Human Interaction-Aware 3D Reconstruction from a Single Image
Supplementary Material

Gwanghyun Kim[1‡] Junghun James Kim[2‡] Suh Yoon Jeon[1‡] Jason Park[1] Se Young Chun[1,2,3†]
[1]Dept. of Electrical and Computer Engineering, [2]IPAI, [3]INMC
Seoul National University, Republic of Korea
{gwang.kim, jonghean12, euniejeon, jsp11235, sychun}@snu.ac.kr

List of Contents

‡Authors contributed equally. †Corresponding author.

A. Details on Methods and Implementation

A.1. Robust Instance Segmentation and SMPL-X Estimation

Instance Segmentation. We adopt a hybrid instance segmentation pipeline that combines detection, pose estimation, and promptable segmentation to produce per-person masks from a single image. This process is designed to be occlusion-aware and ensures that each human instance is segmented consistently, serving as the foundation for downstream SMPL-X fitting.

We begin by applying YOLOv11 [10] to detect human instances in the input image, using only the "person" class to extract tight bounding boxes around each individual. For each detected bounding box, we then estimate 2D keypoints using ViTPose [24] generated from BUDDI [16] project. These keypoints provide reliable localization of body joints and are retained for subsequent matching and alignment purposes. To generate high-quality binary masks for each person, we pass the bounding boxes as prompts to the Segment Anything Model (SAM) [12], which produces accurate per-instance segmentations. These masks are then associated with individual people by solving a Hungarian assignment problem between keypoint-based anchor regions (e.g., head and feet) and the detected masks, ensuring proper instance-level alignment. To further improve segmentation quality, overlapping or duplicate detections are merged based on intersection-over-union (IoU) thresholds. Additionally, in cases of occlusion or ambiguous limb segmentation, we use the predicted keypoints along with SAM's region prompting to correct mismatched or missing parts, particularly in the hand and foot regions. This pipeline enables robust and scalable segmentation of multiple humans in a single view, even in the presence of occlusion or complex poses.

SMPL-X Estimation (RoBUDDI). We adopt BUDDI [16], a diffusion-based prior model, to estimate SMPL-X parameters for multi-human scenes. While BUDDI produces high-quality predictions for individual subjects, it exhibits limited effectiveness in handling collisions and interpenetrations. This is primarily because the penetration constraints are applied only in a second-stage refinement, after collisions have already occurred. As a result, scenes with dense interactions still suffer from body interpenetrations and inaccurate keypoint estimations in occluded regions.

To overcome these limitations, we introduce two physics-inspired supervision terms during optimization: (1) an interpenetration loss that penalizes body collisions between interacting subjects, and (2) a visibility-aware keypoint loss that reduces errors in self- and inter-human occluded areas. These additions enable more robust and physically plausible multi-human pose estimation. We refer to our enhanced approach as RoBUDDI, which integrates these geometry-level constraints into the fitting process. As shown in Tab. S7 and Fig. S18, RoBUDDI achieves both quantitatively superior accuracy and qualitatively improved physical realism compared to the baseline.

The interpenetration loss penalizes unrealistic mesh overlaps between specific body part pairs. We define a set V of tolerance pairs derived from contact regions in the initial SMPL-X meshes (e.g., left thigh vs. right calf). For each pair $(i, j) \in \mathrm{V}$, the nearest surface points are denoted by $s_1^{i,j}$ and $s_2^{i,j}$, and we use their separation $|s_1^{i,j} - s_2^{i,j}|$ in the loss. We apply a soft constraint around a threshold tol, with $T = \max(0.25\,\mathrm{tol}, 10^{-5})$ acting as a smoothing temperature that controls the softness of the penalty:

$$\mathcal{L}_{\mathrm{pen}} = \gamma_{\mathrm{pen}} \cdot \mathrm{mean}_{\mathrm{V}}\Big[T\,\ln\big(1 + e^{(\mathrm{tol}-|s_1^{i,j}-s_2^{i,j}|)/T}\big)\Big]. \quad \text{(S1)}$$

where γ_{pen} sets the overall strength of the term, and T preserves a smooth transition around while enforcing a strong penalty when $|s_1^{i,j} - s_2^{i,j}| < \mathrm{tol}$.

In addition, we introduce a visibility-aware keypoint loss that adaptively downweights occluded joints. Given instance segmentation masks, each projected joint j is assigned:

$$w_j = \begin{cases} 1, & \text{if joint } j \text{ is visible,} \\ \alpha_{\mathrm{occ}}, & \text{if joint } j \text{ is occluded } (\alpha_{\mathrm{occ}} = 0.1). \end{cases}$$

Let $\{\mathbf{u}_j\}$ and $\{\hat{\mathbf{u}}_j\}$ be the ground-truth and estimated 2D joint positions. We define

$$\mathcal{L}_{\mathrm{std}} = \frac{1}{N}\sum_{j=1}^{N} \|\mathbf{u}_j - \hat{\mathbf{u}}_j\|^2, \qquad \mathcal{L}_{\mathrm{vis}} = \frac{1}{N}\sum_{j=1}^{N} w_j\,\|\mathbf{u}_j - \hat{\mathbf{u}}_j\|^2.$$

Here, N denotes the total number of 2D keypoints used in the reprojection loss. We then combine them as below.

$$\mathcal{L}_{\mathrm{kp}} = \gamma_{\mathrm{std}}\,\mathcal{L}_{\mathrm{std}} \;+\; \gamma_{\mathrm{vis}}\,\mathcal{L}_{\mathrm{vis}}$$

Although conceptually similar to the visibility loss used in mesh reconstruction, where misclassified silhouette pixels are penalized, this formulation operates on joint reprojection errors rather than mask-pixel discrepancies, yielding improved robustness to occlusion in keypoint alignment.

RoBUDDI estimates each person's 3D pose in multi-person scenarios, relative to the camera, providing all necessary geometric information for canonicalization. This effectively sets the camera's extrinsic rotation to I and translation to $\vec{0}$.

The original BUDDI optimization takes approximately 60s per image and peaks at 12.48GB of GPU memory on an NVIDIA RTX A6000 GPU (batch size=1). Adding our interpenetration and visibility penalties increases runtime to 77s and peak memory to 15.16GB. All experiments use an interpenetration threshold $tol = 0.02$, a penetration loss weight $\gamma_{\mathrm{pen}} = 15$, an occlusion weight $\alpha_{\mathrm{occ}} = 0.1$, and visibility-aware keypoints and keypoint loss blend factors $\gamma_{\mathrm{std}} = \gamma_{\mathrm{vis}} = 0.5$.

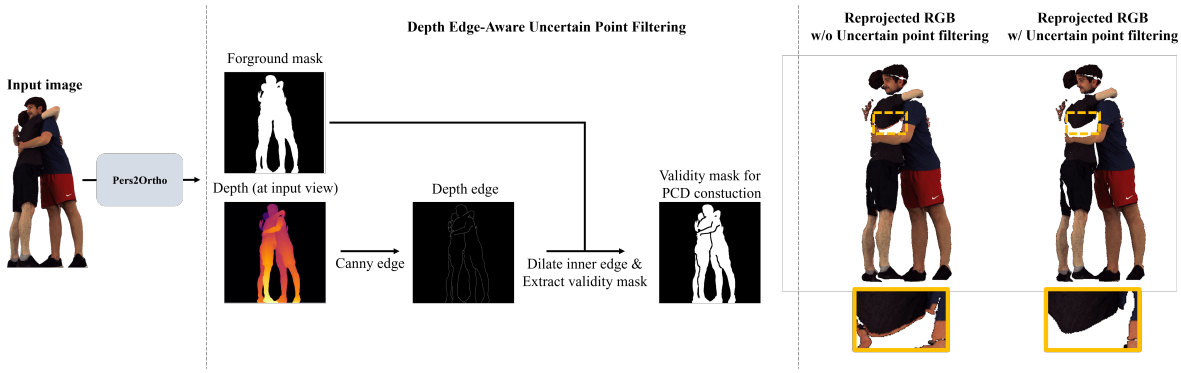


Figure S1. Depth edge-aware filtering removes uncertain boundary regions to improve orthographic projection stability.

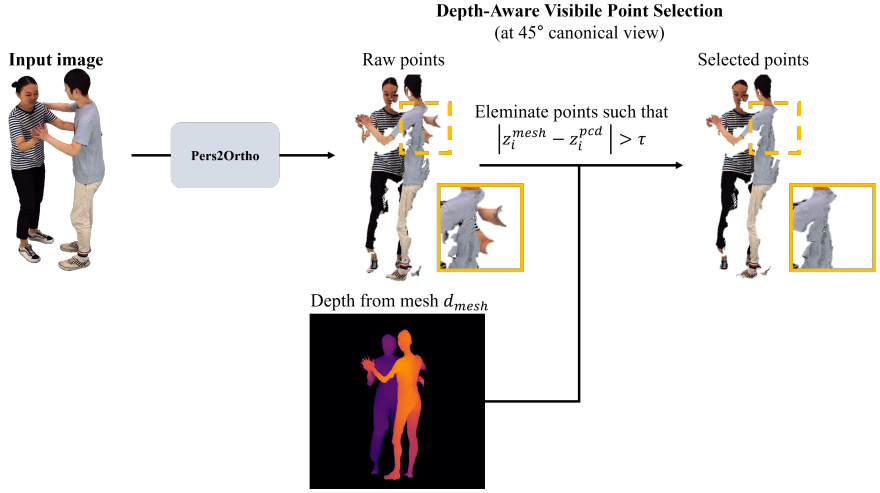


Figure S2. Depth-aware filtering selects front-visible points for clean orthographic projections.

A.2. Canonical Perspective-to-Orthographic View Transform (Pers2Ortho)

In addition to the primary transformation pipeline described in the main paper, we detail two depth-aware filtering strategies designed to suppress projection artifacts during the conversion from perspective to orthographic views.

Depth Edge-Aware Uncertain Point Filtering. As shown in Fig. S1, depth discontinuities often lead to jagged contours and ghosting artifacts near object boundaries. To address this, we detect depth edges using Canny edge detection applied to the rendered depth map. To reduce spurious detections near image borders, we erode the foreground mask before edge extraction. The resulting edge map is then dilated to encompass uncertain boundary regions. A refined validity mask is constructed by excluding these edge-dilated areas from the projection domain, ensuring that only stable, interior pixels are used in the orthographic projection.

Depth-Aware Visible Point Selection. As represented in Fig. S2, to preserve geometric consistency during the projection of partial point clouds (PCD) into orthographic views, we filter for front-visible points using the rendered depth from the mesh as a geometric prior. This strategy eliminates occluded or background points, retaining only those lying in front of the mesh surface and visible from the target camera view.

Specifically, we project 3D world-space points onto the image plane and sample the mesh depth at corresponding pixel locations. A point is retained if: (1) its projected 2D coordinate lies within image bounds, (2) the absolute depth difference between the point and the mesh is below a threshold τ (set to $\tau = 0.02$), and (3) the sampled mesh depth is positive.

Formally, let $p_i \in \mathbb{R}^3$ be a 3D point from the PCD, and z_i^{pcd} and z_i^{mesh} denote the depths from the point cloud and mesh at the projected location, respectively. The point is retained if:

$$|z_i^{\text{mesh}} - z_i^{\text{pcd}}| < \tau, \quad \text{and} \quad z_i^{\text{mesh}} > 0 \tag{S2}$$

This depth-aware visibility filtering yields cleaner foreground silhouettes and reduces projection noise by removing points that are geometrically inconsistent or lie behind the mesh surface.

We optimize the mesh for partial 3D reconstruction over 200 iterations with a learning rate of 0.02. The Pers2Ortho

module, including the reprojection step, takes 16.20 seconds and consumes 14.4GB of VRAM on an NVIDIA A100. We apply a dilation operation with a kernel size of 5 for depth edges.

A.3. Human Group-Instance Multi-View Diffusion (HUG-MVD)

A.3.1. Training Datasets

We leverage one multi-human dataset [25] and two single-human datasets [5, 26] to supervise our model with diverse human poses, interactions, and appearances.

Hi4D [25] is a novel dataset targeting close-range, prolonged human-human interactions with physical contact. Capturing and disentangling such interactions is particularly challenging due to severe occlusions and topological ambiguities. To address this, Hi4D employs individually fitted neural implicit avatars and an alternating optimization scheme that jointly refines surface and pose during close contact. This enables automatic segmentation of fused 4D scans into individual humans. The dataset comprises 100 sequences across 20 subject pairs, totaling over 11K textured 4D scans, all annotated with accurate 2D/3D contact labels and registered SMPL-X models. For our experiments, we extract 1,272 scenes by sampling contact frames with a stride of 16.

CustomHumans [5] contains high-quality static scans of 80 individuals, captured using a multi-view photogrammetry system with 53 RGB (12MP) and 53 IR (4MP) cameras. Each subject performs a set of predefined motions, including T-pose, hand gestures, and squats, in 10-second sequences at 30 FPS. From each sequence, 4–5 high-fidelity frames are selected, yielding over 600 3D scans. Each sample includes a 40K-face mesh, a 4K texture map, and accurately registered SMPL-X parameters, with a wide range of garment styles (120 total).

THuman2.0 [26] offers 500 high-resolution human scans captured using a dense DSLR rig. Each sample consists of a detailed 3D mesh paired with a high-quality texture map, covering a wide range of body shapes and clothing types. This dataset serves as a clean source of diverse clothed human geometry.

A.3.2. Rendering Procedure

For each 3D human scene, we render 16 views in total, comprising 8 orthographic and 8 perspective images. Each view includes RGB images, normal maps, depth maps, and segmentation masks, rendered from both SMPL-X meshes and original scanned meshes. All views share the same azimuth angles to allow consistent comparison across projection types. Orthographic views are rendered with slight variation in elevation (randomly sampled in the range $[-10°, +10°]$) to improve robustness against vertical pose and camera variations. Perspective views, on the other hand, utilize elevation values randomly sampled from a broader range of $[-20°, +45°]$ and distances between 2.0 and 6.0 units to enhance variability and generalization.

Camera extrinsic parameters (rotation $\mathbf{R}$ and translation $\mathbf{T}$) are generated using a virtual camera located at the specified distance and orientation, always looking at the origin. These parameters are derived via the `look_at_view_transform` function in PyTorch3D [18].

Prior to rendering, all meshes are normalized to a canonical space. This is done by computing the bounding box of the vertex positions, determining the center and maximal axis length, and scaling the mesh such that it fits within a unit cube. For datasets like CustomHumans [5] and THuman2.0 [26], we additionally introduce random jittering to the mesh center (on the X and Y axes) to prevent overfitting and encourage generalization. The normalized vertex positions $\mathbf{v}_{\text{norm}}$ are computed by:

$$\mathbf{v}_{\text{norm}} = \frac{\mathbf{v} - \mathbf{c}}{s/2}$$

where $\mathbf{c}$ is the computed center of the bounding box and s is the padded bounding box size.

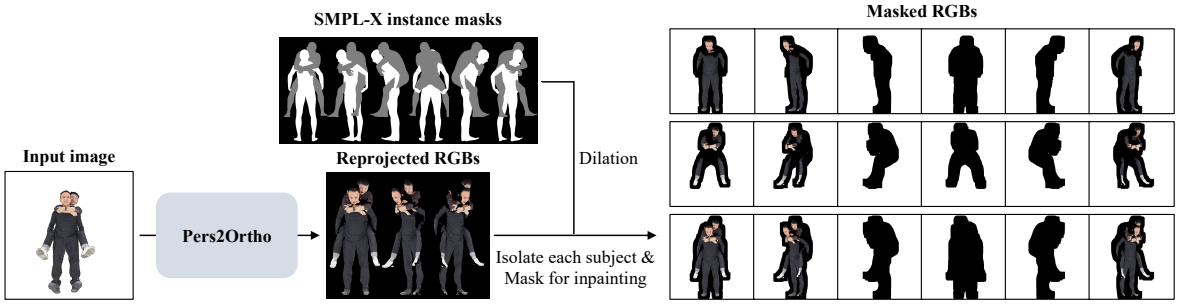
For orthographic rendering, we fix the camera distance at 3.0 units and maintain consistent scale across all views. For perspective rendering, focal lengths are automatically determined based on the normalized mesh size and camera distance, with additional jitter applied up to 20% to simulate realistic monocular variation.

Rendering is performed using the PyTorch3D `MeshRenderer`, configured with either orthographic or perspective camera models. RGB images are rendered using a Phong shading model with ambient or point lighting depending on the dataset. Depth maps are extracted from the rasterizer's z-buffer. Normal maps are generated by interpolating face vertex normals in view space. Instance segmentation masks are computed per-pixel using face-to-instance ID mappings. When contact information is available, contact masks are also rendered for the case of multi-human dataset by identifying faces associated with contact regions and projecting them to image space. Small holes in the resulting binary masks are filled using post-processing.

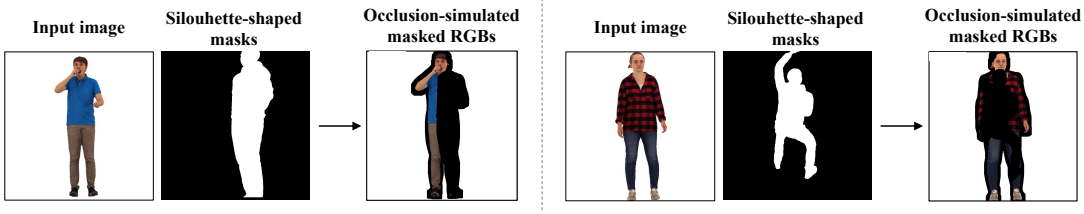
All outputs are rendered at a base resolution of 768×768 pixels. For training, we randomly select a reference view and sample six additional views at fixed relative azimuth angles of $\{0°, 45°, 90°, 180°, 270°, 315°\}$, resulting in a 6-view training input for each instance.

A.3.3. Masking Strategy for Occlusion Simulation

As shown in S3, to simulate realistic visibility and occlusion in multi-human 3D scenes, we construct masked images $x^{(i)}_{\text{mask}}$ from reprojected point clouds $x^{(i)}_{\text{pcd}}$ and visibility



(a) Masking strategy for multi-subject datasets



(b) Masking strategy for single-subject datasets

Figure S3. Illustration of masking strategy for occlusion simulation. (a) For multi-subject datasets, SMPL-X instance masks are used to isolate each subject and specify regions to inpaint. (b) For single-subject datasets, we simulate oclusion through silhouette-shaped masks.

masks M_{vis}, which indicate regions of missing observation in each canonical view $\mathcal{C}_i$.

To define the canonical visibility views, we use the reprojected RGB images captured at $\{0°, 45°, 315°\}$. For all other viewpoints, only the SMPL-X instance masks are used for occlusion simulation without relying on PCD.

During training, we generate masked RGB images $\{x^{(i)}_{\text{mask}}\}^5_{i=0}$ to guide inpainting or occlusion-aware reconstruction networks: background regions are labeled as 1, while occluded or missing areas are labeled as 0 and the visible region with the pixel values. These masks provide supervision for learning to reconstruct or inpaint plausible content in the occluded regions.

In multi-subject datasets such as Hi4D [25], where mutual occlusion naturally occurs, we use SMPL-X instance masks to isolate each subject. These masks are further dilated to account for peripheral structures such as garments and hair, ensuring that edge regions around the human silhouette are adequately covered for inpainting tasks. We apply a dilation operation with a kernel size of 61.

In single-subject datasets such as CustomHumans [5] and THuman2.0 [26], we simulate occlusion by randomly selecting one of two masking strategies with equal probability (0.5): (1) silhouette-shaped masks that resemble human figures, randomly scaled and positioned near the image center to mimic the presence of an occluder, or (2) random hole-based masks, such as freeform or template-driven occlusions, which introduce unstructured masking artifacts. This augmentation scheme enables the model to generalize to a wide range of occlusion scenarios, even in the absence of multiple real subjects.

A.3.4. Training Procedure

The objective of our training procedure is to reconstruct complete RGB and normal images from partially visible, occluded inputs by leveraging a multi-view diffusion model. An overview of the training process is illustrated in Fig. S4. Each training sample consists of six canonical views per scene. The model receives the following inputs: (1) Occluded RGB images reprojected from point clouds using the masking strategies described earlier: $\{x^{(i)}_{\text{mask}}\}^5_{i=0}$, and (2) Corresponding SMPL-X normal maps providing geometric structure: $\{n^{(i)}_{\text{SMPLX}}\}^5_{i=0}$. The supervision targets are: (1) Ground-truth RGB images: $\{x^{(i)}\}^5_{i=0}$, and (2) Ground-truth normal maps: $\{n^{(i)}\}^5_{i=0}$.

We initialize the model from PSHuman [15], a pretrained diffusion model designed to synthesize six RGB and normal views from a single RGB image. Notably, PSHuman is trained exclusively on single-human datasets, including THuman2.0 [26] and CustomHumans [5], and the open-source version of PSHuman does not support SMPL-X conditioning. To overcome this limitation and extend the model to multi-human scenes with explicit geometry input, we integrate ControlNet [27] into the architecture. This enables the model to utilize SMPL-X normal maps as structural guidance during training.

Our model operates in the latent space defined by the variational autoencoder (VAE) from Stable Diffusion 2.1 [19]. Each ground-truth RGB image $x^{(i)}$ and normal map $n^{(i)}$ is encoded into latent variables $z^{(i)}_{\text{rgb}}$ and $z^{(i)}_{\text{normal}}$ us-

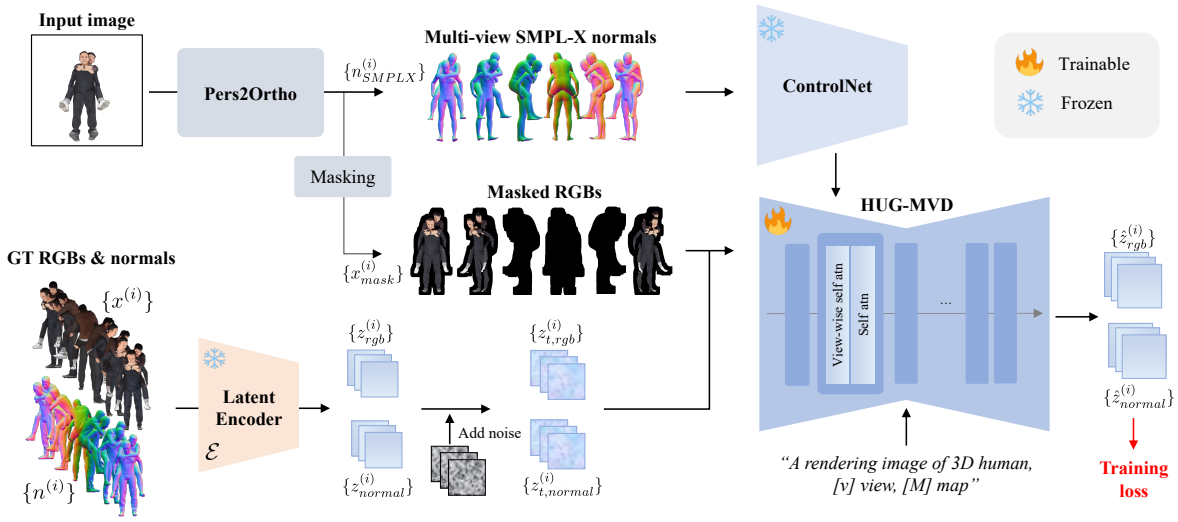


Figure S4. Illustration of the training procedure of HUG-MVD. The model takes occluded RGB images and SMPL-X normal maps from six views and learns to reconstruct complete RGB and normal views via a multi-view diffusion model. SMPL-X guidance is injected via ControlNet, with attention applied across views and modalities for coherent multi-human reconstruction.

ing the VAE encoder. The model then predicts the residual noise added to these latents during the forward diffusion process. The training objective is defined as:

$$\mathcal{L}_{\text{diff}} = \sum_{i=0}^{5} \Big(\mathbb{E}_{t,\epsilon} \Big[\| \epsilon - \epsilon_\theta(z^{(i)}_{t,\text{rgb}}, t, x^{(i)}_{\text{mask}}, n^{(i)}_{\text{SMPLX}}) \|_2^2 \Big] + \mathbb{E}_{t,\epsilon} \Big[\| \epsilon - \epsilon_\theta(z^{(i)}_{t,\text{normal}}, t, x^{(i)}_{\text{mask}}, n^{(i)}_{\text{SMPLX}}) \|_2^2 \Big] \Big) \tag{S3}$$

Here, $z^{(i)}_{t,\text{rgb}}$ and $z^{(i)}_{t,\text{normal}}$ denote the noisy latent variables at timestep t, and ϵ is the Gaussian noise sampled during training. The denoising model ϵ_θ learns to recover the clean signal conditioned on the masked RGB inputs and the SMPL-X geometry.

Two key attention modules are employed to support effective multi-view generation. First, following prior work [14, 15], we apply row-wise multi-view self-attention independently within each modality (RGB or normal), allowing the model to correlate across the six canonical views. Second, to enable information exchange between RGB and normal modalities, we apply self-attention across all latent tokens, allowing cross-modality attention between RGB and normal features at each view. This joint RGB-normal attention mechanism is also inspired by [14, 15].

During training, only the U-Net parameters are optimized, while the remaining components such as the CLIP image encoder and the VAE are kept frozen. Text prompt embeddings are injected via cross-attention using the template: ``a rendering image of 3D human, [V] view, [M] map'', where [V] indicates the view direction (e.g., `front`, `left`, `face`) and [M] specifies the modality (`color` or `normal`).

We also experimented with classifier-free guidance (CFG) by applying conditioning dropout, following prior work [7]. However, since CFG yielded only marginal performance improvements at test time, it was not used during inference.

To address occlusion and mesh collision issues in multi-human scenes, we optionally apply contact-aware binary masks on Hi4D [25] samples. These masks suppress gradient updates in regions of body-to-body contact where the mesh supervision is less reliable due to penetration artifacts or annotation noise.

The model is trained on a single NVIDIA A100 GPU (80GB) with a batch size of 16 and gradient accumulation steps of 8. We use the Adam [11] optimizer with a learning rate of 5×10^{-6}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Training follows a two-stage curriculum: (1) pre-training without inpainting masks for 1,000 steps, followed by (2) fine-tuning with inpainting masks for another 1,000 steps to simulate occlusion. The entire training takes approximately two days, with peak GPU memory usage of around 62GB. A DDPM scheduler with 1,000 diffusion steps is used throughout training.

A.3.5. Inference Procedure

At inference time, the model predicts complete normal maps $\{\hat{n}^{(i)}\}$ and synthesizes complete RGB images $\{\hat{x}^{(i)}\}$ across all views from partially visible RGB inputs and canonical-view SMPL-X normal maps as illustrated in Fig. S5.

To maintain consistency between group-level and instance-level reconstructions, we perform latent-space composition. At each diffusion timestep t, instance-specific latents $z^{(i)}_{t,\text{inst}(k)}$ are softly injected into the group-level la-

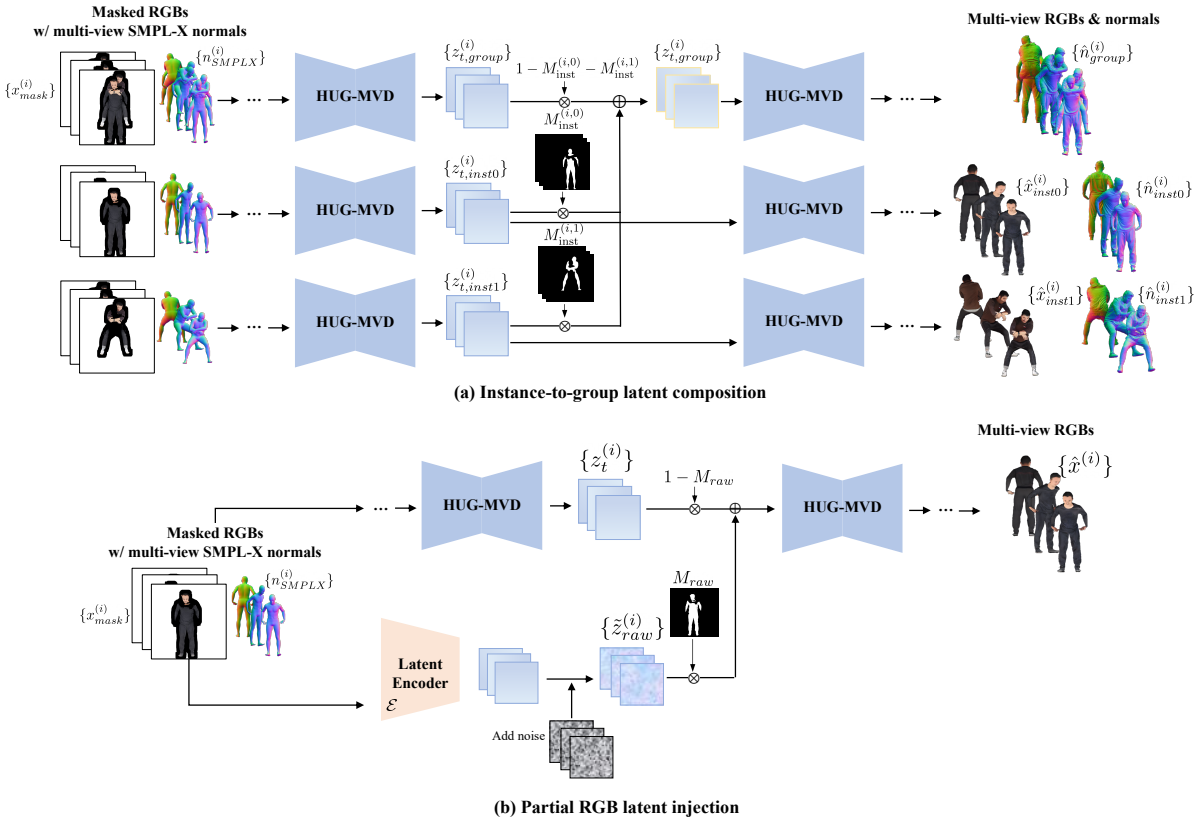


Figure S5. Illustration of the inference procedure of HUG-MVD. (a) Instance-to-group latent composition and (b) partial RGB latent injection for RGB synthesis.

tent $z^{(i)}_{t,\text{group}}$ using binary masks $\mathbb{1}^{(i,k)}_{\text{inst}}$ and a blending ratio $\alpha_{\text{gi}} = 0.8$:

$$z^{(i)}_{t,\text{group}} \leftarrow \sum_{k=1}^{K} \Big[\alpha_{\text{gi}} \cdot \mathbb{1}^{(i,k)}_{\text{inst}} \cdot z^{(i)}_{t,\text{inst}(k)} + (1-\alpha_{\text{gi}}) \cdot \mathbb{1}^{(i,k)}_{\text{inst}} \cdot z^{(i)}_{t,\text{group}}\Big] + \left(1 - \sum_{k=1}^{K} \mathbb{1}^{(i,k)}_{\text{inst}}\right) \cdot z^{(i)}_{t,\text{group}}, \tag{S4}$$

This mechanism allows high-frequency details from instance-specific predictions to be integrated into the global group representation, improving surface continuity in multi-human scenes.

Also, to further enhance RGB quality, we inject latent signals from partially visible RGB inputs. At each diffusion timestep t, we generate a noisy version $\tilde{z}_{\text{raw}}$ of the raw RGB latent (restricted to visible regions) and blend it into the current latent z_t using a binary mask $\mathbf{m}_{\text{raw}}$ and mixing ratio $\alpha_{\text{pcd}} = 0.8$:

$$z_t \leftarrow \mathbf{m}_{\text{raw}} \cdot \big[\alpha_{\text{pcd}} \cdot \tilde{z}_{\text{raw}} + (1-\alpha_{\text{pcd}}) \cdot z_t\big] + (1-\mathbf{m}_{\text{raw}}) \cdot z_t, \tag{S5}$$

This operation is applied exclusively to the RGB branch and aims to reinforce reliable visual priors in visible areas, improving fidelity in occluded or ambiguous regions. We apply this injection selectively to low-confidence views (e.g., non-source views) to avoid overwriting already plausible outputs.

We use a DDIM scheduler with $\eta = 1.0$ and perform 40 denoising steps per sample. We also used $\alpha_{\text{gi}} = \alpha_{\text{pcd}} = 0.8$. Inference for all group-level and instance-level multi-view RGB and normal maps takes 60.16 seconds, using 34.76GB of VRAM on an NVIDIA A100.

A.4. Human Group-Instance Geometry Reconstruction (HUG-GR)

Here, we provide a detailed explanation of the two geometry-level supervision terms—*interpenetration loss* and *visibility loss*. As illustrated in Fig. S6, these losses play complementary roles in enhancing geometric plausibility and part-level visibility consistency during group-instance reconstruction. Z

Interpenetration Loss. To prevent anatomically implausible overlaps between articulated body parts, we define an interpenetration loss that penalizes violations among predefined part pairs $(i, j) \in \text{V}$, where tol is the tolerance set en-

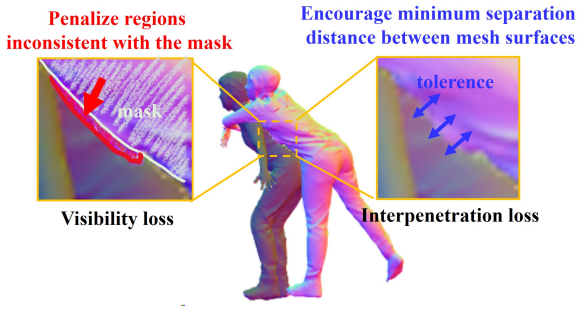


Figure S6. Illustration of the interpenetration loss and visibility loss. The interpenetration loss penalizes body-part collisions (blue arrows), while the visibility loss enforces alignment between rendered masks and ground-truth visibility (red region).

coding pairs subject to collision constraints inspired by [8]. To determine V, we first compute a contact map based on vertex-to-vertex distances on the initial SMPL-X meshes, which identifies all potential contact regions. The resulting contact regions constitutes our tolerance set. For each pair, let $s_1^{i,j}$ and $s_2^{i,j}$ be the closest surface points on parts i and j. We apply a smooth barrier around a tolerance tol (default $= 5 \times 10^{-4}$), with $T = \max(0.25\,\mathrm{tol}, 10^{-5})$ acting as a smoothing temperature:

$$\mathcal{L}_{\text{pen}} = \text{mean}_{(i,j)\in \text{V}} \Big[T \, \ln\big(1 + e^{(\text{tol} - |s_1^{i,j} - s_2^{i,j}|)/T}\big) \Big]. \quad \text{(S6)}$$

This term encourages a minimum surface separation between adjacent parts (e.g., thighs vs. calves), helping to reduce penetration artifacts while preserving flexibility for naturally close configurations such as seated or folded poses.

Visibility Loss. To improve spatial alignment in crowded scenes, we supervise visibility using rendered segmentation masks. For each body part b in instance k, we penalize visibility mismatches using:

$$\mathcal{L}_{\text{vis}} = \frac{1}{2B} \sum_{k=1}^{K} \sum_{b=1}^{B} \frac{E_b^k}{M_b^k + \epsilon}, \quad \text{(S7)}$$

where E_b^k is the number of incorrectly occluded pixels and M_b^k the total visible pixels in the ground truth. This encourages accurate silhouette and occlusion boundaries, particularly in group interactions.

Adaptive Region-Specific Optimization. To balance global stability and preservation of local details, we apply region-specific optimization strategies. In particular, lower learning rates are used for vertices located in semantically and geometrically complex regions such as the hands and face. This allows the model to better preserve high-frequency features provided in the initial SMPL-X mesh in these areas while still allowing flexible carving of geometric features, such as clothing, in other regions. We determine how close a vertice is to a complex region using the optimized SMPL-X joint positions of the hands and face. And use sigmoid blending to derive the actual learning rate. Formally, the vertice-wise adaptive learning rate α_{v} for vertice v is:

$$\alpha_{\text{v}} = \alpha_{\text{base}} \cdot \frac{1}{e^{-(200 d_v + 10)} + 1}, \quad \text{(S8)}$$

Where α_{base} is the base learning rate and d_v is the minimum distance between v and the set of all SMPL-X joint vertices in consideration denoted as $J_{\text{SMPL-X}}$. Then, d_v is:

$$d_v = \min_{j \in J_{\text{SMPL-X}}} \|v - j\| \quad \text{(S9)}$$

Thus, we assign lower learning rates to vertices with smaller d_v (i.e. vertices closer to hands or the face). As illustrated in Fig. 6(b), this adaptive strategy results in sharper reconstructions of fine regions (e.g., fingers, facial contours) while maintaining coherence in broader anatomical parts like the torso or limbs.

We optimize the mesh over 200 iterations with a learning rate of 0.01, $\lambda_{\text{group}} = 1.0$, $\lambda_{\text{inst}} = 0.2$, $\lambda_{\text{pen}} = 30.0$ and $\lambda_{\text{vis}} = 1.0$. HUG-GR takes 125.47 seconds and consumes 7.58GB of VRAM on an A100 GPU.

A.5. Occlusion- and View-Aware Texture Fusion

To construct coherent and high-quality full-body textures, we fuse multi-view RGB predictions into a unified texture. To improve texture fidelity and suppress artifacts, we introduce two important enhancements: *view-aware face restoration* and *occlusion-aware blending*.

View-Aware Face Restoration. Faces captured from extreme angles or under occlusion often exhibit degraded appearance. To address this, as shown in Fig. S7, we first analyze each view using facial landmarks and SMPL-X head orientation to estimate the relative frontalness of the face. Among the six RGB predictions per instance, we select the two most frontal views. If the source view is used for the face, we directly use its content. If not, we perform face inpainting on the most frontal view using CodeFormer [30], where a soft circular mask is generated using warped 5-point facial landmarks. In both cases, the enhanced face region is warped back and blended into the original view using inverse affine transformation. This step improves the final texture synthesis especially in the face region.

Occlusion-Aware Blending. As illustrated in Fig. S8, to prevent ghosting and bleeding artifacts near occlusion boundaries, we employ edge-aware confidence masking guided by view-dependent depth maps. Depth edges are first extracted using the Canny filter, and the resulting edge map is dilated with a fixed kernel to define an exclusion zone. We retain only the pixels that lie within foreground regions and are sufficiently distant from detected depth discontinuities. These reliable pixels are used to generate a binary confidence mask C_i for each view. The final contribution of a view's texture projection T_i is modulated by this mask as $T_i' = C_i \cdot T_i$. This occlusion-aware blending

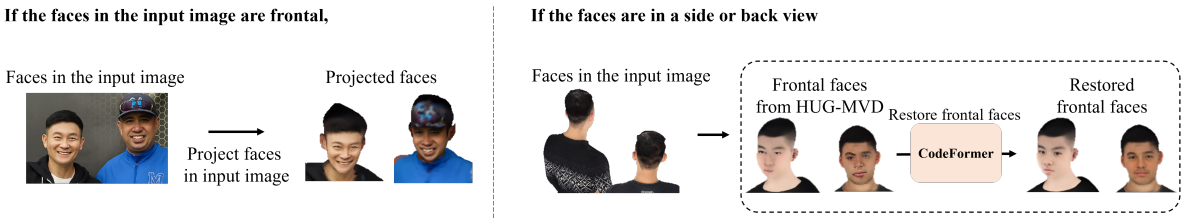


Figure S7. View-aware face restoration enhances frontal views using landmark-guided inpainting.

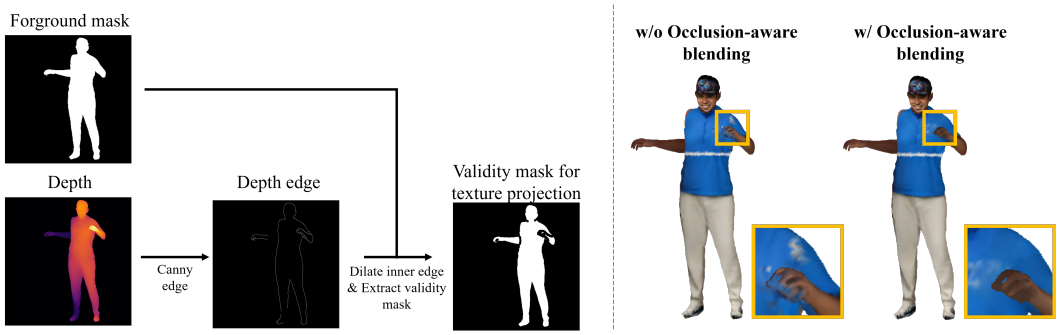


Figure S8. Illustration of the occlusion-aware blending strategy. Edge-aware confidence masks, computed from depth discontinuities, suppress artifacts near occlusion boundaries, resulting in cleaner silhouettes and improved cross-view consistency.

strategy effectively suppresses unstable regions near self-occlusion edges, resulting in cleaner object silhouettes and improved consistency across views. We apply a dilation operation with a kernel size of 21.

Our texture fusion takes 14.49 seconds and consumes 4.95GB of VRAM on an A100 GPU.

B. Evaluation Details

B.1. Evaluation Settings

We evaluate our method in comparison to prior works across three categories: methods of single human reconstruction from a single image, methods of multi-human reconstruction from multi-view images, and methods of multi-human reconstruction from videos. Since there is no publicly available baseline implementation that directly performs multi-human reconstruction from a single image as represented in Tab. S1, to ensure a fair comparison, we follow the evaluation protocol of [3] by adapting related methods in each category under consistent settings. To isolate the effect of SMPL-X prediction from the reconstruction process, all main comparisons are conducted using ground-truth SMPL-X. Results using predicted SMPL-X are included in Sec. C.1 of the supplementary.

Single human reconstruction from single image. For methods of originally designed for single human reconstruction [6, 15, 23, 28], we adapted them to the multi-human setting as follows. Ground-truth instance segmentation masks were used to isolate each person in the input image. Each individual was then reconstructed independently using the corresponding method. Since the outputs lie in different coordinate frames, we performed a canonicalization procedure to align all reconstructions into a shared space. Specifically, for each instance, we first predicted the SMPL-X mesh using the method's native estimator. We then computed a similarity transformation—comprising scale, rotation, and translation—that aligns the ground-truth SMPL-X mesh to the predicted one. The ground-truth SMPL-X mesh was transformed into the predicted space before reconstruction, and the reconstruction output was transformed back to the ground-truth space via the inverse transformation, allowing the reconstructed scene to be composited consistently. We also evaluate PSHuman-multi, which applies the single-person reconstruction pipeline PSHuman [15] directly to uncropped multi-person images. Since we use ground-truth SMPL-X for all evaluations, we omit the SMPL-X optimization process for baselines that originally involve it.

Multi-human reconstruction from multi-view image. For multi-view baselines [29], we provided only a single view as input for inference, to ensure comparability with our single-image reconstruction setting.

Multi-human reconstruction from videos. Similarly, for video-based baselines [9], we provided only a single image as the first frame for inference.

B.2. Evaluation Dataset

MultiHuman [9]. To facilitate both quantitative and qualitative evaluation of reconstructed meshes, we rendered perspective-view images from the MultiHuman dataset using a multi-view setup. Our evaluation covers a total of 20 two-person scenes, including 6 closely interactive cases (sequences 8, 23, 24, 250, 252, 253) and 14 naturally interactive scenes (sequences 12, 16, 17, 18, 19, 20, 22, 30, 226, 244, 249, 251, 255, 256). For ablation studies, we focus on the closely interactive cases.

For each scene, we rendered the meshes from 4 distinct camera viewpoints, generated by sampling a random azimuth and adding fixed offsets of $\{0°, 90°, 180°, 270°\}$, resulting in views uniformly distributed around the subject. The elevation angles were randomly sampled in the range $[-20°, 45°]$, and the camera-to-subject distances were sampled uniformly from $[2.0, 6.0]$, simulating varying levels of zoom and perspective distortion.

To ensure scale-invariant and consistently framed rendering, each mesh was normalized to fit within a unit cube centered at the origin. This was achieved by computing the mesh's axis-aligned bounding box and uniformly scaling it based on the maximum side length.

In-the-wild. For our qualitative evaluation with in-the-wild images, we leveraged OpenAI's Sora service to obtain a diverse set of test images. Sora performs a web-based image search for user-specified concepts, reconstructs novel scenes by referencing those search results, and synthesizes new images that reflect real-world variation. The resulting Sora outputs—whose content is derived from Internet-sourced photos—were then used as our "in-the-wild" evaluation set, ensuring that our method is tested on unconstrained, naturally diverse imagery.

B.3. Evaluation Metrics

We employ a comprehensive set of metrics to evaluate both the geometric and texture quality of reconstructed multi-human meshes. These metrics cover surface accuracy, physical realism, and perceptual quality. Here, P and Q are point clouds sampled from the predicted and ground-truth meshes.

Chamfer Distance (CD). Chamfer Distance measures the bidirectional discrepancy between the predicted and ground-truth surfaces. We uniformly sample 100,000 points from each mesh surface and compute the average closest-point distance from the predicted points to the ground-truth surface and vice versa. The final CD score is defined as:

$$\mathrm{CD}(P,Q) = \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2 + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|q-p\|_2.$$

A lower CD indicates a more accurate reconstruction that closely matches the geometry of the ground truth in both completeness and precision.

Point-to-Surface Distance (P2S). P2S measures the unidirectional accuracy of the predicted surface with respect to the ground-truth shape. Specifically, it measures the average Euclidean distance from each point sampled on the pre-

Table S1. Comparison of recent 3D human reconstruction methods. HUG3D supports multi-human reconstruction from a single image with both geometry and texture.

Method	Multi- or Single Human	Input Type	Geometry	Texture	Publicly Available
ECON [23]	Single human	Single image	✓	✗	✓
SiTH [6]	Single human	Single image	✓	✓	✓
SIFU [28]	Single human	Single image	✓	✓	✓
PSHuman [15]	Single human	Single image	✓	✓	✓
DeepMultiCAP [29]	Multi-human	Multi-view images	✓	✗	✓
Multiply [9]	Multi-human	Video	✓	✓	✓
Cha et al. [3]	Multi-human	Single image	✓	✗	✗
HUG3D (Ours)	Multi-human	Single image	✓	✓	✓ (upon acceptance)

dicted mesh to the closest point on the ground-truth surface:

$$\text{P2S}(P \to Q) = \frac{1}{|P|} \sum_{p \in P} \min_{q \in Q} \|p - q\|_2^2.$$

P2S emphasizes surface accuracy without penalizing missing parts, and lower values indicate closer alignment to the reference shape.

Normal Consistency (NC). NC measures the angular similarity between surface normals on the predicted and ground-truth meshes. For each point, we compare the normal vector at that point with the normal vector at the closest point on the opposite surface. The final score is averaged bidirectionally:

$$\text{NC}(P, Q) = \frac{1}{2|P|} \sum_{p \in P} \left(1 - \langle \mathbf{n}_p, \mathbf{n}_{\text{NN}(p,Q)} \rangle\right) + \frac{1}{2|Q|} \sum_{q \in Q} \left(1 - \langle \mathbf{n}_q, \mathbf{n}_{\text{NN}(q,P)} \rangle\right),$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product between unit normals, and $\text{NN}(\cdot)$ returns the nearest neighbor in the opposite set. A higher NC indicates better preservation of surface orientations and local detail.

F-score. F-score evaluates both precision and recall of the predicted surface points with respect to a ground-truth reference under a distance threshold τ. We use $\tau = 1$cm. Precision measures the percentage of predicted points that lie within τ of the ground-truth surface, while recall measures the converse. F-score is defined as the harmonic mean of the two:

$$\text{F-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

This metric rewards reconstructions that are both accurate and complete.

Bounding Box IoU (bbox-IoU). We compute the 3D Intersection-over-Union (IoU) of axis-aligned bounding boxes of the predicted and ground-truth meshes:

$$\text{IoU}_{\text{bbox}} = \frac{\text{vol}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{vol}(B_{\text{pred}} \cup B_{\text{gt}})},$$

where B_{pred} and B_{gt} are the predicted and ground-truth bounding boxes, respectively. This metric evaluates global layout similarity and spatial coverage.

L2 Normal Error. We assess surface detail preservation by computing the per-pixel L_2 distance between rendered normal maps of the predicted and ground-truth meshes. This is done across four orthographic views at azimuth angles $\{0°, 90°, 180°, 270°\}$:

$$\text{L2-NormErr} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{n}_i^{\text{pred}} - \mathbf{n}_i^{\text{gt}}\|_2^2.$$

We also report this error computed within occluded regions, to specifically assess reconstruction quality under visual occlusion.

Contact Precision (CP). To evaluate the physical plausibility of multi-human reconstruction, we measure the alignment of predicted inter-body contact regions with the ground truth. This metric quantifies how accurately the predicted contact points reflect the true contact between two human bodies.

Let $\hat{M}_1$ and $\hat{M}_2$ be the predicted meshes, and M_1 and M_2 the corresponding ground-truth meshes. Denote their vertex sets as $\hat{V}_1$, $\hat{V}_2$, V_1, and V_2, respectively. A vertex is considered in contact if it lies within a threshold distance δ from the other mesh.

First, we define the ground-truth contact region $\mathcal{C}_{\text{gt}}$ as:

$$\mathcal{C}_{\text{gt}} = \left\{ v \in V_1 \cup V_2 \;\middle|\; \min_{v' \in V_2 \cup V_1} \|v - v'\|_2 < \delta \right\},$$

and similarly, the predicted contact region $\mathcal{C}_{\text{pred}}$ as:

$$\mathcal{C}_{\text{pred}} = \left\{ \hat{v} \in \hat{V}_1 \cup \hat{V}_2 \;\middle|\; \min_{\hat{v}' \in \hat{V}_2 \cup \hat{V}_1} \|\hat{v} - \hat{v}'\|_2 < \delta \right\}.$$

We then compute precision by counting the proportion of predicted contact points that are close to the ground-truth contact region:

$$\text{CP} = \frac{1}{|\mathcal{C}_{\text{pred}}|} \sum_{\hat{v} \in \mathcal{C}_{\text{pred}}} \mathbf{1}\left[\text{NN}(\hat{v}, V_{\text{gt}}) \in \mathcal{C}_{\text{gt}}\right],$$

where $\text{NN}(\hat{v}, V_{\text{gt}})$ denotes the nearest vertex to $\hat{v}$ among all ground-truth vertices.

We set the contact threshold $\delta = 0.01$ meter. A higher CP indicates better prediction of physically plausible inter-human contacts.

Texture Fidelity. To assess the perceptual quality of the reconstructed texture, we evaluate the rendered mesh images against ground-truth renderings using three standard image similarity metrics: PSNR, SSIM, and LPIPS.

Given the predicted image $\hat{I}$ and the ground-truth image I rendered from the same view, we compute:

Peak Signal-to-Noise Ratio (PSNR):

$$\text{PSNR}(I, \hat{I}) = 10 \cdot \log_{10} \left(\frac{(L_{\max})^2}{\text{MSE}(I, \hat{I})} \right),$$

where $L_{\max} = 255$ and MSE denotes the mean squared error between pixel values. A higher PSNR indicates better reconstruction.

Structural Similarity Index Measure (SSIM).

$$\text{SSIM}(I, \hat{I}) = \frac{(2\mu_I \mu_{\hat{I}} + c_1)(2\sigma_{I\hat{I}} + c_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + c_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + c_2)},$$

where μ, σ^2, and $\sigma_{I\hat{I}}$ denote means, variances, and covariances of local patches. SSIM captures perceptual similarity in terms of luminance, contrast, and structure.

Learned Perceptual Image Patch Similarity (LPIPS). LPIPS compares features from a pretrained deep network (e.g., AlexNet) between $\hat{I}$ and I, and correlates better with human judgment of perceptual similarity. Lower LPIPS indicates better quality.

These metrics are computed over four rendered views $\{0°, 90°, 180°, 270°\}$, under orthographic projection. We also report masked versions of these metrics that are evaluated only on occluded foreground regions, allowing more fine-grained assessment under challenging interaction scenarios.

Occlusion-aware Metrics. To evaluate reconstruction quality under challenging visibility conditions, we compute occlusion-aware variants of image-based metrics and surface normal metrics by restricting the evaluation to regions occluded by another human instance.

Let $I^{(i)}$ and $\hat{I}^{(i)}$ denote the ground-truth and predicted images of instance $i \in \{0, 1\}$, and let $M^{(j)}$ be the binary mask of the other instance $j \neq i$. A pixel (x, y) is considered occluded in instance i if it belongs to $M^{(j)}$ and the corresponding pixel in $I^{(i)}$ is not background (i.e., not white):

$$\mathcal{O}^{(i)} = \left\{ (x, y) \,\middle|\, M^{(j)}(x, y) = 1 \ \wedge \ I^{(i)}(x, y) \neq \textit{background} \right\}.$$

We then compute each metric by applying the occlusion mask $\mathcal{O}^{(i)}$ to both predicted and ground-truth images:

$$\text{Occ-PSNR}^{(i)} = \text{PSNR}\left(\hat{I}^{(i)}|_{\mathcal{O}^{(i)}}, I^{(i)}|_{\mathcal{O}^{(i)}} \right),$$

$$\text{Occ-SSIM}^{(i)} = \text{SSIM}\left(\hat{I}^{(i)}|_{\mathcal{O}^{(i)}}, I^{(i)}|_{\mathcal{O}^{(i)}} \right)$$

For surface normal comparisons, let $N^{(i)}$ and $\hat{N}^{(i)}$ be the ground-truth and predicted normal maps of instance i. The occlusion-aware L_2 Normal Error is defined as:

$$\text{Occ-L2-NormErr}^{(i)} = \frac{1}{|\mathcal{O}^{(i)}|} \sum_{(x,y) \in \mathcal{O}^{(i)}} \left\| \hat{N}^{(i)}(x, y) - N^{(i)}(x, y) \right\|_2^2.$$

All occlusion-aware metrics are averaged over both instances and across the four canonical viewpoints to provide a robust estimate of reconstruction performance in visually occluded regions.

C. Additional Results of 3D Multi-Human Reconstruction

C.1. Qualitative Comparison Including Additional Baselines

In addition to the baselines presented in the main paper, we include two additional baselines for comparison: DeepMultiCap [29], a method designed for multi-human reconstruction from multi-view images, and Multiply [9], a method for multi-human reconstruction from videos. Fig. S9 presents additional qualitative comparisons on the in-the-wild images, while Fig. S10 shows results on the MultiHuman dataset. In the in-the-wild setting, where SMPL-X predictions are used instead of ground-truth, our method continues to produce high-quality reconstructions, demonstrating robustness to SMPL-X estimation errors. Across all baselines, we observe common failure modes: incomplete geometry and missing textures in occluded regions, severe interpenetration or failure to preserve contact due to the lack of inter-person modeling, and inability to correct perspective distortion in images with complex viewpoints. In contrast, HUG3D consistently delivers robust multi-human reconstructions that preserve contact, correct geometric distortion, and hallucinate plausible textures even under severe occlusion.

C.2. Results with Predicted SMPL-X

Table S2. End-to-end evaluation using predicted masks, SMPL-X parameters, and camera estimates from RoBUDDI. Despite operating on predicted inputs, HUG3D outperforms existing baselines.

Method	CD↓	PSNR↑	CP↑
SIFU	26.268	10.037	0.006
SiTH	24.607	9.332	0.012
PSHuman	23.315	9.251	0.011

In the main paper, we evaluate reconstruction performance using ground-truth SMPL-X parameters to decouple reconstruction quality from pose estimation errors, following common protocols in prior work (e.g., PSHuman [15], ECON [23]). To further assess robustness under realistic conditions, we additionally report end-to-end results using predicted masks, SMPL-X parameters, and camera estimates obtained via RoBUDDI. As shown in Tab. S2, HUG3D consistently outperforms all baselines even when operating on predicted inputs.

C.3. Separate Results for Each Instance

Table S3 shows per-instance comparisons of geometry and texture metrics. Our method consistently outperforms baselines across all measures, achieving better geometric accuracy (e.g., lowest CD, P2S, and Norm $L2$; highest NC and F-score) and texture quality (highest PSNR/SSIM, lowest LPIPS). This instance-level analysis further highlights the effectiveness of our unified framework in capturing both fine-grained geometry and high-quality appearance.

C.4. Results Depending on Level of Interaction

Tables S4–S6 compare results across two interaction levels: Closely interactive and Naturally interactive. Our method consistently outperforms others in geometry, texture, and occluded regions. It demonstrates superior geometric fidelity (e.g., CD, NC, F-score), texture quality (PSNR, SSIM, LPIPS), and robustness under occlusions, regardless of interaction level. These results highlight the resilience and generalizability of our approach under varying interaction conditions.

C.5. Scalability to Larger Human Groups

We further evaluate the scalability of HUG3D on scenes containing three or more interacting humans. Although the model is trained only on single-person and pair interactions, it generalizes to larger groups through the joint diffusion and reconstruction framework. As shown in Fig. S11, HUG3D successfully reconstructs plausible multi-person interactions with three or more subjects while preserving consistent geometry and contact relationships.

C.6. Generalization to Out-of-Distribution Humans

To assess the robustness of our method, we tested HUG3D on novel human inputs, including stylized 3D characters and children—categories not present during training. As shown in Fig. S12, while minor mismatches in body proportions may occur due to distribution shifts, our model still generates geometrically plausible and semantically coherent outputs. These results highlight the strong generalization ability of HUG3D, even in challenging and unseen scenarios.

C.7. Results from Multiple Views

Figs. S15 and S16 show qualitative renderings of our reconstructed textured 3D mesh from a broad set of viewpoints. We visualize both training views (with gray backgrounds) and novel views (with white backgrounds), sampled across varying camera positions: elevations of {-45°, 0°, 45°} and azimuths of {0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°}. These results demonstrate the model's strong generalization capability to unseen perspectives for both normal maps and RGB images.

C.8. Robustness to Intermediate Errors

As shown in Fig. S13 and Fig. S14, our approach demonstrates robustness to inaccuracies in intermediate stages such as segmentation, depth estimation, diffusion predictions, and SMPL-X estimation. This robustness is enabled by our multi-view diffusion prior and physics-based, interaction-aware geometry reconstruction.

While our approach remains stable under moderate errors (the top two rows of Fig. S14), reconstruction quality degrades when the SMPL-X initialization is heavily corrupted, as shown in the last row of Fig. S14.

C.9. Videos

We provide an accompanying supplementary video that better visualizes the key advantages of our method, HUG3D. The video highlights that HUG3D produces physically plausible, high-fidelity 3D reconstructions of interacting people from a single image.

Figure S9. Additional qualitative comparison on multi-human 3D reconstruction from a single in-the-wild image. HUG3D outperforms baselines by correcting perspective distortion, preserving inter-human contact, and hallucinating plausible textures under heavy occlusion.

Table S3. Quantitative comparison of geometry and texture for each instance

Method	CD ↓	P2S ↓	NC ↑	F-score ↑	bbox-IoU ↑	Norm $L2$ ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SIFU	6.367	2.292	0.753	30.203	0.659	0.018	17.222	0.882	0.127
SiTH	9.642	3.166	0.712	21.740	0.541	0.024	16.090	0.881	0.143
PSHuman	16.876	6.384	0.614	9.561	0.402	0.039	13.720	0.857	0.188
DeepMultiCap	13.314	2.952	0.754	18.898	0.442	0.026	15.25	0.880	0.161
Ours	**3.531**	**1.719**	**0.816**	**42.946**	**0.801**	**0.012**	**18.659**	**0.894**	**0.102**

Table S4. Quantitative comparison of geometry depending on level of interaction.

Interaction	Method	CD ↓	P2S ↓	NC ↑	F-score ↑	bbox-IoU ↑	Norm $L2$ ↓	CP ↑
Closely	SIFU	7.267	2.750	0.724	24.335	0.757	0.033	0.117
	SiTH	10.908	3.491	0.697	19.216	0.694	0.044	0.281
	PSHuman	14.920	5.518	0.616	10.572	0.631	0.065	0.049
	DeepMultiCap	9.6697	2.745	0.764	20.471	0.606	0.039	0.123
	Ours	**4.315**	**2.121**	**0.811**	**37.243**	**0.838**	**0.022**	**0.326**
Natural	SIFU	4.895	2.069	0.768	31.510	0.788	0.026	0.076
	SiTH	8.486	3.044	0.714	21.877	0.715	0.038	0.068
	PSHuman	15.884	6.350	0.617	9.370	0.671	0.070	0.017
	DeepMultiCap	17.081	2.463	0.741	17.018	0.470	0.052	0.064
	Ours	**3.340**	**1.585**	**0.816**	**42.957**	**0.849**	**0.018**	**0.184**

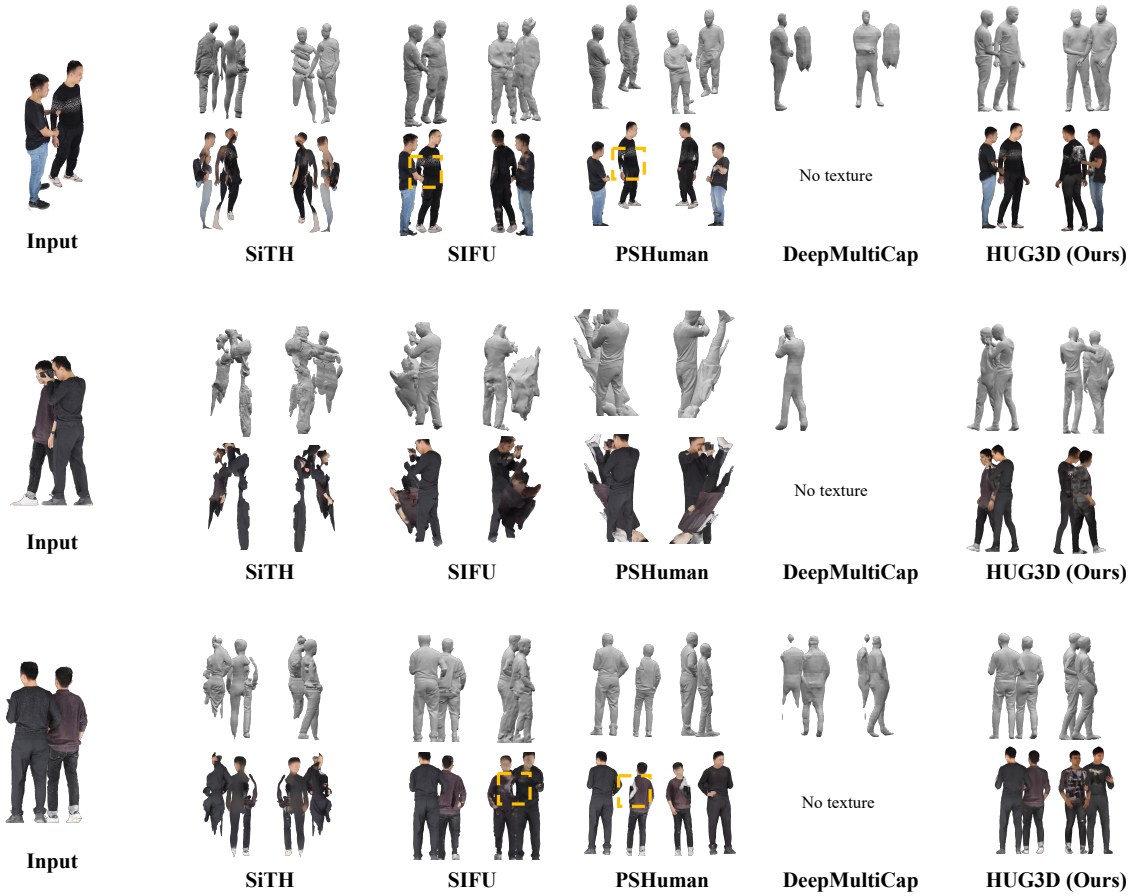


Figure S10. Additional qualitative comparison on multi-human 3D reconstruction from a single image in the MultiHuman dataset. Yellow boxes highlight broken geometry, missing texture, and incorrect inter-human interactions. HUG3D outperforms baselines by correcting perspective distortion, preserving inter-human contact, and hallucinating plausible textures under heavy occlusion.

Table S5. Quantitative comparison of texture depending on level of interaction.

Interaction	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Closely	SIFU	14.369	0.781	0.223
	SiTH	13.683	0.785	0.243
	PSHuman	11.722	0.747	0.293
	Ours	**16.454**	**0.805**	**0.179**
Natural	SIFU	15.586	0.799	0.192
	SiTH	13.851	0.790	0.228
	PSHuman	11.278	0.740	0.309
	Ours	**16.741**	**0.818**	**0.166**

Table S6. Quantitative comparison within occluded regions depending on level of interaction.

Interaction	Method	Norm $L2$ ↓	PSNR ↑	SSIM ↑
Closely	SIFU	0.223	5.745	0.569
	SiTH	0.218	5.900	0.551
	PSHuman	0.258	4.344	0.529
	DeepMultiCap	0.219	-	-
	Ours	**0.153**	**8.082**	**0.610**
Natural	SIFU	0.184	6.359	0.554
	SiTH	0.187	6.557	0.532
	PSHuman	0.249	4.757	0.501
	DeepMultiCap	0.216	-	-
	Ours	**0.138**	**8.358**	**0.599**

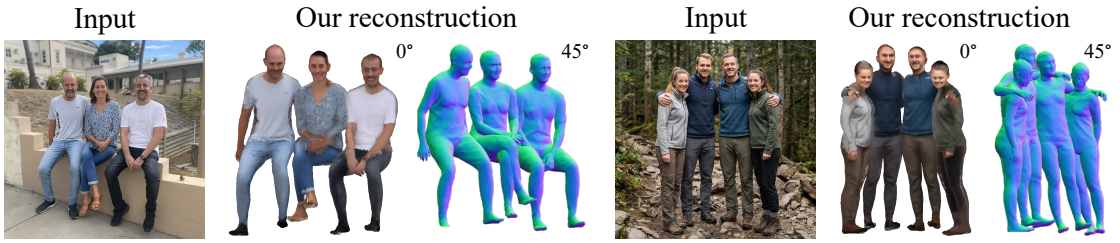


Figure S11. Qualitative results on scenes with three or more interacting humans. HUG3D reconstructs plausible multi-person interactions while preserving consistent geometry and contact relationships, demonstrating its scalability to larger human groups.

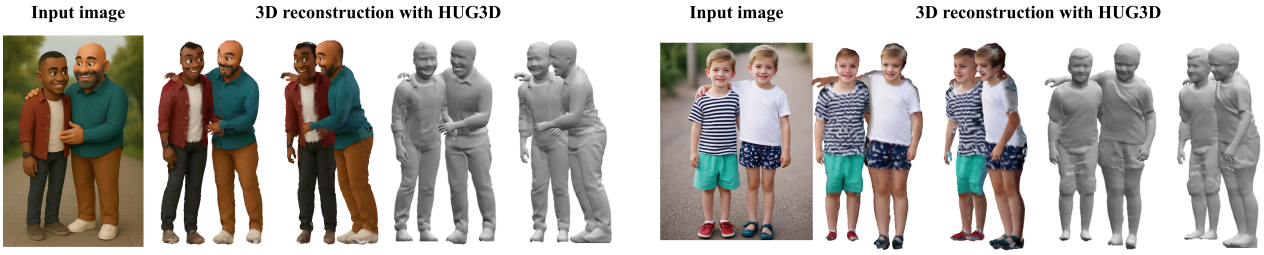


Figure S12. Qualitative results demonstrating HUG3D's generalization capability to novel human types, including stylized 3D characters and children. Despite domain differences, our method produces structurally plausible and semantically consistent outputs.

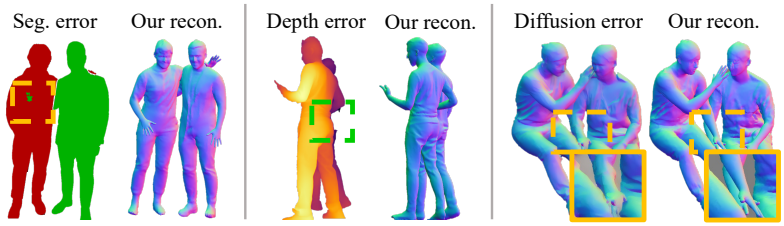


Figure S13. Robustness to errors in intermediate predictions. Despite inaccuracies in segmentation, depth estimation, or diffusion outputs, HUG3D maintains plausible multi-human reconstructions with consistent geometry and interaction relationships.

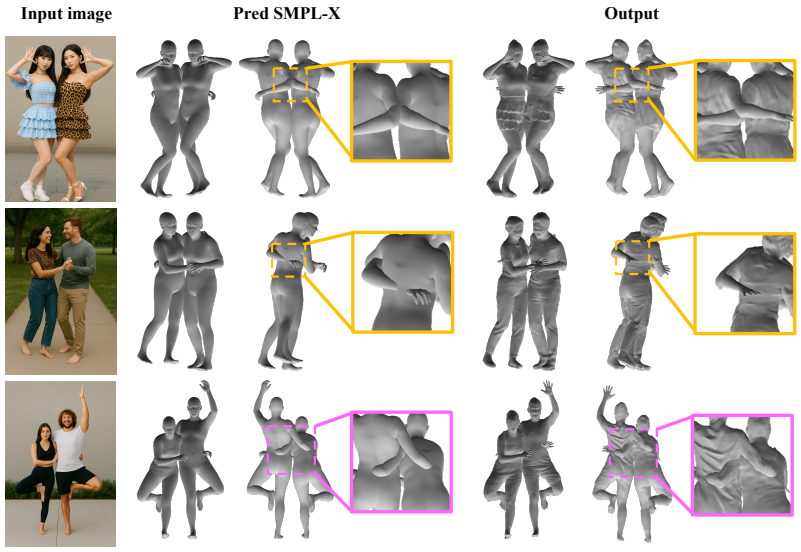


Figure S14. Robustness to SMPL-X estimation errors. The top two rows demonstrate that our method reduces interpenetration artifacts even with inaccurate SMPL-X estimates. The last row shows a failure case arising from severely corrupted SMPL-X initialization.

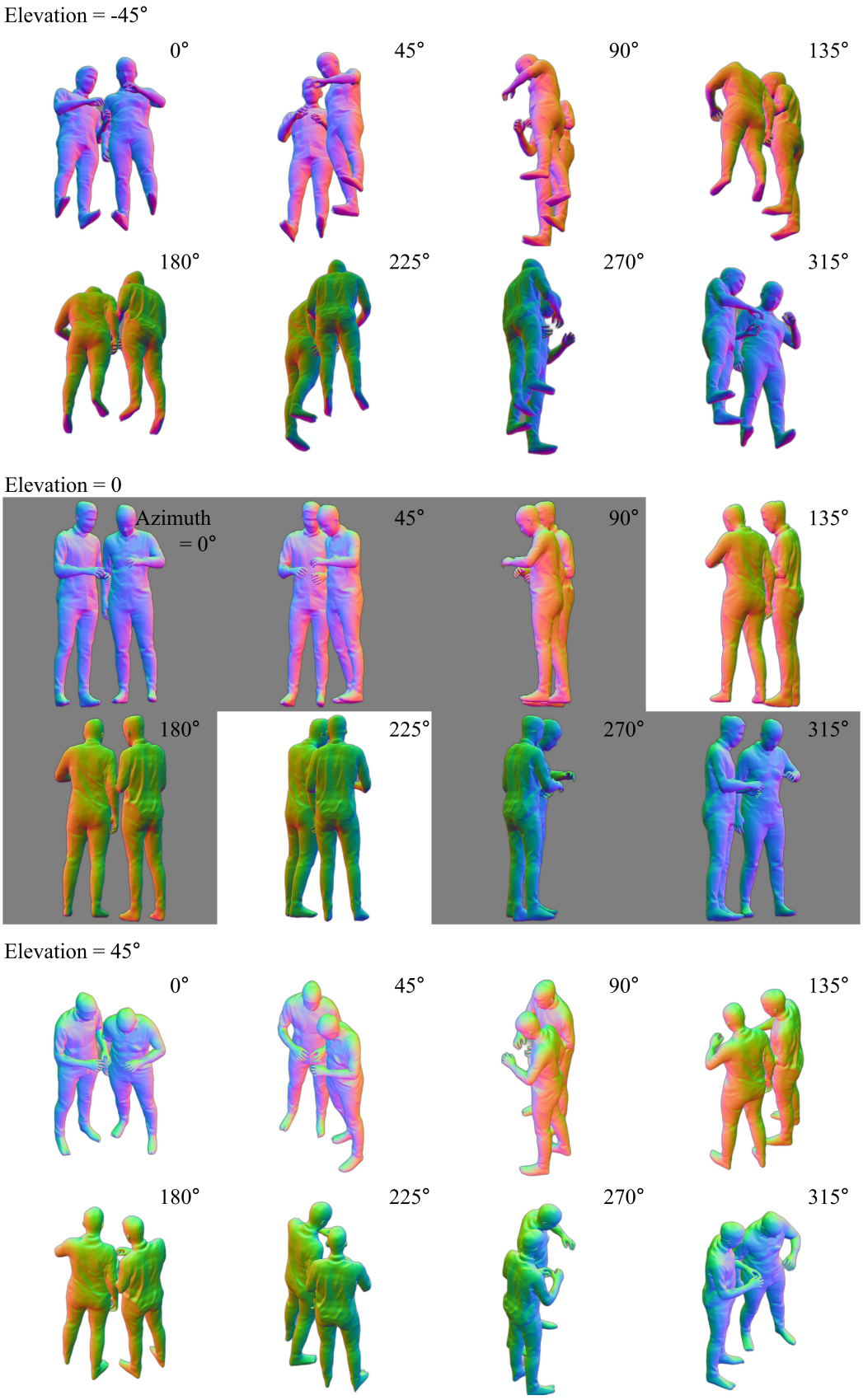


Figure S15. Normal maps rendered from multiple viewpoints of our reconstructed textured 3D mesh, including both training views (gray background) and novel views (white background).

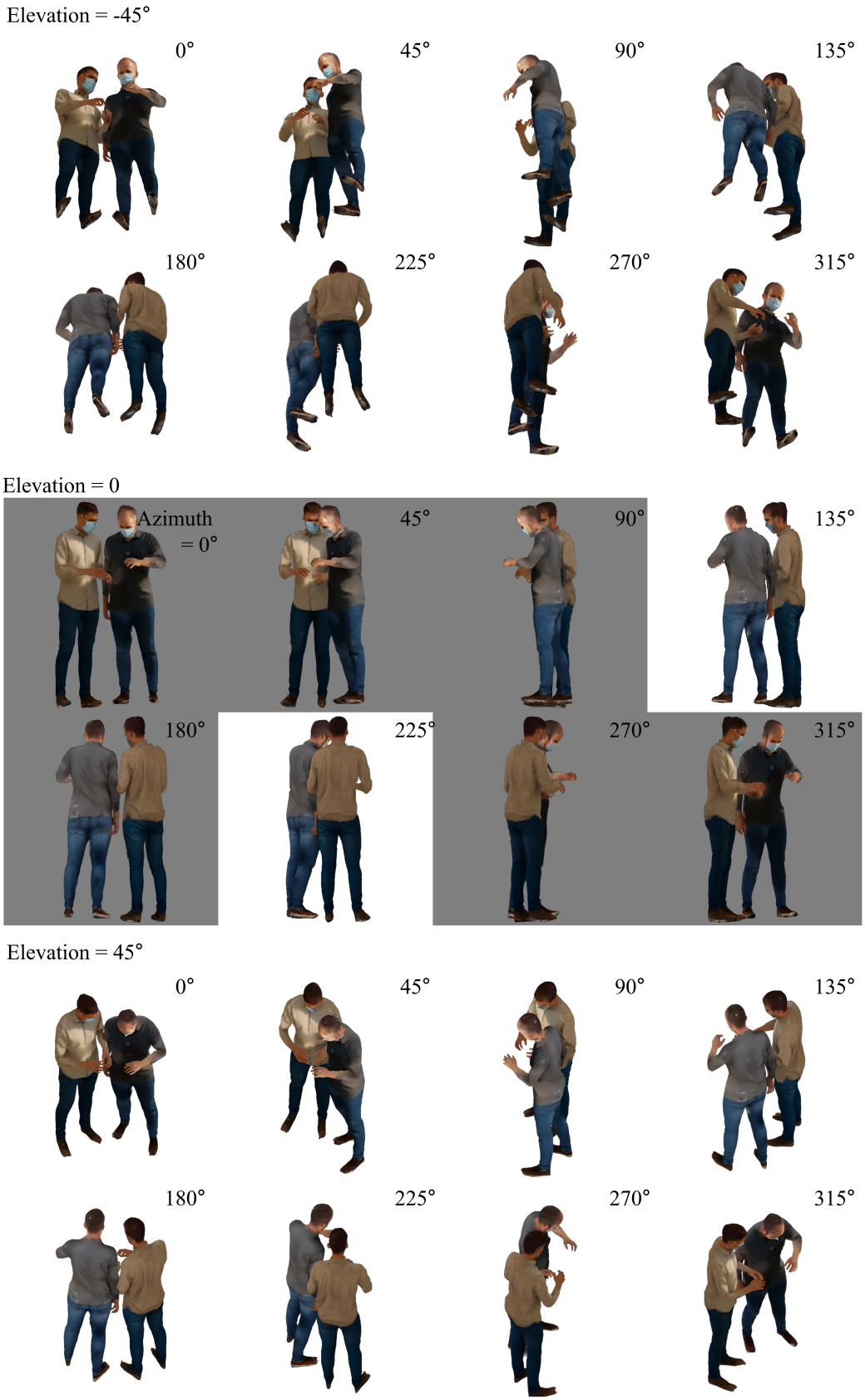


Figure S16. RGB renderings from multiple viewpoints of our reconstructed textured 3D mesh, including both training views (gray background) and novel views (white background).

D. Results from Each Component

Fig. S17 presents qualitative outputs from each stage of our framework. (1) *SMPL-X Estimation and Instance Segmentation* produce parametric body models and segmentation masks. (2) *Canonical Perspective-to-Orthographic View Transformation (Pers2Ortho)* enables reprojection of RGB images to a shared canonical view. (3) *Human Group-Instance Multi-View Diffusion (HUG-MVD)* generates multi-view consistent RGB and normal maps. (4) *Human Group-Instance Geometry Reconstruction (HUG-GR)* reconstructs accurate 3D meshes of multiple human subjects. (5) *Occlusion- and View-Aware Texture Fusion* synthesizes high-quality textured meshes by integrating multi-view information while handling occlusions and viewpoint variations.

E. Additional Ablation Studies and Analysis

E.1. Robust SMPL-X Estimation (RoBUDDI)

Table S7. Quantitative comparison of SMPL-X fitting accuracy on the MultiHuman dataset [29]. Metrics include mean per-joint position error (MPJPE), its Procrustes-aligned variant (PA-MPJPE), and mean vertex error (MVE).

Method	MPJPE ↓	PA-MPJPE ↓	MVE ↓
BEV	13.178	12.704	10.570
BUDDI	13.162	12.695	10.591
Ours (RoBUDDI)	**13.139**	**12.673**	**10.566**

We evaluate our proposed RoBUDDI on the MultiHuman dataset [29] and compare it against BEV [20] and BUDDI [16]. As shown in Tab. S7, RoBUDDI achieves lower MPJPE, PA-MPJPE, and MVE, demonstrating superior accuracy in 3D pose and shape estimation.

In addition to quantitative improvements, our method shows qualitative benefits as illustrated in Fig. S18. While BUDDI suffers from interpenetration artifacts between closely interacting subjects (yellow arrows), our RoBUDDI, enhanced with interpenetration and visibility-aware losses, yields more physically plausible and realistic 3D reconstructions.

E.2. Canonical Perspective-to-Orthographic View Transform (Pers2Ortho)

Depth Edge-Aware Uncertain Point Filtering. As shown in Fig. S1, removing uncertain points helps reduce jagged contours and ghosting artifacts near object boundaries after reprojection.

Depth-Aware Visible Point Selection. As shown in Fig. S2, this strategy filters out occluded or background points, retaining only those in front of the mesh surface and visible from the target camera view.

E.3. Human Group-Instance Multi-View Diffusion (HUG-MVD)

Perspective multi-view diffusion vs. Orthographic multi-view diffusion. We compare the results of multi-view diffusion models trained directly on perspective images with those trained on orthographic images in Fig. S19 with the same training settings. Due to the limited amount of ground-truth group data, the geometric complexity of group scenes, and the constrained capacity of the base diffusion model, models trained on perspective images often fail to produce plausible outputs (Fig. S19(a)). These failures manifest as artifacts such as twisted limbs and mixed clothing textures. In contrast, our multi-view diffusion model trained on orthographic images performs significantly better under limited data settings (Fig. S19(b)). This highlights the effectiveness of our strategy, which first transforms perspective images into orthographic views and then applies multi-view diffusion in a canonical space.

Comparison with state-of-the-art diffusion-based inpainting. Fig. S20 shows a detailed comparison between current state-of-the-art diffusion-based inpainting methods [13, 19] and our HUG-MVD, which achieves multi-view consistent inpainting and generation. Unlike existing inpainting approaches, HUG-MVD (1) produces multi-view consistent results, (2) maintains pose consistency without anatomical errors, and (3) additionally performs normal map inpainting—representing a significant advancement.

While several prior works address multi-view consistent inpainting, their tasks differ and are less suitable for human occlusion completion. For instance, MVInpainter [2] requires a consistent multi-view background and an inpainted object visible in the first view, and Instant3Dit [1] demands a 3D object as input. In contrast, our method only requires a single occluded human image as input, which is a more challenging and realistic scenario.

Training MVD with occlusion-aware masks and freeform masks. In Fig. S21, we evaluate our inpainting training strategy for HUG-MVD by comparing mask generation schemes. Since freeform masks do not reflect real world occlusion patterns, it biases the model towards learning mask artifacts and producing visibly unnatural inpainted regions. In contrast, our method enables more natural, artifact-free inpainting.

E.4. Occlusion- and View-Aware Texture Fusion

View-Aware Face Restoration. As shown in Fig. S7, our method effectively refines facial regions captured from extreme angles or under occlusion, such as back views, which often exhibit degraded appearances.

Occlusion-Aware Blending. As illustrated in Fig. S8, our method effectively prevents ghosting and bleeding artifacts near occlusion boundaries.

Table S8. Comparison of inference time and memory usage across baseline methods. Our method achieves superior performance while operating within the range of existing methods.

Metric	SIFU	ECON	SiTH	PSHuman	DeepMultiCap	MultiPly	HUG3D (Ours)
Elapsed Time (s)	333.79	80.09	148.01	128.47	42.35	27907.67	216.32
Peak VRAM (GB)	7.31	5.44	17.79	32.12	1.37	5.75	34.76

Table S9. Average elapsed time and peak VRAM usage for each pipeline stage.

Metric	Pers2Ortho	HUG-MVD	HUG-GR	Texture Fusion
Elapsed Time (s)	16.20	60.16	125.46	14.49
Peak VRAM (GB)	14.40	34.76	7.58	4.95

Table S10. Wilcoxon signed-rank test results (p-values) across all evaluation metrics, confirming statistically significant improvements of our method over baselines.

Method	CD	P2S	NC	F-score	bbox-IoU	Norm $L2$	CP	PSNR	SSIM	LPIPS	Occ.Norm $L2$	Occ.PSNR	Occ.SSIM
SIFU	1.8e-09	3.3e-08	3.9e-14	3.9e-10	1.6e-07	2.0e-39	1.1e-05	3.8e-27	1.7e-38	5.1e-37	5.7e-10	5.9e-10	1.9e-11
SiTH	3.6e-14	3.8e-14	3.6e-14	5.8e-14	4.6e-13	1.4e-51	8.2e-04	1.1e-48	2.2e-38	1.5e-51	1.3e-14	4.2e-14	1.4e-13
PSHuman	3.6e-14	3.6e-14	3.6e-14	5.2e-14	2.7e-13	1.4e-51	1.2e-08	1.4e-51	1.7e-51	1.4e-51	1.9e-18	1.4e-18	5.2e-17
DeepMultiCap	7.8e-13	1.1e-08	2.6e-12	1.3e-13	1.4e-13	1.4e-46	5.3e-06	3.5e-50	1.2e-27	5.2e-47	5.9e-11	1.1e-14	1.5e-14

E.5. Interaction-Aware Modeling

Table S11. Ablation study on interaction-aware modeling. Removing either component degrades performance, confirming that both stages are essential for accurate multi-human interaction reconstruction.

Method	CD↓	P2S↓	NC↑
w/o HUG-MVD	10.810	6.993	0.529
w/o HUG-GR	11.746	7.920	0.524
Ours	**10.627**	**6.877**	**0.537**

Tab. S11 reports metrics computed specifically within contact regions to isolate the contribution of each stage. *w/o HUG-MVD* is trained solely on single-human data and removes the joint multi-human diffusion inference, eliminating the generative priors required for modeling interactions. *w/o HUG-GR* removes the interaction-aware geometric reconstruction stage by disabling group-normal and physics-based losses, which reduces physical plausibility and contact precision. These results confirm that interaction-aware modeling in both stages is essential.

E.6. Efficiency Analysis

We provide a comparison of inference time and memory usage across baseline methods in Tab. S8. While our end-to-end inference time of 216 seconds per image is within the range of existing methods, this represents a reasonable trade-off, as our approach substantially outperforms them in reconstruction fidelity and physical plausibility. The primary runtime overhead arises from contact-mask computation in HUG-GR, which is necessary for accurate interaction modeling. Disabling it reduces runtime by 58.6% but degrades interaction realism (Fig. 6(b)). Moreover, the runtime scales linearly with the number of subjects, rather than exponentially.

We also measured the elapsed time and peak VRAM usage for each stage in our proposed method as shown in Tab. S9 with NVDIA A100. We observed that the most significant bottlenecks were identified to be HUG-GR (time-wise) and HUG-MVD (peak VRAM-wise), with each stage consuming 125.46 seconds and 34.76 GB of VRAM respectively.

E.7. Statistical Significance Analysis

We conducted Wilcoxon signed-rank tests [22] to assess statistical significance across all metrics in Tabs. 1, 2, 3. As shown in Tab. S10, the p-values confirm that our method consistently outperforms all baselines with statistically significant (p value ¡ 0.001) differences across geometry, texture, and occlusion handling metrics.

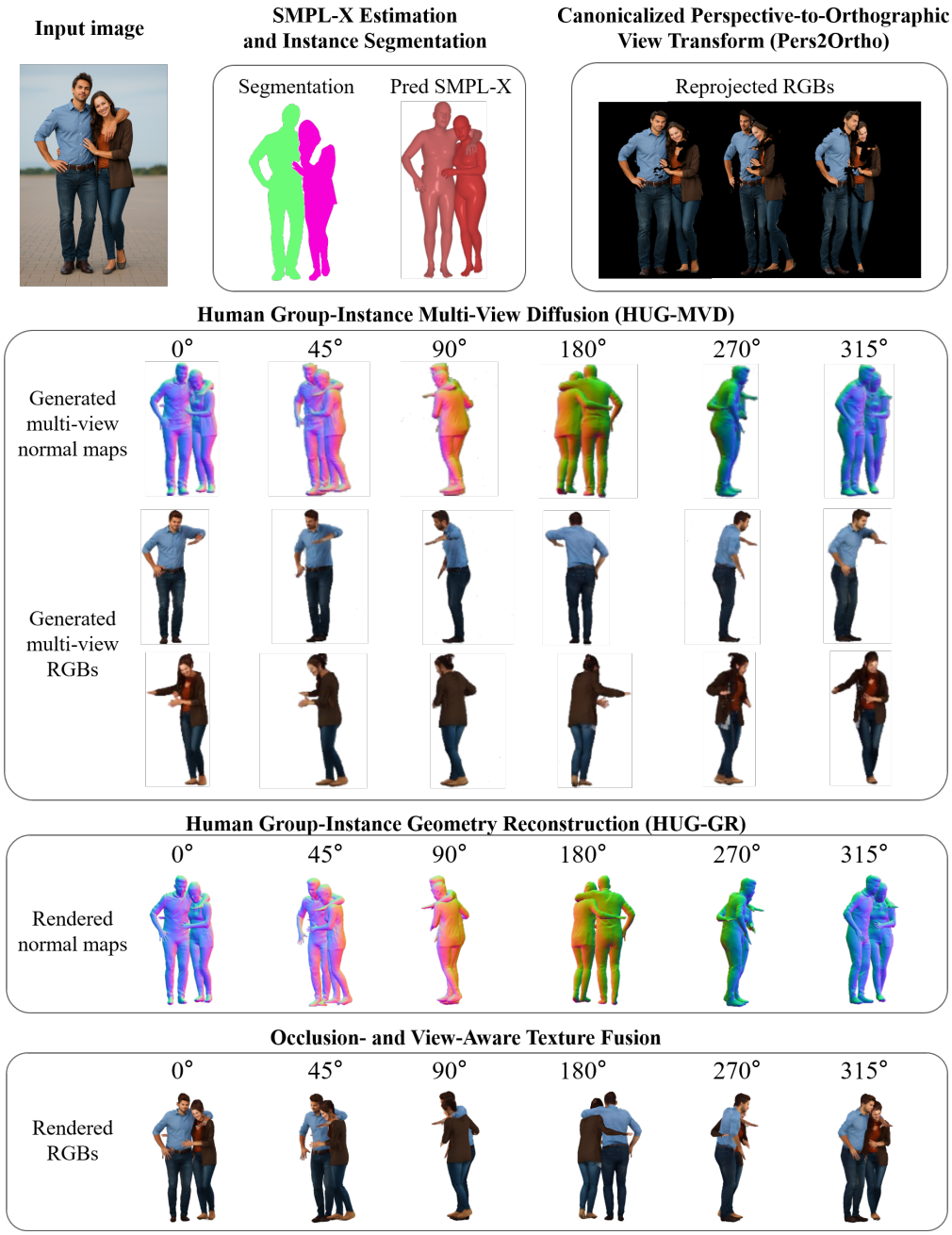
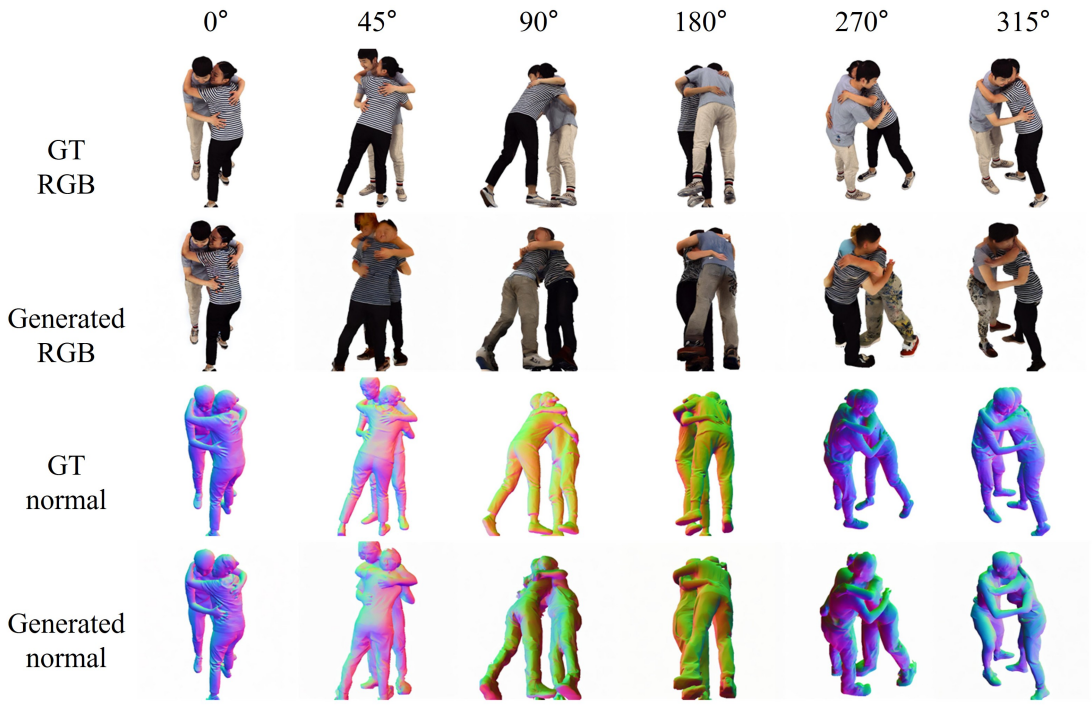
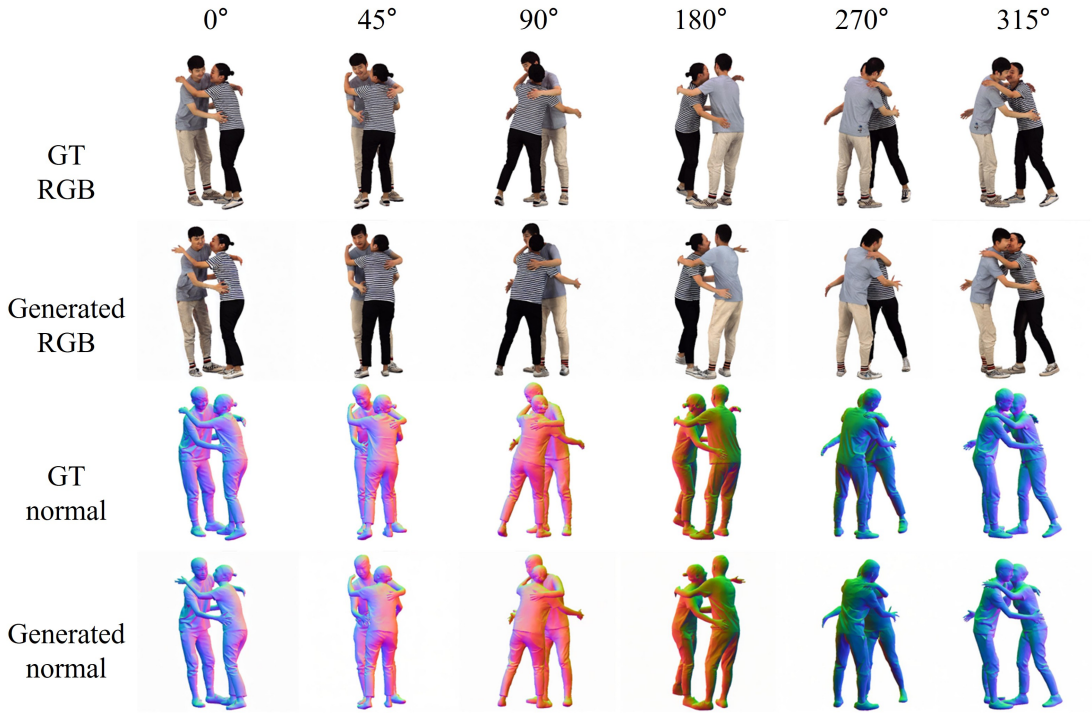


Figure S17. Example of results from each component of our HUG3D.

Figure S18. Qualitative comparison of our SMPL-X fitting (RoBUDDI) against BUDDI [16]. BUDDI exhibits visible interpenetrations between interacting subjects (yellow arrows), whereas our RoBUDDI produces more physically plausible results.



(a) Results of multi-view diffusion models trained directly on **perspective** images



(b) Results of multi-view diffusion models trained on **orthographic** images **(HUG-MVD)**

Figure S19. Comparison of results from multi-view diffusion models trained directly on perspective images vs. models trained on orthographic images.

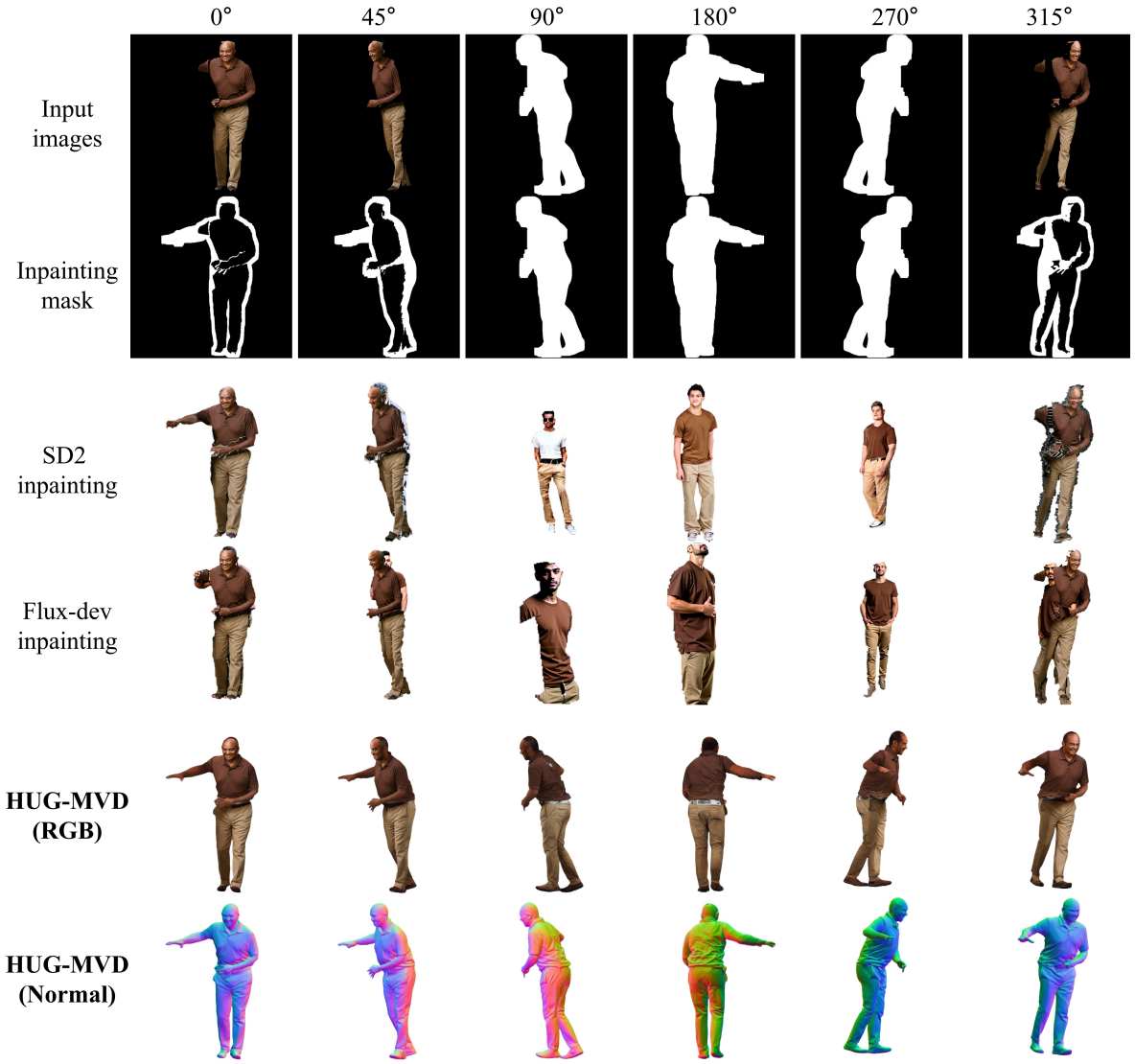


Figure S20. Comparison between state-of-the-art diffusion-based inpainting methods and our HUG-MVD framework for multi-view consistent inpainting and generation.

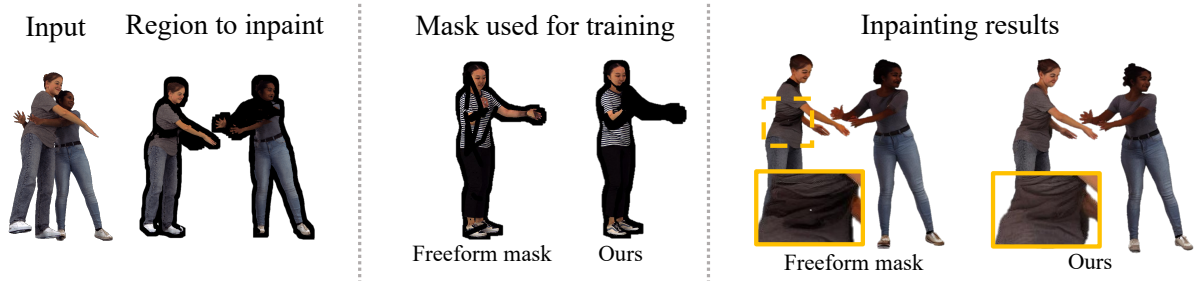


Figure S21. Ablation on mask types for training HUG-MVD. Our occlusion-simulated masks enhance inpainting compared to freeform masks.

F. Licenses for Existing Assets

F.1. Libraries

The libraries used in this work are shown in Tab. S12.

F.2. Datasets

The datasets used in this work are shown in Tab. S13.

F.3. Pretrained Models

The pretrained models used in this work are shown in Tab. S14.

G. Limitations, Impact and Safeguards

G.1. Limitations

While our approach demonstrates strong performance across various scenarios, we acknowledge several aspects that offer room for future improvement.

First, our method is trained under ambient lighting assumptions with consistent illumination across multiple views. In certain challenging cases such as low-light scenes or strong lighting contrast, minor failures may occur, as shown in the top-left of Fig. 4 in the main paper, the back side of the person appears relatively dark. We note that our method is a proof of concept, and these issues can potentially be mitigated through data augmentation, more diverse training data, or incorporation of synthetic lighting variations.

Second, our method focuses on inter-human occlusion but does not yet explicitly model object-induced occlusions. In cases where a person is holding or interacting with a object, the model may fail to recognize the object as separate and instead reconstruct it as part of the body, resulting in distorted geometry (see bottom-right of Fig. 4 in the main paper). We plan to address this limitation in future work by incorporating object-aware reasoning.

Third, in cases where the model relies on predicted SMPL-X input, errors in the pose estimation can lead to discrepancies between the input image and the reconstructed mesh. The model may tend to follow the predicted SMPL-X pose, resulting in slightly misaligned geometry with the input image. Nonetheless, as shown in Tab. S7 and Fig. S18, our method remains robust under such conditions and outperforms existing baselines.

Finally, there is currently no publicly available baseline that directly matches our problem setting—multi-human reconstruction from a single image. To allow meaningful comparisons, we carefully adapted related methods across different input modalities (e.g., single-human or multi-view approaches). Although these comparisons are not perfectly aligned, they offer reasonable context. We also note that relevant recent work such as [3] was not included due to the lack of released implementation.

G.2. Impact and Safeguards

This work can have significant potential across fields such as virtual reality, gaming, telepresence, digital fashion, and medical imaging. However, the ability to generate lifelike 3D representations from minimal input raises important ethical concerns around consent, data ownership, and control over digital likenesses. Moreover, the generated normal maps or 3D mesh can be used to infer sensitive biological data of the individual. It is therefore essential to limit access to the model through controlled licensing agreements and establish guidelines centered on the consent of the input image provider to minimize these concerns.

References

[1] Amir Barda, Matheus Gadelha, Vladimir G Kim, Noam Aigerman, Amit H Bermano, and Thibault Groueix. Instant3dit: Multiview inpainting for fast editing of 3d objects. *arXiv preprint arXiv:2412.00518*, 2024. 20

[2] Chenjie Cao, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvinpainter: Learning multi-view consistent inpainting to bridge 2d and 3d editing. *arXiv preprint arXiv:2408.08000*, 2024. 20

[3] Junuk Cha, Hansol Lee, Jaewon Kim, Nhat Nguyen Bao Truong, Jaeshin Yoon, and Seungryul Baek. 3d reconstruction of interacting multi-person in clothing from a single image. *WACV*, 2024. 10, 11, 26

[4] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. *CVPR*, 2020. 27

[5] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. *CVPR*, 2023. 4, 5, 27

[6] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. *CVPR*, 2024. 10, 11

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS*, 2021. 6

[8] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. *CVPR*, 2024. 8

[9] Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. Multiply: Reconstruction of multiple people from monocular video in the wild. *CVPR*, 2024. 10, 11, 13

[10] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 2

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *ICCV*, 2023. 2

[13] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 20

Table S12. Libraries used in the paper

Library	Link to license
Pytorch [17]	https://github.com/pytorch/pytorch/blob/main/LICENSE
Pytorch3D [18]	https://github.com/facebookresearch/pytorch3d/blob/main/LICENSE
Diffusers [21]	https://github.com/huggingface/diffusers/blob/main/LICENSE

Table S13. Datasets used in the paper

Dataset	Link to license
Hi4D [25]	https://hi4d.ait.ethz.ch
CustomHumans [5]	https://custom-humans.ait.ethz.ch/
THuman2.0 [26]	https://github.com/ytrock/THuman2.0-Dataset/blob/main/THUman2.1_Agreement.pdf
MultiHuman [29]	https://github.com/y-zheng18/MultiHuman-Dataset

Table S14. Pretrained Models used in the paper

Pretrained model	Link to license
Stable Diffusion 2.1 Unclip [19]	https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL
PSHuman [15]	https://github.com/pengHTYX/PSHuman/blob/main/LICENSE.txt
ControlNet [27]	https://github.com/lllyasviel/ControlNet/blob/main/LICENSE
CodeFormer [30]	https://github.com/sczhou/CodeFormer/blob/master/LICENSE
Face detector [4]	https://github.com/serengil/retinaface/blob/master/LICENSE

[14] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wei Xue, Wenhan Luo, et al. Era3d: high-resolution multiview diffusion using efficient row-wise attention. *NeurIPS*, 2024. 6

[15] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024. 5, 6, 10, 11, 13, 27

[16] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *CVPR*, 2024. 2, 20, 23

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 27

[18] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 4, 27

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 5, 20, 27

[20] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. *CVPR*, 2022. 20

[21] Patrick von Platen, Luke Lewis, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022. 27

[22] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945. 21

[23] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. *CVPR*, 2023. 10, 11, 13

[24] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022. 2

[25] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. *CVPR*, 2023. License: Non-commercial academic use only. See https://hi4d.ait.ethz.ch/. 4, 5, 6, 27

[26] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. *CVPR*, 2021. 4, 5, 27

[27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *ICCV*, 2023. 5, 27

[28] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. *CVPR*, 2024. 10, 11

[29] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. *ICCV*, 2021. 10, 11, 13, 20, 27

[30] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *NeurIPS*, 2022. 8, 27