geometry priors. These pipelines preserve explicit interaction

Method	HSI Reasoning	Feed-forward	Multi-Human	Runtime
PROX [4]	✔	✘	✘	73 sec.
HolisticMesh [30]	✔	✘	✘	5 min.
PhySIC [32]	✔	✘	✔	20 sec.
Human3R [2]	✘	✔	✔	0.24 sec.
UniSH [10]	✘	✔	✘	0.36 sec.
Ours	✔	✔	✔	0.38 sec.

Table 1: GRAFT achieves strong multi-human HSI reasoning at competitive feed-forward inference speeds.

reasoning but remain slow and sensitive to initialization. More recently, feed-forward video methods such as Human3R [2] and UniSH [10] achieve fast inference, but lack explicit HSI reasoning. As a result, they struggle with complex interactions (e.g., leaning on a wall, or lying on a couch) where multiple body parts must respect fine-grained surface constraints simultaneously. GRAFT is designed to achieve the best of both worlds, combining feed-forward efficiency with explicit 3D HSI reasoning, faithfully recovering the full range of complex body–scene contacts while maintaining competitive runtime (Table 1).

2.2 Human-Scene Interaction Prior

Early human-prior models such as Pose-NDF [23], NRDF [6], and VPoser [17] learn pose distribution/manifolds in isolation, without explicit scene grounding. Explicit HSI-prior methods [5, 34] condition on scene geometry, but follow a two-stage *generate-then-optimize* paradigm: they first predict plausible contact maps or placement distributions, then run costly test-time optimization to fit candidate poses to each scene. This decouples scene sensing from pose correction, making the pipeline slow and fragile—the optimizer must bridge the gap between a coarse contact prediction and a full-body solution, often falling into local minima. Moreover, these methods encode the scene with dense body-centric representations ($\sim$4K–10K parameters [5, 19, 34]), which risk overfitting to training-set geometry and scale poorly to partial observations.

GRAFT instead operates as a *corrective* prior: given an existing (possibly wrong) human–scene configuration, it directly predicts interaction gradients that move the mesh toward a plausible state, unifying scene sensing and pose correction in a single forward pass. Its compact 24-token HSI representation ($\sim$500 geometric parameters) captures the dominant interaction signal with far fewer dimensions, improving generalization and enabling the model to function as a geometry-only plug-in prior on top of other methods (Sec. 4).

2.3 Iterative Refinement

Iterative residual refinement has seen widespread adoption in diverse vision systems [21, 22, 31]. This paradigm has naturally been adopted in human-centric tasks. ReFit [27] refines SMPL estimation in a recurrent loop by reprojecting 3D keypoints onto 2D image features. WiLoR [18] iteratively projects an initial mesh onto multi-scale image features to regress pose and shape residuals. Learned Vertex Descent (LVD) [3] predicts 3D per-vertex displacements from localized feature projections, effectively amortizing an optimization descent.

GRAFT is the first work exploring the iterative refinement for 3D *human–scene relationship*: geometric probes are recomputed from the updated mesh at every step, so as the body moves, distances, normals, and contact patterns shift—giving the network a direct physical gradient toward plausible interaction. Because this feedback is purely geometric, GRAFT can also operate without any visual features and still serve as a plug-in HSI prior atop any method (Table 2)—a capability no prior iterative method possesses.

3 Method

Given a single RGB image $\mathcal{I}_h \in \mathbb{R}^{H\times W\times 3}$ of humans in a scene, our goal is to coherently reconstruct both the 3D scene geometry $\mathcal{P}_s \in \mathbb{R}^{H\times W\times 3}$ and the human body meshes $\mathcal{M}_i = \text{SMPLX}(\boldsymbol{\Theta}_i)$, with i indexing the person instances. Each per-human parameter vector is $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{\beta}\}$, with pose $\boldsymbol{\theta} \in \mathbb{R}^{165}$, translation $\boldsymbol{t} \in \mathbb{R}^3$, and shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$. In contrast to one-shot methods, GRAFT iteratively predicts *interaction gradients*—learned corrective parameter updates—by repeatedly probing the local scene geometry and attending to image tokens, progressively driving each mesh toward a geometry-grounded solution.

Our framework has two core components (Fig. 2). We start from a coarse metric initialization of the human and scene using separate foundation models (Sec. 3.1), where the two predictions are typically still incoherent. We then apply GRAFT (Sec. 3.2), an iterative refinement transformer that leverages compact HSI tokens and geometric probes capturing human–scene point cloud relations, with optional Geometry-Aware Cross-Attention grounded in image evidence. Training objectives are described in Sec. 3.3.

3.1 HSI Initialization

To initialize iterative refinement, we use two foundation models: MapAnything [8] for scene geometry and features, and NLF [20] for initial human meshes $\mathcal{M}_i^{\text{init}}$. To recover occluded scene geometry behind humans, we first remove them from $\mathcal{I}_h$ using OmniEraser [29], similar to [32], producing $\mathcal{I}_s$. We then apply MapAnything to $\mathcal{I}_s$ and $\mathcal{I}_h$ jointly, yielding coherent scene pointmaps $\mathcal{P}_s$ and $\mathcal{P}_h$ in a shared camera frame; these features are reused in later stages. We recover metric scale using a monocular depth estimator [25], following [12, 13].

For coarse alignment, we project each mesh's head joint to 2D, retrieve the corresponding 3D scene point from $\mathcal{P}_h$, and compute a per-human depth-ratio scale $s_i = p_{i,z}^{\text{scene}} / p_{i,z}^{\text{head}}$ to bring the mesh into the metric frame of $\mathcal{P}_s$, yielding $\{\mathcal{M}_i^0\}_{i=1}^{N_h}$. This provides only a coarse initialization; GRAFT then independently refines each $\boldsymbol{\Theta}_i$ using local geometric evidence and learned interaction priors, naturally supporting multi-person scenes with shared weights.

3.2 Iterative Interaction Transformer

Our model takes as input the current interaction state defined by the scene pointmap $\mathcal{P}_s$ and human parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{\beta}\}$ and predicts the *interaction gradient*: a corrective $\Delta\boldsymbol{\Theta} = \{\Delta\boldsymbol{\theta}, \Delta\boldsymbol{t}, \Delta\boldsymbol{\beta}\}$ with an additional uniform scale s to compensate residual metric-scale mismatch from initialization. We propose parameter- and data-efficient architecture design by leveraging (a) a *compact HSI token representation* that captures localized 3D cues and (b) an *efficient Geometry-Aware Cross-Attention mechanism* that grounds these cues in visual evidence. In each recurrent step, the model closes the loop by re-sampling metric spatial evidence from the scene, enabling precise contact resolution that single-pass methods fail to achieve.

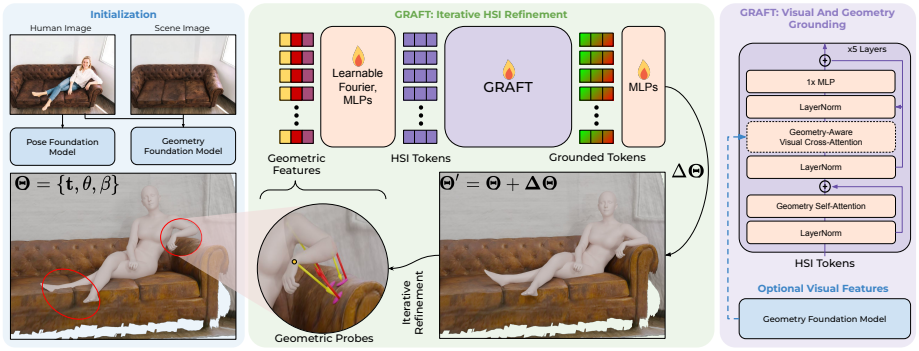

Fig. 2: Overview of GRAFT. *Left:* Foundation models initialize human meshes (NLF [20]) and scene geometry (MapAnything [8]), often yielding misalignments. *Center:* Geometric probes (nearest-neighbor scene points for body joints and body vertices) encode local contact cues (such as position and surface normals) into compact HSI tokens. GRAFT uses these tokens to predict iterative updates $\boldsymbol{\Theta}' = \boldsymbol{\Theta} + \Delta\boldsymbol{\Theta}$, correcting penetration and floating artifacts. *Right:* GRAFT alternates geometric self- and visual cross-attention for HSI-prior-guided refinement, optionally fusing image features.

HSI Tokenization with Geometric Probing Transformers are well suited for HSI refinement because they jointly model long-range body-part dependencies and can fuse geometry with image evidence through self-/cross-attention. We encode the current interaction state $(\mathcal{P}_s, \boldsymbol{\Theta})$ as a compact set of *HSI tokens* and keep this representation low-dimensional for both *generalization* and *efficiency*: it discourages overfitting to high-dimensional noise, encourages learning the underlying interaction manifold, and reduces quadratic attention cost. Our tokenization aggregates local geometric cues around the body into a fixed, small number of tokens (instead of raw points or dense maps), using ∼500 parameters versus ∼4K for BPS (basis-point-set distance) representations [19,34] and ∼ 10K for POSA (dense body-surface contact) representations [5].

Each token is built from one or more *geometric probes* that capture the local 3D human–scene relationship at a specific body location. Each probe is a spatial query anchored at a 3D point $\mathbf{p}$ (e.g., a joint or surface vertex). Since signed distance fields are ill-defined for partial observations such as pointmaps, for each $\mathbf{p}$, we retrieve the closest scene point $\mathbf{p}^*$ and compute the relative offset $\mathbf{v}_s = \mathbf{p}^* - \mathbf{p}$ together with the corresponding scene normal $\mathbf{n}^*$. We also express $\mathbf{p}$ in body-relative coordinates. Fused with visual features via cross-attention, these signals enable reasoning about local contact, penetration, and support. All geometric vectors $\{\mathbf{v}_s, \mathbf{n}^*, \mathbf{p}\}$ are encoded using learnable Fourier features [11].

We define 24 fixed HSI tokens for each human: 21 body joint tokens (non-root joints), 2 hand tokens (left/right), and 1 full-body token. Each token concatenates (i) geometric probe features and (ii) token-specific SMPL-X parameter context, then projects them to a shared embedding space via an MLP. In compact form,

$$\mathbf{z}_k^t = \phi_k\big([\mathbf{g}_k^t; \mathbf{p}_k^t]\big), \qquad \mathbf{Z}^t = \{\mathbf{z}_k^t\}_{k=1}^{24},$$

where $\mathbf{g}_k^t$ are geometric probe encodings and $\mathbf{p}_k^t$ are token-specific SMPL-X parameter features. Each *body joint token* carries one probe at its joint together with the joint's 6D rotation; each *hand token* aggregates five distal-joint probes and the corresponding finger rotations; and the *full-body token* aggregates 27 surface probes at fixed vertices uniformly sampled from the SMPL-X template mesh, together with global orientation, translation, and shape parameters. Tokenization details and the MLP architecture are provided in the supplementary.

Visual Anchors and Geometry-Aware Attention To ground each token in image evidence, we use two MapAnything feature streams: a scene stream from the human-removed image $\mathcal{I}_s$ and an interaction stream from the original image $\mathcal{I}_h$. For each token, we define one or more *visual anchors*: 3D body points $\mathbf{a} \in \mathbb{R}^3$ projected to both images by $\mathbf{u} = \pi(\mathbf{a})$. Around each $\mathbf{u}$, we sample multi-scale MapAnything features from both streams and concatenate them as the token context. Body tokens use one anchor per body joint, hand tokens use one anchor per hand (mean of distal joints), and the full-body token uses 27 surface anchors. We sample 3×3 MapAnything token neighborhoods around each body/hand anchor and 1×1 (single-token) features for the 27 full-body anchors to preserve efficiency.

We use a 5-layer transformer with standard multi-head attention, alternating: (i) self-attention among HSI tokens and (ii) constrained cross-attention to sampled visual features. In self-attention, geometry-aware tokens exchange global interaction context (e.g., support-contact dependencies between limbs and torso). In cross-attention, token k attends only to features sampled at its own anchors, yielding a sparse connectivity pattern that tightly couples geometric probes with local visual evidence. Let $\mathbf{V}^t = \{\mathbf{v}_k^t\}_{k=1}^{24}$ denote anchor features sampled from both streams. One attention block can be written as follows, with $\widehat{\mathbf{Z}}^t$ denoting the geometry-grounded interaction state:

$$\widetilde{\mathbf{Z}}^t = \texttt{Self-Attn}(\mathbf{Z}^t), \qquad \widehat{\mathbf{Z}}^t = \texttt{Cross-Attn}(\widetilde{\mathbf{Z}}^t, \mathbf{V}^t),$$

Iterative Interaction Refinement At iteration t, we decode step-wise updates directly from the geometry-grounded state $\widehat{\mathbf{Z}}^t$. We decode 6D rotations for the 21 body joints from the body tokens, 6D rotations for the 15 joints in each hand from the hand tokens, and global orientation from the full-body token, together with shape and translation updates. In compact form,

$$(\Delta\boldsymbol{\Theta}^t, s^t) = \Psi(\widehat{\mathbf{Z}}^t), \qquad \boldsymbol{\Theta}^{t+1} = \boldsymbol{\Theta}^t + \Delta\boldsymbol{\Theta}^t.$$

Lightweight MLP heads implement Ψ, predicting $\Delta\boldsymbol{\theta}^t$ (body, hand joint, and global rotations), $\Delta\boldsymbol{t}^t$, and $\Delta\boldsymbol{\beta}^t$, while a separate head on the full-body token predicts a uniform scale s^t applied to the vertices. We maintain the rotation representation in 6D for stable updates, then convert to axis-angle for the SMPL-X forward pass to produce the refined mesh $\mathcal{M}^t$.

This updated state is fed back into the next iteration: we recompute geometric probes, resample visual anchors, and re-run the transformer with shared

weights across timesteps. Over T iterations, the meshes progressively reduce interpenetration and improve contact and alignment with the scene geometry. We describe the supervision used to train this iterative trajectory in Sec. 3.3.

Fast Differentiable Scale Update In addition to pose, translation, and shape updates, GRAFT predicts a per-step uniform scale s^t applied to mesh vertices. During training, naively re-fitting SMPL-X shape coefficients at every step to match scaled vertices is prohibitively expensive. We therefore use a closed-form linear approximation: because shape blendshapes act linearly on vertices, a uniform scaling can be absorbed into shape parameters as

$$\boldsymbol{\beta}_s = s \cdot (\boldsymbol{\beta} + \mathbf{c}) - \mathbf{c},$$

where $\mathbf{c}$ is the least-squares projection of the mean body template into $\boldsymbol{\beta}$-space, computed once offline and fixed for a given body model (SMPL-X). Translation scales directly and rotations are unchanged, replacing complex per-step fitting with simple vector arithmetic. This yields an efficient, fully differentiable update for multi-step rollout supervision. Full derivation is provided in the Suppl..

3.3 Training Objectives

We train GRAFT on paired human–scene data with ground-truth SMPL-X annotations, supervising the full iterative trajectory by applying losses at every refinement step t rather than only at the final output. This per-step supervision exposes the network to a richer set of off-manifold states and encourages monotonic improvement across iterations.

Training queries. At each iteration we sample a mixture of NLF-initialized queries (realistic misalignments) and ground-truth parameters with random perturbations. This teaches the model to both (i) correct large initialization errors and (ii) remain stable near the optimum by predicting near-zero updates when the input is already well aligned.

Per-step losses. At each step t, we combine three complementary terms:

$$\mathcal{L} = \sum_{t=1}^{T} \left(\lambda_p \left\| \boldsymbol{\theta}^t - \boldsymbol{\theta}^* \right\|_2^2 + \lambda_v \left\| \mathcal{V}^t - \mathcal{V}^* \right\|_2^2 + \lambda_n \left\| \widetilde{\mathcal{V}}^t - \widetilde{\mathcal{V}}^* \right\|_2^2 \right),$$

where $\boldsymbol{\theta}^t$ denotes the predicted 6D joint rotations, $\mathcal{V}^t$ are camera-relative vertices (capturing the joint effect of $\boldsymbol{t}$ and scale s), and $\widetilde{\mathcal{V}}^t$ are mean-normalized vertices that isolate shape supervision from global placement. The rotation loss directly penalizes articulation errors in the continuous 6D space, while the two vertex losses decouple global positioning from body proportions. Notably, we do *not* employ explicit contact or interpenetration losses. Our geometric probes already encode local human–scene distances and normals at every step, providing dense spatial evidence from which contact and non-penetration constraints emerge implicitly through regression to ground-truth poses. Moreover, such geometric

Fig. 3: GRAFT as a learned HSI prior. Starting from an initial state (green mesh), we apply a translation and pose perturbations(red); after each perturbation, GRAFT—operating with *no visual features*—projects the state back to a geometrically valid human–scene interaction (green). Our geometric probes encode contact and penetration cues that drive each correction step. We refer readers to the supplementary video for a comprehensive visualization.

losses are ill-defined for the partial, single-view point clouds we operate on, where occluded surfaces introduce spurious nearest-neighbor correspondences. We find that rich per-step geometric input combined with trajectory supervision is sufficient to learn plausible interactions without auxiliary HSI losses (Sec. 4).

Training strategy. We stabilize iterative training with two techniques: *curriculum rollout*, which gradually increases the number of supervised refinement steps from zero to T, and *visual-anchor dropout*, which randomly drops per-token visual neighborhoods from both scene and interaction streams to prevent over-reliance on appearance. All hyperparameters (loss weights $\lambda_p, \lambda_v, \lambda_n$, mixture ratios, and training schedules) are provided in Sec. 4.

4 Experiments

4.1 Implementation Details

Datasets. We train GRAFT on a pseudo-labeled HSI dataset InHabitants [9], comprising 75k images and 97k human instances. Other datasets do not offer both realistic RGB and accurate interaction annotations: RICH [7] offers accurate motion captures and 3D scene scans, but is limited to <15 indoor/outdoor scenes, PROX-D [4] is captured on 12 indoor scenes and uses optimization pseudo-labels with unnatural and inaccurate poses, BEDLAM [1] lacks realistic interactions (humans float above furniture), and HUMANISE [28] provides no photorealistic RGB. GRAFT is inherently bounded by pseudo-label quality; scaling to richer data is a promising future direction.

Optimization. We train with Adam for 150k iterations on a single H100 GPU. The peak learning rate is 1×10^{-4}, with linear warmup followed by cosine decay.

Training queries and perturbations. Following Sec. 3.3, each training step uses a mixture of NLF-initialized queries and GT-based queries. For GT-based

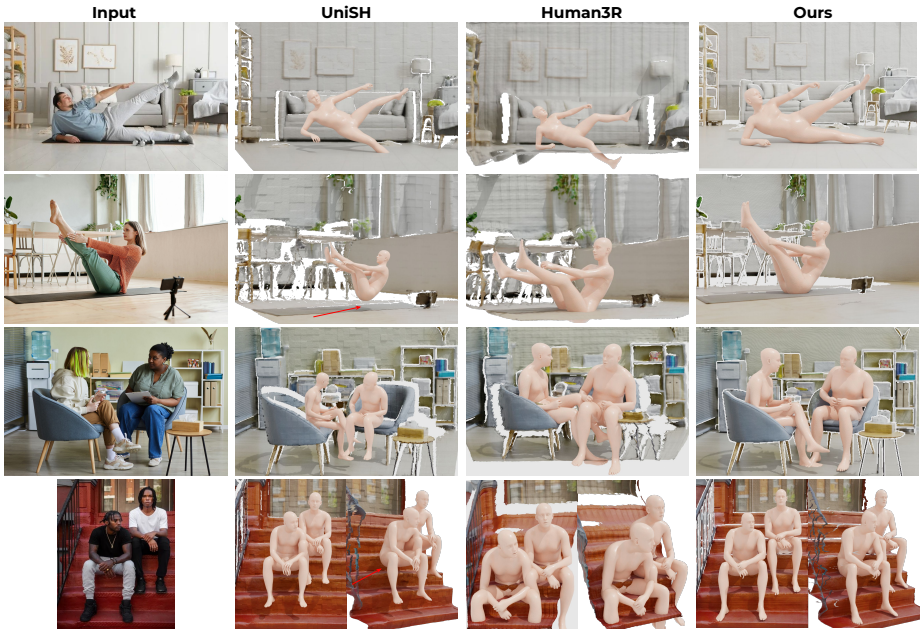

Fig. 4: In-the-wild qualitative comparison. We compare GRAFT against feed-forward methods, UniSH [10] and Human3R [2] on unconstrained internet images. While prior approaches localize humans in the scene, they lack explicit human–scene interaction modeling, often resulting in weak support, hovering or penetration. In contrast, GRAFT recovers physically coherent interactions with scene-consistent contact. We refer readers to the supplementary for more results.

queries, we keep clean GT states with probability 0.2, and with probability 0.8 we add Gaussian perturbations to improve robustness around the interaction manifold: translation noise with standard deviation 0.1 m, SMPL-X shape (β) noise with standard deviation 0.03, pose rotation noise with standard deviation 7°, and global orientation noise with standard deviation 3°.

Rollout training and scaling. We supervise all refinement steps and train with curriculum rollout, increasing from single-step supervision to full T=3-step supervision. Full rollout supervision starts after the first 10k training iterations. As in Sec. 3.2, refinement weights are shared across iterations. we set T=3 for both training and inference, as additional iterations yield only marginal gains (Fig. 6). During training, we also apply random global scale augmentation by sampling a factor from $[0.85, 1.15]$ and use our fast differentiable scale update (Sec. 3.2) to keep this operation efficient. We apply visual-anchor dropout with probability 0.35 per token.

4.2 Evaluation Protocol

Baselines. We train our method from scratch and compare against both optimization based and feed-forward baselines. Specifically, we evaluate PROX and

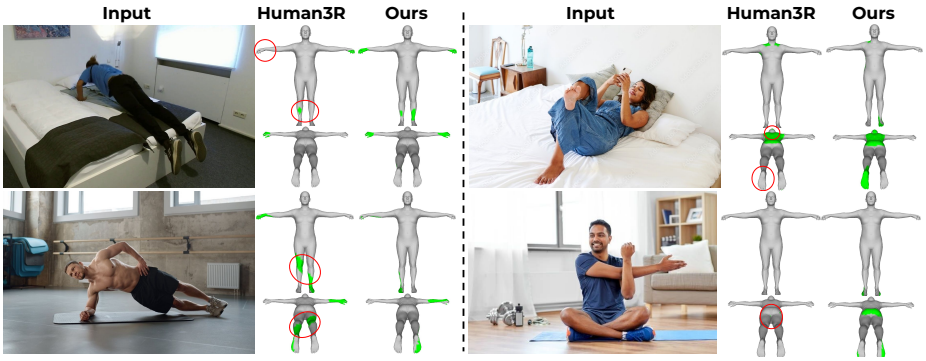

Fig. 5: Contact maps from geometry. By reconstructing accurate 3D human–scene interaction, GRAFT derives contact directly by spatial proximity, yielding sharp and reliable contact regions.

PhySIC (test-time optimization) as well as Human3R and UniSH (feed-forward methods trained on synthetic/large-scale video data). Human3R and UniSH are video-based models and, in the absence of dedicated single-image HSI reconstruction baselines, they are the closest feed-forward baselines for comparison. For fair comparison, and because interaction evaluation requires complete scene geometry, we pass the inpainted scene image $\mathcal{I}_s$ through the depth backbones used by the baselines (CUT3R [24] and Pi3 [26]) and robustly align scene points to their incomplete predictions, so all methods are evaluated under the same complete-geometry protocol.

Metrics. In addition to standard contact precision/recall/F1, we introduce two geometric interaction metrics that provide a continuous measure of human–scene relative positioning. Contact precision/recall/F1 remains our hard-constraint evaluation, while the new metrics quantify geometric consistency continuously. Using the PROX contact-vertex set [4], for each contact vertex we compute the vector from the human vertex to its nearest scene point in both prediction and ground truth. We then report (i) **V2S** (Vector-to-Scene): the weighted mean Euclidean error between predicted and ground-truth vectors (reported in mm), and (ii) **D2S** (Direction-to-Scene): the weighted mean angular error between these vectors (reported in degrees). Per-vertex weights are density-aware, so high-vertex-density regions (e.g., hands) do not dominate the error term.

Following the evaluation protocol of PhySIC, we report quantitative results on RICH-100 and PROX-test, both of which provide ground-truth human and scene scans. In addition, we present qualitative results on PiGraphs and demonstrate strong generalization on curated in-the-wild internet images.

4.3 Qualitative Evaluation

Compared with UniSH and Human3R, GRAFT produces more realistic human–scene interactions, especially at support and contact regions (e.g., feet–floor and

Test Data	Method	Feed-forward	Human Pose Metrics ↓			Contact Metrics ↑		
			PA-MPJPE	V2S	D2S	Precision	Recall	F1 score
RICH-100	PROX [4]	✘	120.24	538.67	79.38	0.069	0.253	0.108
	PhySIC [32]	✘	46.50	237.69	41.65	0.538	0.695	0.606
	UniSH [10]	✔	69.30	326.80	36.23	0.329	0.356	0.342
	Human3R [2]	✔	48.84	274.05	44.68	0.282	0.531	0.368
	Human3R + Ours	✔	48.84	240.83	40.68	0.378	0.655	0.479
	Ours	✔	49.30	222.54	36.80	0.441	0.784	0.565
PROX-test	PROX [4]	✘	76.98	224.49	59.60	0.458	0.144	0.219
	HolisticMesh [30]	✘	81.30	170.64	54.48	0.427	0.348	0.383
	PhySIC [32]	✘	44.17	148.05	47.11	0.550	0.424	0.479
	UniSH [10]	✔	59.02	256.38	60.34	0.362	0.175	0.236
	Human3R [2]	✔	59.25	200.35	59.36	0.228	0.324	0.268
	Human3R + Ours	✔	59.25	187.04	54.24	0.350	0.434	0.387
	Ours	✔	**54.83** (↓7.5%)	**188.71** (↓5.8%)	**51.67** (↓13.0%)	**0.526** (↑130.7%)	**0.627** (↑93.5%)	**0.572** (↑113.4%)

Table 2: Quantitative comparison on RICH-100 and PROX-test. GRAFT consistently improves interaction quality over recent feed-forward baselines, with the strongest gains on human-scene metrics and comparable PA-MPJPE. Gains are computed w.r.t. Human3R. Compared with optimization-heavy methods, GRAFT achieves similar interaction quality with much faster inference (Tab. 1).

body–furniture interfaces). While these baselines often localize the human reasonably in the scene, they frequently fail to capture fine interaction structure, leading to weak support, hover artifacts, or penetration near local geometry. In contrast, our iterative refinement predicts interaction gradients from geometry-grounded local probes and updates parameters incrementally rather than regressing the full pose from scratch, enabling reliable reconstruction even for challenging articulated poses. Because refinement is applied per human on shared scene geometry with shared weights, GRAFT naturally handles multi-person scenes while preserving coherent relative placement, and generalizes strongly to in-the-wild images despite training on a comparatively small pseudo-labeled dataset. As shown in Fig. 5, accurate contact maps emerge directly from the 3D reconstruction via spatial proximity, without contact-specific post-processing.

Fig. 3 provides a qualitative illustration of the learned HSI prior. Beginning from an initial reconstruction, we apply successive perturbations—first to global translation, then to body pose—and run GRAFT with no visual features. In both cases, GRAFT projects the perturbed state back onto the manifold of geometrically valid human–scene interactions, recovering plausible contact and support from geometry alone. This confirms that geometric probes are sufficient to capture a strong, generalizable interaction prior, and explains why the geometry-only mode transfers effectively as a plug-in prior to other feed-forward methods. We refer readers to the supplementary video for a more comprehensive visualization of this behavior across diverse scenes and perturbation magnitudes.

4.4 Quantitative Evaluation

As shown in Table 2, GRAFT consistently improves interaction quality over recent feed-forward baselines on both PROX-test and RICH-100, with the largest gains in contact and geometric interaction metrics, while keeping PA-MPJPE competitive. The reusable-refinement effect is also clear: applying GRAFT in

geometry-only mode (no visual features) on top of Human3R improves interaction quality on both datasets — on PROX-test, F1 improves from 0.268 → 0.387 (with V2S 200.35 → 187.04), and on RICH-100, F1 improves from 0.368 → 0.479 (with V2S 274.05 → 240.83) — demonstrating that GRAFT functions as a plug-in HSI refinement prior (see Fig. 3). Compared with optimization-heavy methods, GRAFT remains slightly behind the strongest optimization baseline on some metrics, but preserves feed-forward runtime (Table 1).

Fig. 6 shows the convergence behavior across refinement iterations. We observe the most significant F1 jump at the first iteration, indicating that a single update already corrects most coarse interaction errors. Further iterations continue to improve performance, but with diminishing returns, yielding marginal gains per step while runtime grows steadily.

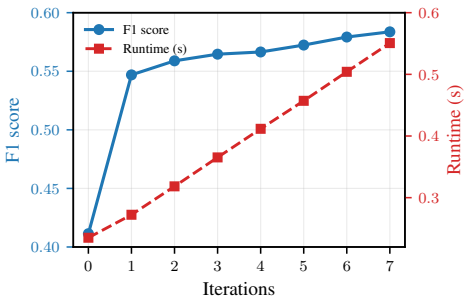

Fig. 6: Refinement iterations: quality–runtime trade-off. Contact F1 improves sharply at the first refinement iteration, while additional iterations provide marginal gains. Runtime increases approximately linearly with iteration count.

We conduct a perceptual study comparing GRAFT, UniSH, and Human3R via an online survey. Each participant evaluates a fresh random sample of 20 images (2 PROX-D, 3 PiGraphs, ∼100 in-the-wild) as interactive 3D reconstructions, selecting the most plausible human–scene interaction; method order is randomized and model names are hidden. Across n=53 participants (912 responses), GRAFT was chosen 591 times (64.8%), well above the 33% chance baseline (Fig. 3).

4.5 Ablation Study

Table 3 analyzes each design choice by measuring the degradation it causes relative to the full model; heatmap shading highlights the severity of each drop. As a data-control, Human3R retrained on our pseudo-labeled set still lags clearly behind GRAFT, confirming our gains are architectural. Notably, *Ours (Init.)* achieves a lower PA-MPJPE than the full model. This reflects an expected trade-off: PA-MPJPE is Procrustes-aligned and scene-agnostic, whereas GRAFT is supervised with pseudo-GT that intentionally adjusts articulation for scene consistency, diverging from mocap GT in pose space. The moderate PA-MPJPE increase (43.51→49.30, on par with Human3R's 48.84) is accompanied by large interaction gains (F1 0.411→0.565, V2S 240→223 mm), confirming that GRAFT learns scene-grounded rather than scene-agnostic articulation; this gap would shrink with higher-fidelity training data. *Rollout supervision* and *geometry grounded features* are the two most critical components: removing either causes the largest drops across interaction metrics (rollout: F1 0.565 → 0.358; geometric features: F1 0.565 → 0.389, V2S 222.54 → 273.33). Without rollout, the model

Method	Human Pose Metrics ↓			Contact Metrics ↑		
	PA-MPJPE	V2S	D2S	Precision	Recall	F1 score
RICH-100 Dataset						
Human3R*	54.78	334.73	47.86	0.310	0.444	0.365
Ours (Init.)	43.51	240.30	37.15	0.349	0.499	0.411
w/o rollout	51.30	274.83	50.74	0.270	0.532	0.358
w/o geometric feat.	48.76	273.33	42.19	0.315	0.509	0.389
w/o scale pred.	50.29	239.45	36.45	0.420	0.642	0.508
w/o GT-training	65.56	231.76	37.56	0.425	0.752	0.544
w/o visual feat.	49.60	223.00	33.99	0.449	0.777	0.569
Ours	49.30	222.54	36.80	0.441	0.784	0.565
PROX-test Dataset						
w/o visual feat.	52.07	185.65	49.63	0.424	0.538	0.474
Ours	51.55	188.26	52.67	0.408	0.658	0.504

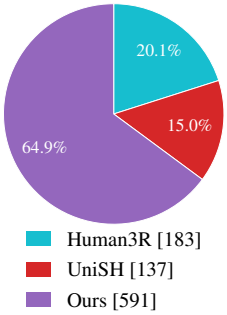

Table 3: *(Left)* Ablation study on RICH-100, with an additional visual-feature ablation on PROX. Heatmap shading indicates degradation severity among design-choice variants (worst , second , third). *Human3R retrained on the same training data as GRAFT. *(Right)* User study: GRAFT is preferred by participants in 64.8% of trials (912 responses, 53 participants), nearly twice the 33% chance rate.

cannot learn multi-step correction trajectories; without scene-grounded tokens, it loses direct evidence of contact and penetration. *Explicit scale prediction* provides a consistent moderate improvement (F1 0.565 → 0.508): it decouples scale from translation, whereas without it the model must absorb metric-scale errors through coupled translation and β updates, where the β–scale mapping is nonlinear. *GT-training* queries are essential for stability near the optimum: without them the model over-corrects already-good initializations (PA-MPJPE 49.30 → 65.56). Finally, *visual features* have a limited effect on RICH but provide a clearer gain on PROX, mainly through higher recall. High-dimensional visual tokens risk overfitting on our comparatively small training set, whereas low-dimensional geometric probes generalize more reliably; yet when multiple geometrically plausible corrections exist, image evidence disambiguates between possible solutions.

5 Conclusion

We presented GRAFT, a learned human–scene interaction prior that amortizes costly geometry-based optimization into a feed-forward transformer. By encoding the interaction state into compact, scene-grounded HSI tokens and refining them in a recurrent loop, GRAFT matches the interaction quality of optimization-based methods at ∼50× lower runtime while substantially outperforming existing feed-forward baselines. Because the model reasons over explicit 3D geometry, it also serves as a universal plug-in prior: applied in geometry-only mode atop Human3R, it improves contact F1 by up to 44% without retraining or access to image features, confirming that the learned interaction manifold transfers across methods. Current limitations include dependence on upstream scene reconstruction and human mesh estimation quality, the rigid-scene assumption of

the nearest-point geometric probes (*e.g.*, deformable surfaces are not modeled), and difficulty with heavily occluded multi-person interactions.

Acknowledgements. We thank the whole RVH team for the support, and especially Chuqiao Li for creating the GRAFT logo. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting PYM and YX. YC is funded by the Westlake Education Foundation. NK was supported by Bosch Industry on Campus Lab at the University of Tübingen. NK thanks the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program for support. IS and GPM were supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 409792180 (Emmy Noether Programme, project: Real Virtual Humans). GPM is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645 and is supported by the Carl Zeiss Foundation.

Author contributions. PYM led the project, including the core idea development, method design, implementation, experimental study, and writing. YX provided close supervision and contributed to project and method design, and writing. IS provided supervision and contributed to method design and writing. YC provided valuable insights for debugging Human3R baselines and helped with writing. NK supported the use of InHabit in this work. GPM supervised the project and contributed to idea development, project framing, and writing.

Supplementary Material

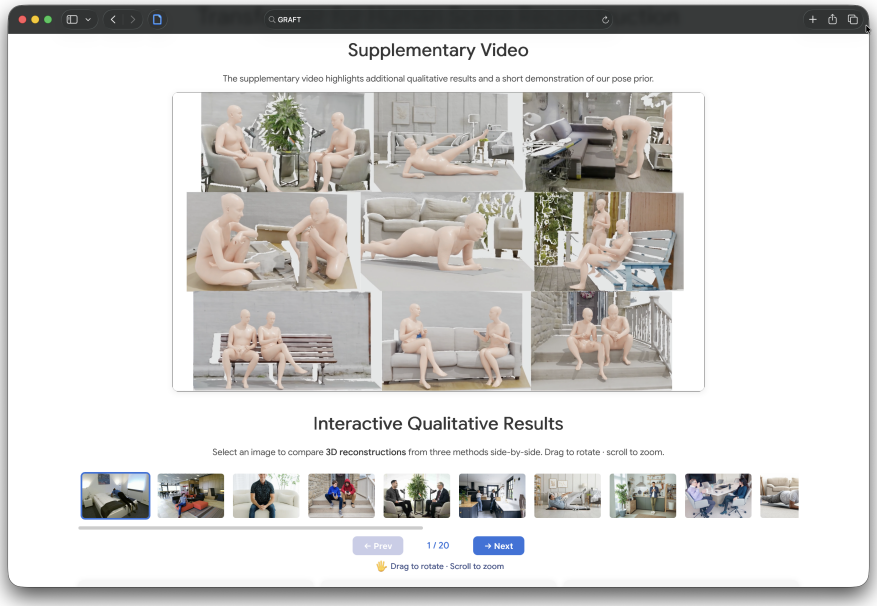

Fig. S1: Supplementary webpage. The webpage includes the supplementary video with reconstruction renderings, a demonstration of our pose prior, and interactive 3D results and method comparisons.

We encourage readers to visit the supplementary webpage shown in Fig. S1, which presents the supplementary video with reconstruction renderings, a demonstration of our pose prior, and interactive 3D results and comparisons. This supplementary material provides additional details complementing the main paper. Sec. S1 gives the etymology of the GRAFT acronym. Sec. S2 provides additional qualitative comparisons. Sec. S3 describes the user study setup and interface. Sec. S4 presents the full inference and training algorithms. Sec. S5 derives the fast differentiable scale update used during rollout training. Sec. S6 defines the V2S and D2S interaction metrics. Sec. S7 details hyperparameters, loss weights, and the model architecture.

S1 GRAFT: Definition

graft *verb.*
1. To join or integrate an element with another so as to bring about a close union.

Here: aligning human pose with scene geometry for physically plausible interaction. The name reflects the core idea: the model iteratively refines human–scene interactions by predicting corrective updates to pose, translation, and shape, yielding reconstructions that are better grounded in geometry.

S2 Additional Qualitative Results

Fig. S2 provides additional side-by-side comparisons of GRAFT, Human3R, and UniSH on in-the-wild scenes. Baselines frequently fail to capture fine interaction structure near contact regions, leading to hover artefacts or penetration, while GRAFT's iterative geometric refinement yields more accurate interactions.

S3 User Study Details

Participants (n=53) were each shown 20 scenes randomly sampled from a pool of PROX-D, PiGraphs, and ∼100 in-the-wild internet images, giving each participant a unique session. For each scene, all three reconstructions (GRAFT, UniSH, Human3R) were displayed side by side as interactive 3D viewers in a browser, with method order randomised and model names hidden. Participants selected the reconstruction with the most plausible human–scene interaction based on contact realism and pose accuracy; see Fig. S3 for the interface.

S4 Algorithm

Alg. S1 gives the full GRAFT inference procedure and Alg. S2 the training loop.

S5 Fast Differentiable Scale Update

Since uniform scaling leaves pose rotations unchanged, we ignore pose and define a simplified SMPL-X parameterization:

$$\mathrm{SMPLX}(\boldsymbol{\beta}) = \mathbf{T} + \mathbf{S}^\top \boldsymbol{\beta}, \tag{S1}$$

where $\mathbf{T} \in \mathbb{R}^N$ is the mean body template in a canonical pose ($N = 10475 \times 3$), $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are the shape coefficients, and $\mathbf{S} \in \mathbb{R}^{10\times N}$ are the shape blend shape directions. Given a predicted scale factor s, we seek scaled shape coefficients $\boldsymbol{\beta}_s$ satisfying:

$$\mathbf{T} + \mathbf{S}^\top \boldsymbol{\beta}_s = s\big(\mathbf{T} + \mathbf{S}^\top \boldsymbol{\beta}\big). \tag{S2}$$

Were $\mathbf{T}$ absent, the solution would be trivially $\boldsymbol{\beta}_s = s\,\boldsymbol{\beta}$. The template introduces an additive offset that must be absorbed into $\boldsymbol{\beta}$-space.

To this end, we project $\mathbf{T}$ into shape space via least squares:

$$\mathbf{c} = \arg\min_{\mathbf{c}} \|\mathbf{S}^\top \mathbf{c} - \mathbf{T}\|_2^2 = (\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{T}, \tag{S3}$$

Fig. S2: Qualitative comparison with Human3R and UniSH. Each row shows the scene image, followed by the reconstructions produced by each method. GRAFT (right) achieves tighter foot–ground contact and more accurate global placement, particularly in cluttered or geometrically complex environments.

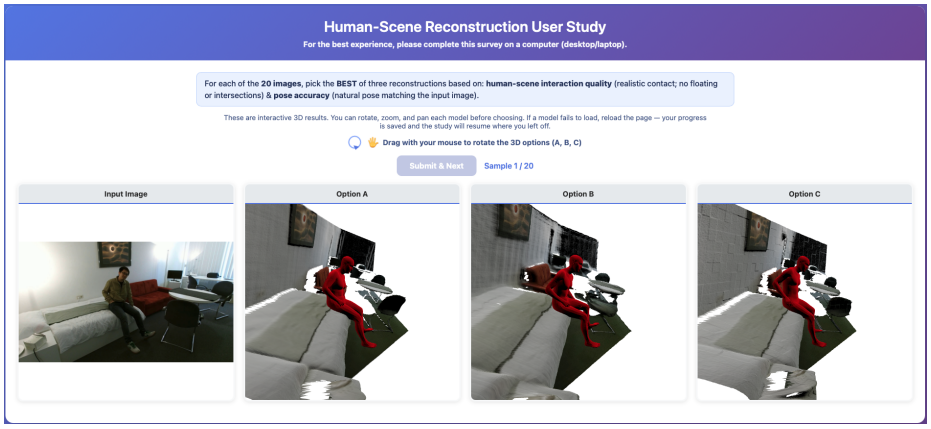

Fig. S3: User study interface. The input image (left) is shown alongside three interactive 3D reconstructions that participants freely rotate before selecting the most plausible human–scene interaction.

so that $\mathbf{T} \approx \mathbf{S}^\top \mathbf{c}$. Substituting this approximation into Eq. (S2):

$$\mathbf{S}^\top \mathbf{c} + \mathbf{S}^\top \boldsymbol{\beta}_s = s\big(\mathbf{S}^\top \mathbf{c} + \mathbf{S}^\top \boldsymbol{\beta}\big). \tag{S4}$$

Factoring out $\mathbf{S}^\top$ on both sides and cancelling it (as $\mathbf{S}$ has full row rank), we obtain:

$$\boldsymbol{\beta}_s = s \cdot (\boldsymbol{\beta} + \mathbf{c}) - \mathbf{c}. \tag{S5}$$

This is a closed-form, fully differentiable update: $\mathbf{c}$ is computed offline once per body model, translations scale directly, and rotations are unchanged. It replaces expensive per-step shape refitting with simple vector arithmetic, enabling efficient multi-step rollout supervision.

S6 Interaction Metric Details: V2S and D2S

We introduce two geometry-continuous metrics to complement the hard-threshold contact F1 score. Both are computed over the PROX contact-vertex set: approximately 1000 vertices distributed across body regions commonly involved in human–scene contact, including the hands, feet, back, and thighs.

Vector-to-Scene (V2S). For each contact vertex v in the set, we compute the 3D displacement vector $\mathbf{d}(v) = \mathbf{p}_v^* - \mathbf{p}_v$ from the vertex position to its nearest scene point, both in the prediction and in the ground truth. V2S reports the weighted mean Euclidean error between these predicted and ground-truth displacement vectors (in mm):

$$\text{V2S} = \frac{\sum_v w_v \, \|\mathbf{d}^{\text{pred}}(v) - \mathbf{d}^{\text{gt}}(v)\|_2}{\sum_v w_v}. \tag{S6}$$

Algorithm S1 GRAFT inference

Require: Scene image $\mathcal{I}_s$, human image $\mathcal{I}_h$, refinement steps T
Ensure: Scene pointmap $\mathcal{P}_s$, refined human parameters $\{\mathbf{\Theta}_i^T\}_{i=1}^{N_h}$

$(\mathcal{P}_s, \mathcal{P}_h, \mathcal{F}_s, \mathcal{F}_h) \leftarrow \texttt{MapAnything}(\mathcal{I}_s, \mathcal{I}_h)$
$\{\mathbf{\Theta}_i^{\text{init}}\}_{i=1}^{N_h} \leftarrow \texttt{NLF}(\mathcal{I}_h)$
for $i = 1$ to N_h **do**
  $\mathbf{p}_i^{\text{scene}} \leftarrow \mathcal{P}_h\big[\pi(\mathbf{p}_i^{\text{head}})\big]$ ▷ 3D scene point at projected head location
  $s_i^0 \leftarrow p_{i,z}^{\text{scene}} / p_{i,z}^{\text{head}}$
  $\mathbf{\Theta}_i^0 \leftarrow \texttt{MetricAlign}(\mathbf{\Theta}_i^{\text{init}}, s_i^0)$
for $t = 0$ to $T-1$ **do**
  for $i = 1$ to N_h **do**
    $\mathbf{g}_{i,k}^t \leftarrow \texttt{GeometricProbe}_k(\mathcal{P}_s, \mathbf{\Theta}_i^t), \quad k = 1, \ldots, 24$
    $\mathbf{z}_{i,k}^{t,0} \leftarrow \phi_k([\mathbf{g}_{i,k}^t; \mathbf{p}_{i,k}^t]), \quad \mathbf{Z}_i^{t,0} = \{\mathbf{z}_{i,k}^{t,0}\}_{k=1}^{24}$
    $\mathbf{V}_i^t \leftarrow \{\mathbf{v}_{i,k}^t\}_{k=1}^{24} \leftarrow \texttt{VisualAnchors}(\mathcal{F}_s, \mathcal{F}_h, \mathbf{\Theta}_i^t)$
    for $\ell = 1$ to 5 **do**
      $\widetilde{\mathbf{Z}}_i^{t,\ell} \leftarrow \texttt{SelfAttn}_\ell(\mathbf{Z}_i^{t,\ell-1})$
      $\mathbf{Z}_i^{t,\ell} \leftarrow \texttt{GeomCrossAttn}_\ell(\widetilde{\mathbf{Z}}_i^{t,\ell}, \mathbf{V}_i^t)$
    $(\Delta\mathbf{\Theta}_i^t, s_i^t) \leftarrow \Psi(\mathbf{Z}_i^{t,5})$ ▷ interaction gradient
    $\mathbf{\Theta}_i^{t+1} \leftarrow \mathbf{\Theta}_i^t + \Delta\mathbf{\Theta}_i^t$
    $\mathcal{M}_i^{t+1} \leftarrow s_i^t \ \text{SMPLX}(\mathbf{\Theta}_i^{t+1})$
return $\mathcal{P}_s$ and $\{\mathbf{\Theta}_i^T\}_{i=1}^{N_h}$

Direction-to-Scene (D2S). D2S reports the weighted mean angular error between the predicted and ground-truth displacement vectors (in degrees):

$$\text{D2S} = \frac{\sum_v w_v \angle\big(\mathbf{d}^{\text{pred}}(v),\, \mathbf{d}^{\text{gt}}(v)\big)}{\sum_v w_v}. \tag{S7}$$

Together, V2S captures errors in the magnitude and direction of human–scene proximity, while D2S isolates directional misalignment independently of scale.

Density-aware vertex weighting. The SMPL-X mesh is non-uniform: hand regions contain far more vertices than coarser regions such as the thigh. Without correction, high-density regions would dominate both metrics regardless of physical significance. Each vertex v is therefore weighted by its Voronoi area on the mesh surface, so that dense regions (e.g. fingers) contribute less and coarser regions (e.g. thigh) contribute more, making the metrics independent of tessellation density.

S7 Implementation Details

Learning-rate schedule. We train with a batch size of 16. Linear warmup runs for the first 2k iterations (from 0 to 1×10^{-4}), followed by cosine decay to 2×10^{-7} over the remaining 148k iterations.

Loss weights. Camera-relative vertex loss weight $\lambda_v = 7.0$; centered-vertex loss weight $\lambda_n = 5.0$. Rotation weights per parameter group: global orientation 5.0,

Algorithm S2 GRAFT training

Require: Batch $\mathcal{B}$, iteration k, rollout horizon T=3
Ensure: Updated GRAFT weights
$(\mathcal{I}_s, \mathcal{I}_h, \mathcal{P}_s, \mathcal{F}_s, \mathcal{F}_h) \leftarrow \texttt{PrepareBatch}(\mathcal{B})$
$(\boldsymbol{\Theta}^*, \mathcal{V}^*, \widetilde{\mathcal{V}}^*) \leftarrow \texttt{PrepareGT}(\mathcal{B})$
$\boldsymbol{\Theta}^{\text{nlf}} \leftarrow \texttt{NLFInit}(\mathcal{B})$
Sample $\boldsymbol{\Theta}^0$ from $\{\boldsymbol{\Theta}^{\text{nlf}},\ \boldsymbol{\Theta}^*,\ \boldsymbol{\Theta}^* + \epsilon\},\ \epsilon \sim \mathcal{N}(0, \Sigma)$ ▷ NLF / GT / perturbed GT
Apply visual-anchor dropout to sampled anchor neighbourhoods
$R \leftarrow 1$ if $k < 10$k, else $R \leftarrow T$
$\mathcal{L} \leftarrow 0$
for $t = 0$ to $R - 1$ **do**
  $(\Delta\boldsymbol{\Theta}^t, s^t) \leftarrow GRAFT_\omega(\mathcal{P}_s, \mathcal{F}_s, \mathcal{F}_h, \boldsymbol{\Theta}^t)$ ▷ shared weights over t
  $\boldsymbol{\Theta}^{t+1} \leftarrow \boldsymbol{\Theta}^t + \Delta\boldsymbol{\Theta}^t$
  $\mathcal{V}^{t+1} \leftarrow s^t\ \text{SMPLX}(\boldsymbol{\Theta}^{t+1})$
  $\widetilde{\mathcal{V}}^{t+1} \leftarrow \mathcal{V}^{t+1} - \text{mean}(\mathcal{V}^{t+1})$
  $\mathcal{L} \leftarrow \mathcal{L} + \lambda_p \|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^*\|_2^2 + \lambda_v \|\mathcal{V}^{t+1} - \mathcal{V}^*\|_2^2 + \lambda_n \|\widetilde{\mathcal{V}}^{t+1} - \widetilde{\mathcal{V}}^*\|_2^2$
$\omega \leftarrow \texttt{Adam}(\omega, \nabla_\omega \mathcal{L})$
Update learning rate by linear warmup followed by cosine decay

body pose 2.0, left-hand pose 0.5, right-hand pose 0.5. Translation and body shape are supervised implicitly through the vertex losses. No explicit contact or interpenetration losses are used.

Geometric feature encoding. Each geometric probe yields three 3D quantities: the displacement to the nearest scene point, the surface normal at that point, and the probe position in body-relative coordinates. Each is encoded by a separate learnable Fourier feature encoder (64 output dimensions, with the raw input concatenated), and the results are concatenated with the current joint state before being lifted to the transformer dimension by a two-layer MLP. Hand and full-body tokens additionally compress their per-probe features through a shared small MLP before concatenation.

Model architecture. GRAFT uses a transformer width of 512 with 5 layers and 8 attention heads, totalling 16.2M parameters across transformer layers, tokenizer MLPs, and decoder heads—a deliberately lightweight design. Visual features are sampled from four levels of MapAnything: the post-ViT output, two intermediate activations within the alternating transformer, and its final output. At each spatial position in the sampled neighbourhood, the four level features are each linearly projected to 128 dimensions, concatenated, and passed through a 512→512 linear layer, yielding one 512-dimensional token per position. For body and hand anchors we sample a 3×3 neighbourhood, producing 9 tokens per stream; for the 27 full-body surface anchors we sample a single token each (1×1). Features are extracted from two streams—the scene image $\mathcal{I}_s$ and the interaction image $\mathcal{I}_h$—so each HSI token cross-attends to $2n^2$ context tokens in total. A learnable *stream embedding* (one per stream, added after the per-level fusion) lets the model distinguish scene from interaction evidence. All decoder heads Ψ

are two-layer MLPs with hidden dimension 256, and all MLPs throughout the model use GELU activations.

References

1. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion (2023) 9
2. Chen, Y., Chen, X., Xue, Y., Chen, A., Xiu, Y., Gerard, P.M.: Human3r: Everyone everywhere all at once. arXiv preprint arXiv:2510.06219 (2025) 2, 3, 4, 10, 12
3. Corona, E., Pons-Moll, G., Alenyà, G., Moreno-Noguer, F.: Learned Vertex Descent: A New Direction for 3D Human Model Fitting (Jul 2022) 4
4. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D Human Pose Ambiguities with 3D Scene Constraints (Aug 2019) 3, 9, 11, 12
5. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D Scenes by Learning Human-Scene Interaction (Apr 2021) 4, 6
6. He, Y., Tiwari, G., Birdal, T., Lenssen, J.E., Pons-Moll, G.: Nrdf: Neural riemannian distance fields for learning articulated pose priors. In: Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024) 4
7. 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: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 13274–13285 (Jun 2022) 9
8. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P.: MapAnything: Universal feed-forward metric 3D reconstruction. In: International Conference on 3D Vision (3DV). IEEE (2026) 5, 6
9. Kister, N., YM, P., Sárándi, I., Wang, J., Khoreva, A., Pons-Moll, G.: Inhabit: Leveraging image foundation models for scalable 3d human placement. https://virtualhumans.mpi-inf.mpg.de/inhabit/ (2026), project website 9
10. Li, M., Li, P., Zhang, Z., Lu, J., Zhao, C., Xue, W., Liu, Q., Peng, S., Zhang, W., Luo, W., Liu, Y., Guo, Y.: Unish: Unifying scene and human reconstruction in a feed-forward pass (2026), https://arxiv.org/abs/2601.01222 2, 3, 4, 10, 12
11. Li, Y., Si, S., Li, G., Hsieh, C.J., Bengio, S.: Learnable fourier features for multi-dimensional spatial positional encoding (2021), https://arxiv.org/abs/2106.02795 6
12. Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos (2024) 5
13. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views (2025) 5
14. Liu, Z., Lin, J., Wu, W., Zhou, B.: Joint optimization for 4d human-scene reconstruction in the wild (2025), https://arxiv.org/abs/2501.02158 2, 3
15. Müller, L., Choi, H., Zhang, A., Yi, B., Malik, J., Kanazawa, A.: Reconstructing people, places, and cameras. arXiv:2412.17806 (2024) 2, 3
16. Patel, P., Black, M.J.: CameraHMR: Aligning People with Perspective (Nov 2024) 3
17. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2019) 4
18. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild (Mar 2025) 4

19. Prokudin, S., Lassner, C., Romero, J.: Efficient learning on point clouds with basis point sets (2019) 4, 6
20. Sárándi, I., Pons-Moll, G.: Neural localizer fields for continuous 3d human pose and shape estimation. In: Advances in Neural Information Processing Systems (NeurIPS) (2024) 3, 5, 6
21. Teed, Z., Deng, J.: RAFT: Recurrent All-Pairs Field Transforms for Optical Flow (Aug 2020) 4
22. Teed, Z., Deng, J.: DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras (Feb 2022) 4
23. Tiwari, G., Antic, D., Lenssen, J.E., Sarafianos, N., Tung, T., Pons-Moll, G.: Pose-NDF: Modeling Human Pose Manifolds with Neural Distance Fields (Jul 2022) 4
24. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3D Perception Model with Persistent State (Jan 2025) 11
25. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details (2025), https://arxiv.org/abs/2507.02546 5
26. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: $π^3$: Permutation-Equivariant Visual Geometry Learning (Sep 2025) 11
27. Wang, Y., Daniilidis, K.: ReFit: Recurrent Fitting Network for 3D Human Recovery (Aug 2023) 4
28. Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes (2022) 9
29. Wei, R., Yin, Z., Zhang, S., Zhou, L., Wang, X., Ban, C., Cao, T., Sun, H., He, Z., Liang, K., Ma, Z.: Omnieraser: Remove objects and their effects in images with paired video-frame data. arXiv preprint arXiv:2501.07397 (2025), https://arxiv.org/abs/2501.07397 5
30. Weng, Z., Yeung, S.: Holistic 3D Human and Scene Mesh Estimation from Single View Images. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2021) 3, 12
31. Xu, H., Barath, D., Geiger, A., Pollefeys, M.: ReSplat: Learning Recurrent Gaussian Splats (Oct 2025) 4
32. Yalandur Muralidhar, P., Xue, Y., Xie, X., Kostyrko, M., Pons-Moll, G.: PhySIC: Physically Plausible 3D Human-Scene Interaction and Contact from a Single Image. In: SIGGRAPH Asia 2025 Conference Papers (2025) 2, 3, 5, 12
33. Yi, H., Huang, C.H.P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., Black, M.J.: Human-Aware Object Placement for Visual Environment Reconstruction. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2022) 3
34. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: PLACE: Proximity Learning of Articulation and Contact in 3D Environments (Nov 2020) 4, 6