

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 HOIGS: HUMAN-OBJECT INTERACTION GAUSSIAN SPLATTING FROM MONOCULAR VIDEOS

005 **Anonymous authors**

006 Paper under double-blind review



018 **Figure 1: Comparison between our method and previous approaches.** This figure compares
019 rendering results between ExAvatar (Moon et al. (2024)), a human-centric model, and Ex4DGS (Lee
020 et al. (2024)), which uses a single motion field for all motions. ExAvatar reconstructs only humans,
021 while Ex4DGS fails to represent contact in interaction scenarios, producing artifacts and noise around
022 contact regions.

023 ABSTRACT

025 Reconstructing dynamic scenes with complex human–object interactions is a fun-
026 damental challenge in computer vision and graphics. Existing Gaussian Splatting
027 methods either rely on human pose priors, neglecting dynamic objects, or approxi-
028 mate all motions within a single field, limiting their ability to capture interaction-rich
029 dynamics. To address this gap, we propose Human-Object Interaction Gaussian
030 Splatting (HOIGS), which explicitly models interaction-induced deformation be-
031 tween humans and objects through a cross-attention based HOI module. Distinct
032 deformation baselines are employed to extract complementary motion features: hex-
033 plane for humans and Cubic Hermite Spline (CHS) for objects. By integrating these
034 heterogeneous features, HOIGS effectively captures interdependent motions and im-
035 proves deformation estimation in scenarios involving occlusion, contact, and object
036 manipulation. Comprehensive experiments on multiple datasets demonstrate that
037 our method consistently outperforms state-of-the-art human-centric and 4D Gaus-
038 sian approaches, highlighting the importance of explicitly modeling human–object
039 interactions for high-fidelity reconstruction. The video results of HOIGS are avail-
040 able at: <https://anonymous.4open.science/w/HOIGS-0F47/>

041 1 INTRODUCTION

044 Reconstructing videos of scenes that involve complex interactions between humans and objects
045 and synthesizing novel viewpoints constitute a central research problem in computer vision and
046 graphics. These techniques can be extended to various applications, including virtual reality, the
047 metaverse, and 3D animation. However, the inherent limitations of monocular cameras and the need
048 to accurately model intricate interactions between humans and objects remain major challenges for
049 achieving high-quality reconstruction. Addressing these issues is essential for enabling realistic scene
050 understanding and representation.

051 Recent approaches on human-centric video scene reconstruction (Kocabas et al. (2024); Moon et al.
052 (2024); Hu et al. (2024c); Qian et al. (2024); Liu et al. (2024); Hu et al. (2024a); Wen et al. (2024);
053 Kim et al. (2025)) have combined human pose estimation with 3D Gaussian Splatting (3DGS)
(Kerbl et al. (2023)) to model dynamic scenes. Typically, SMPL (Loper et al. (2023)) parameters are

054 regressed in advance for each frame, and a canonical space is defined using a T-pose as the reference.
 055 Within this space, 3D Gaussian parameters are established and trained using feature planes and MLPs.
 056 Subsequently, deformation to each frame’s 3D space is performed via Linear Blend Skinning (LBS)
 057 ([Loper et al. \(2023\)](#)), allowing for scene reconstruction and rendering. These methods have evolved
 058 into specialized models focused on humans and static backgrounds, achieving reliable performance
 059 when accurate human pose priors are available. However, existing approaches mainly focus on
 060 modeling humans alone, and thus fail to reconstruct complete scenes that involve objects beyond
 061 the human body. As a result, dynamically moving objects are often treated as static background
 062 or even disappear from the reconstructed scene. Even when deformations of objects are modeled
 063 separately, the interactions between humans and objects are not sufficiently considered in dynamic
 064 scenarios, which leads to artifacts and noisy results in the interaction regions, as shown in [Fig. 1](#).
 065 Consequently, accurately reconstructing scenes that involve both humans and objects requires new
 066 modeling paradigms that extend beyond conventional human-centric frameworks.
 067

068 Recent studies on 4D Gaussian Splatting extend beyond humans to encompass arbitrary moving
 069 objects, offering the advantage of general applicability. However, they generally exhibit lower
 070 reconstruction performance for humans compared to human-centric models. These approaches
 071 typically either define a canonical space and learn an implicit function that deforms it into the world
 072 coordinate system ([Wu et al. \(2024\)](#); [Jung et al. \(2023\)](#); [Bae et al. \(2024\)](#)), or explicitly parameterize
 073 object motions and optimize the corresponding parameters ([Yang et al. \(2023b\)](#); [Li et al. \(2024a\)](#); [Lee](#)
 074 [et al. \(2024\)](#)). Nevertheless, they do not explicitly model interactions between objects and instead
 075 treat all moving entities within a single motion field, which limits their ability to capture complex
 076 interactions. As a result, implicit methods struggle to represent long-term or highly non-linear
 077 motions in a stable manner, while explicit methods fail to handle scenarios such as contact and object
 manipulation, as ignoring the mutual interactions between motions limits their ability to capture
 realistic dynamics.

078 To overcome these limitations, we propose Human-Object Interaction Gaussian Splatting (HOIGS), a
 079 unified framework for reconstructing complex video scenes that involve both humans and dynamic
 080 objects. Unlike previous approaches that either model only human motion or employ a single motion
 081 field for all entities, our framework explicitly incorporates human–object interactions to achieve more
 082 faithful deformation modeling.

083 At the core of our framework lies the HOI module, which adopts a mutual attention mechanism
 084 to capture the bidirectional dependencies between human features and object motion features at
 085 each frame. Specifically, the module receives temporally varying human features, derived from the
 086 dynamic components of the hexplane representation, together with object motion features, obtained
 087 by embedding velocity vectors and their associated parameters. By explicitly learning how these two
 088 types of features influence one another, the HOI module effectively overcomes the shortcomings of
 089 prior methods that modeled humans and objects independently, which often resulted in artifacts and
 090 unstable reconstructions in interaction-rich scenes.

091 Furthermore, we design different deformation baselines tailored to humans and objects. For objects,
 092 we employ the Cubic Hermite Spline (CHS) to capture continuous motion trajectories, embedding the
 093 velocity vectors of keyframe Gaussians along with additional learnable parameters to construct robust
 094 object motion features. For humans, we utilize hexplane as the deformation baseline, where time-
 095 varying parameters are leveraged to represent fine-grained human deformation in both spatial and
 096 temporal domains. The extracted features from both humans and objects are subsequently integrated
 097 within the HOI module, which outputs offset vectors for each entity. This design ultimately enables
 098 our framework to achieve accurate and stable deformation estimation, even under complex scenarios
 099 involving close contact, mutual manipulation, or other intricate human–object interactions.

100 In summary, our main contributions are as follows:

- 101 • We propose an entity-aware cross-attention HOI module that enforces motion consistency
 102 between humans and objects. By attending to their features, it captures interdependent
 103 dynamics and improves reconstruction during contact and manipulation.
- 104 • We design distinct strategies for humans and objects using tailored deformation baselines.
 105 Hexplane encodes temporal and spatial features for human motion, while Cubic Hermite
 106 Splines (CHS) embed velocity vectors and learnable parameters for objects. This separation
 107 enables accurate and expressive motion representations for both entities.

108 • We conduct extensive experiments on diverse human–object interaction scenes and demon-
 109 strate that our method achieves more accurate reconstruction compared to existing human-
 110 centric and 4D Gaussian approaches.
 111

112 **2 RELATED WORKS**
 113

114 **2.1 HUMAN MODELING**
 115

116 Research on realistic human modeling has long been pursued. Early parametric models enabled
 117 efficient estimation of human pose, exemplified by HMR ([Kanazawa et al. \(2018\)](#)), but struggled
 118 to capture clothing and accessories. To address this, implicit function-based methods ([Huang et al.
 119 \(2020\)](#); [Saito et al. \(2019; 2020\)](#); [Xiu et al. \(2022; 2023\)](#)) were proposed, which recover fine details
 120 such as hair and clothing but remain limited in global consistency and rendering efficiency. These
 121 methods mainly focused on human geometry with little attention to human-object interactions. With
 122 Neural Radiance Fields (NeRF) ([Mildenhall et al. \(2021\)](#)), several works applied it to human modeling
 123 ([Peng et al. \(2021\)](#); [Jiang et al. \(2022\)](#); [Weng et al. \(2022\)](#); [Alldieck et al. \(2022\)](#); [Liao et al. \(2023\)](#);
 124 [Guo et al. \(2023\)](#)), achieving realistic appearance and view consistency but still suffering from high
 125 training cost and slow rendering. In terms of human-object interactions, some attempts ([Fan et al.
 126 \(2024\)](#)) introduced objects, yet dynamic interactions were not fully captured. Recently, 3D Gaussian
 127 Splatting (3DGS) ([Kerbl et al. \(2023\)](#)) emerged as a new representation and has been applied to
 128 human reconstruction ([Kocabas et al. \(2024\)](#); [Moon et al. \(2024\)](#); [Hu et al. \(2024c\)](#); [Liu et al. \(2024\)](#);
 129 [Hu et al. \(2024a\)](#)). However, most efforts still regard objects as static. Recent approaches leverage
 130 multi-view data for high-fidelity results. For example, Animatable Gaussian [Li et al. \(2024b\)](#) and
 131 GASPACHO [Mir et al. \(2025\)](#) utilize multi-view setups to create relightable avatars and disentangle
 132 human-object interactions. However, these methods depend on complex capture settings, whereas
 133 real-world scenarios are predominantly monocular. To address this, we propose HOIGS, a model
 134 for stable human reconstruction that explicitly captures human–object interactions from monocular
 135 inputs.

136 **2.2 DYNAMIC SCENE MODELING**
 137

138 The field of dynamic scene rendering and reconstruction has seen a paradigm shift from initial NeRF-
 139 based methods ([Park et al. \(2021a;b\)](#); [Wu et al. \(2022\)](#); [Fridovich-Keil et al. \(2023\)](#)) to the more recent
 140 3D Gaussian Splatting framework. Previous studies such as HOSNeRF ([Liu et al. \(2023\)](#)) effectively
 141 modeled human-object interactions by controlling human motion through skeleton-based models
 142 such as SMPL and leveraging object state embeddings. Nevertheless, the implicit representation
 143 inherent to NeRF led to significant computational overhead in training and rendering, and limited the
 144 ability to represent detailed features in large-scale environments. To address this efficiency bottleneck,
 145 a line of work has emerged that extends 3DGS to the temporal domain, known as 4D Gaussian
 146 Splatting (4DGS) ([Wu et al. \(2024\)](#); [Yang et al. \(2023b\)](#)). Although these methods achieve real-time
 147 rendering speeds, they face persistent issues. Most 4DGS approaches rely on Structure-from-Motion
 148 for Gaussian initialization, which is fundamentally ill-suited for dynamic subjects as it operates on
 149 the assumption of a static world. This leads to inaccurate point cloud generation for moving objects.
 150 Moreover, the MLP-based implicit deformation fields used to capture motion, while adequate for
 151 simple trajectories, often result in over-smoothed or unnatural movements when applied to complex,
 152 in-the-wild scenarios. Therefore, we propose an explicit, spline-based motion model. This approach
 153 allows us to model intricate temporal movements with high fidelity, achieving high-quality rendering
 154 even in dynamic scenes that include complex human-object interactions.

155 **3 METHOD**
 156

157 As shown in Fig. 2, we reconstruct the scene by independently modeling the deformations of humans
 158 and objects, and then incorporating interaction-aware transformations through the HOI module.
 159 Object deformations are estimated using a Cubic Hermite Spline (CHS). Human deformations are
 160 based on hexplane features, where time-invariant spatial features are used to learn the texture of the
 161 canonical T-pose, and Linear Blend Skinning (LBS) is subsequently applied to deform the canonical
 representation into each world space. Using these deformation baselines, we independently model

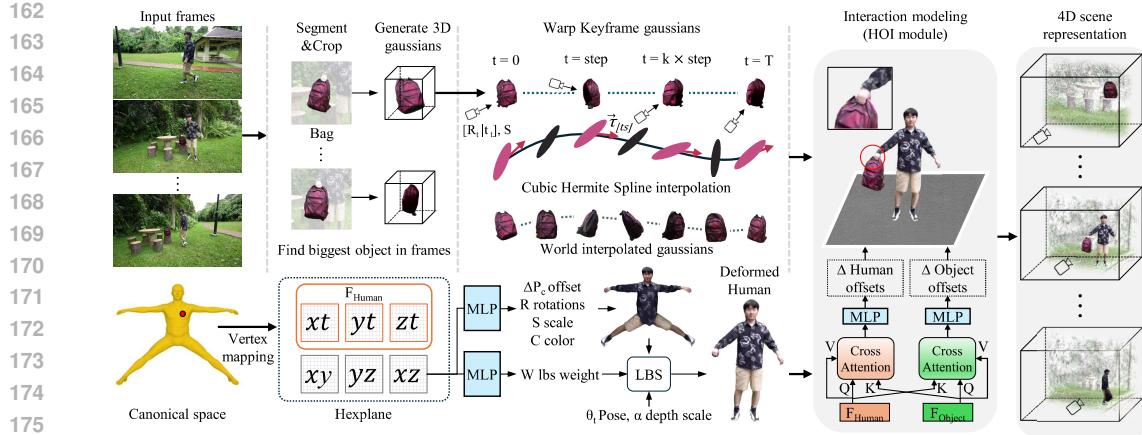


Figure 2: **Overview of the Proposed Framework.** Given an input video sequence, we first extract object-specific information, which is then used to reconstruct the 3D object shape via a diffusion prior. Based on the reconstructed shape, we initialize 3D Gaussians for each keyframe and use spline-based deformation as the baseline, where time-invariant and time-varying hexplane features are employed for canonical humans and interaction modeling, respectively. The final deformation is modeled through the HOI module, which learns interactions using human features and object motion features.

humans and objects and estimate their approximate positions for each frame, from which motion features are extracted. Finally, the extracted human and object features are fed into the HOI module, which accounts for interaction-driven transformations and determines the final positions of humans and objects in the reconstructed interaction scene.

3.1 OBJECT DEFORMATION

Object Initialization. We begin by segmenting the object of interest and cropping the object region from a representative frame of the entire sequence. Next, we apply a diffusion model initialized from DreamScene4D Chu et al. (2024) and guided by SDS loss to generate a canonical 3D Gaussian point cloud of the object. However, the 3D Gaussians generated through this diffusion prior may differ from the actual object geometry. While diffusion models can generate plausible 3D shapes from images, they often fail to precisely recover the true object structure. To address this, we align the diffusion-based canonical 3D Gaussians with the world space. First, we estimate a per-frame warping scale S_t by minimizing the discrepancy between the projected 3D bounding box and the 2D mask bounding box for each frame t . We then compute the global scale S as the average over all frames:

$$S_t = \arg \min_{S_t} \|\text{BBox}_{\text{proj}}(S_t \cdot \text{Gaussians}) - \text{BBox}_{\text{mask}}^t\|, \quad S = \frac{1}{T} \sum_{t=1}^T S_t.$$

where T is the total number of frames. We then transform the scaled Gaussians to the world coordinate system using the COLMAP camera extrinsics R_t and \mathbf{t}_t as follows:

$$\mathbf{x}_{\text{world}}^t = R_t^{-1}(\mathbf{x}_{\text{cam}}^t - \mathbf{t}_t).$$

From the warped Gaussians G_k of each keyframe, we extract each Gaussian's mean and color value, while initializing the remaining 3D Gaussian parameters with identity values.

Object Deformation. Based on the redefined mean and color from the keyframes, we construct the object's 3D Gaussians and use them to model the object deformation. To represent the continuous motion of the object over time, we model the mean values of each Gaussian as control-point-based curves. Specifically, we define a Cubic Hermite Spline function $\text{CHS}(t, \mathbf{m})$, and estimate the position of an object Gaussian at time t , denoted as $\mathbf{M}(t)$, as follows:

This initialization ensures that the explicit 3D Gaussian deformation model aligns with the actual object geometry and structural information. From the warped Gaussians of each keyframe, we extract

each Gaussian’s mean and color value, while initializing the remaining 3D Gaussian parameters with identity values. We apply a diffusion prior with SDS loss to reconstruct the object from a representative frame of the entire sequence. The reconstructed object is then warped using the camera parameters of each keyframe to initialize the corresponding 3D Gaussians. However, the 3D Gaussians generated through the diffusion prior may differ from the actual object geometry. While diffusion models can generate plausible 3D shapes from images, they often fail to precisely recover the true object structure. To address this, we introduce an explicit 3D Gaussian deformation model that aligns the diffusion-based initialization with the actual object geometry and structural information. From the warped Gaussians G_k of each keyframe, we extract each Gaussian’s mean and color value, while initializing the remaining 3D Gaussian parameters with identity values. Based on the redefined mean and color from the keyframes, we construct the object’s 3D Gaussians and use them to model the object deformation. To represent the continuous motion of the object over time, we model the mean values of each Gaussian as control-point-based curves. Specifically, we define a Cubic Hermite Spline function $CHS(t, \mathbf{m})$, and estimate the position of an object Gaussian at time t , denoted as $M(t)$, as follows:

$$M(t) = CHS(t, \mathbf{m}), \quad (1)$$

where $\mathbf{m} = \{m_k \mid m_k \in \mathbb{R}^3\}_{k \in [0, N_{key}-1]}$ is a learnable set of control points representing the mean positions of the Gaussians at each key frame, and N_{key} denotes the number of key frames. $CHS(t, \mathbf{m})$ is formulated as

$$CHS(t, \mathbf{m}) = (2t_r^3 - 3t_r^2 + 1)m_{\lfloor t_s \rfloor} + (t_r^3 - 2t_r^2 + t_r)\tau_{\lfloor t_s \rfloor} + (-2t_r^3 + 3t_r^2)m_{\lfloor t_s \rfloor + 1} + (t_r^3 - t_r^2)\tau_{\lfloor t_s \rfloor + 1}, \quad (2)$$

where $t_r = t_s - \lfloor t_s \rfloor$, $t_s = t_n(N_{key} - 1)$, $t_n = \frac{t}{N_f - 1}$ and N_f denotes the number of all frames. $m_{\lfloor t_s \rfloor}$ denotes the mean of the 3D Gaussians corresponding to the $\lfloor t_s \rfloor$ -th key frame.

In the standard formulation, $\tau_{\lfloor t_s \rfloor}$ represents the tangent vector with respect to the means of the surrounding Gaussians, which is typically approximated as $\tau_{\lfloor t_s \rfloor} = \frac{1}{2}(m_{\lfloor t_s \rfloor + 1} - m_{\lfloor t_s \rfloor - 1})$. Instead of using this fixed approximation, we reinterpret $\tau_{\lfloor t_s \rfloor}$ as a *velocity vector* and employ it as a learnable parameter. By embedding this velocity, we construct motion features that better capture the dynamic behavior of objects over time.

The position parameter \mathbf{m} between key frames is estimated via spline interpolation using both the Gaussian positions m_k at the key frames and the corresponding velocity vectors $\tau_{\lfloor k \rfloor}$. Only the Gaussians at the key frames are directly optimized during training. Once the intermediate Gaussians are estimated and rendered, the resulting gradients from the loss function are backpropagated to update the parameters of the corresponding key frame Gaussians. Among the Gaussian parameters, rotation and opacity are defined as time-dependent variables. The rotation parameter is modeled using Spherical Linear Interpolation based on the Gaussian rotations at each key frame, enabling smooth transitions over time. The opacity parameter varies with time to account for occluded regions caused by object motion. In contrast, the scale parameter is kept constant across all corresponding Gaussians at different key frames.

3.2 HUMAN DEFORMATION

Hexplane-based Deformation. We model human deformation using hexplane features. Specifically, we adopt time-invariant spatial features f from hexplane to learn the texture of the canonical T-pose mesh T_c in the canonical space. The features f are processed by an MLP head ψ to learn the Gaussian properties in the canonical space. This representation serves as the baseline for human deformation. The canonical human representation is then deformed into the posed world space using Linear Blend Skinning (LBS) as follows:

$$\psi_h(f(T_c)) = (c, o, \Delta P_c, R, S, W), \quad (3)$$

$$P_{def} = \alpha * LBS(P_c, \theta, W), \quad (4)$$

where θ denotes the set of SMPL-X pose parameters and α is a learnable scale parameter for human pose. Equation (3) extracts the Gaussian properties (color c , opacity o , position offset ΔP_c , rotation R , scale S and skinning weights W) from the canonical hexplane features, while Equation (4) applies the LBS function to obtain the deformed positions P_{def} of the Gaussians in the posed space.

270 To ensure that the reconstructed human representation matches the actual geometry, we further apply
 271 a depth supervision loss:

$$272 \quad \mathcal{L}_{\text{depth}} = \|D_{\text{render}} - D\|_1, \quad (5)$$

273 where D_{render} is the rendered depth map from the deformed Gaussians and D is the depth obtained
 274 from an off-the-shelf metric depth estimation model and further scaled using the COLMAP point
 275 cloud. This depth-guided supervision constrains the learnable scale parameter α and improves
 276 geometric fidelity in the reconstructed human shape.

279 3.3 CROSS-ATTENTION INTERACTION MODULE

281 **Feature Extraction.** We extract time-varying features from both humans and objects to learn their
 282 interactions. For humans, instead of relying on time-invariant texture features from the canonical
 283 space, we utilize time-varying features from hexplane. Furthermore, since it is not possible to know
 284 in advance which body parts are involved in object interactions, we divide the human body into 16
 285 parts and extract hexplane features for each part.

286 For objects, the features are derived from the velocity embeddings associated with each keyframe in
 287 the deformation process, which capture the local motion information at those frames. In addition,
 288 we embed learnable parameters for each keyframe to represent latent motion characteristics that
 289 cannot be fully captured by velocity alone. These velocity vectors and learnable parameters are then
 290 projected together with the corresponding time values, enabling the construction of object motion
 291 features. This formulation allows us to obtain continuous motion features for objects across all frames,
 292 rather than being limited to discrete keyframes.

293 **Interaction Module.** The proposed Interaction module (HOI module) takes time-varying features
 294 of humans and objects as inputs and explicitly models their interactions. Let the human and object
 295 features be denoted as F_{Human} and F_{Object} . To capture interdependencies between the two, we apply
 296 *mutual attention*, where queries, keys, and values are defined as:

$$297 \quad Q_h, K_h, V_h = F_{\text{Human}} W_h^Q, F_{\text{Human}} W_h^K, F_{\text{Human}} W_h^V, \quad (6-7)$$

$$299 \quad Q_o, K_o, V_o = F_{\text{Object}} W_o^Q, F_{\text{Object}} W_o^K, F_{\text{Object}} W_o^V.$$

300 Cross-attention is then performed in both directions, from human to object and from object to human,
 301 while incorporating a distance mask B into the attention computation:

$$302 \quad F'_{\text{Human}} = \text{softmax} \left(\frac{Q_h K_o^\top}{\sqrt{d}} + B \right) V_o, \quad F'_{\text{Object}} = \text{softmax} \left(\frac{Q_o K_h^\top}{\sqrt{d}} + B^\top \right) V_h. \quad (8)$$

304 The distance mask B filters out distant objects based on their 3D spatial proximity to the human.
 305 Specifically, for each object, we compute the Euclidean distance between the object's Gaussian center
 306 $c_{\text{obj}}^{\text{world}}$ and the human pelvis position $p_{\text{pelvis}}^{\text{world}}$ in world coordinates. B_{ij} encodes the relative distance
 307 between the i -th human token and the j -th object token:

$$310 \quad B_{ij} = \begin{cases} -\infty & \text{if } \|c_{\text{obj}}^{\text{world}} - p_{\text{pelvis}}^{\text{world}}\| \geq \tau, \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

312 where τ is set to the human arm length (derived from the SMPL-X model). When $B_{ij} = -\infty$, the
 313 corresponding object tokens are masked out during attention, effectively excluding non-interacting
 314 background objects.

316 This process yields updated features F'_{Human} and F'_{Object} that embed interaction cues. Finally, F'_{Human}
 317 is used to regress Δ SMPL-X refinements (body pose, hand pose), while F'_{Object} is used to predict
 318 ΔG_{object} , i.e., corrections for Gaussian-based object motion. In this way, the HOI module augments
 319 the baseline deformations (hexplane+LBS for humans and CHS for objects) with interaction-aware
 320 adjustments, enabling accurate reconstruction of human-object interaction scenes.

321 3.4 OPTIMIZATION

323 For background modeling, we employ the standard 3D Gaussian Splatting (3DGS) technique. During
 training, we isolate the background by masking out the object and human regions, allowing the static

324 Gaussian background to be optimized using a photometric loss. For human modeling, we regress
 325 the SMPL-X parameters by Moon et al. (2022), and incorporate an SMPL-X-based avatar model
 326 to ensure natural interaction with the object. For each frame, we extract the SMPL-X parameters
 327 and define a canonical T-pose human avatar. This canonical avatar is then deformed to match each
 328 frame using LBS. During training, image-based loss metrics such as SSIM, LPIPS, and L1-norm
 329 were utilized to compare the Gaussian renderer’s output with the human region in the image.

330 **Object Motion Optimization.** We model the motion of objects using CHS to ensure continuity in
 331 position interpolation. A CHS is a piecewise cubic polynomial that is defined by both the positions and
 332 the first derivatives (tangents) at key points in time. By specifying the starting and ending slopes for
 333 each spline segment, CHS guarantees smooth transitions between key frames, maintaining continuity
 334 not only in the object’s position but also in its velocity. In other words, the object’s trajectory over
 335 time remains continuous and smooth, without abrupt jumps or changes in speed. This property is
 336 crucial for accurately modeling temporal motion in a realistic and stable manner.

337 **Integrated Optimization.** We train our model using an integrated optimization objective that
 338 combines multiple loss terms. Specifically, the overall loss function is formulated as:

$$\mathcal{L} = \gamma \mathcal{L}_{\text{object motion}} + \beta \mathcal{L}_{\text{human}} + \sigma \mathcal{L}_{\text{scene}} + \mathcal{L}_{\text{depth}}, \quad (9)$$

341 where $\mathcal{L}_{\text{object motion}}$, $\mathcal{L}_{\text{human}}$, and $\mathcal{L}_{\text{scene}}$ are the loss components for the object’s motion, the human-
 342 related factors, and the scene context, respectively. Here, γ , β , and σ are hyperparameters that control
 343 the relative weight of each loss term during training. By tuning these hyperparameters, we balance the
 344 influence of each component on the training objective. This integrated optimization approach ensures
 345 that the model simultaneously accounts for object motion accuracy, human interaction plausibility,
 346 and scene consistency during learning.

348 4 EXPERIMENTS

350 4.1 IMPLEMENTATION DETAILS

352 We use ExAvatar (Moon et al. (2024)) as the baseline human rendering model, and all hyperparameters
 353 are kept identical to those used in ExAvatar. For object deformation using splines (Ahlberg et al.
 354 (2016); De Boor & De Boor (1978)), we fix the time interval to 4 for all scenes. Training is conducted
 355 using an NVIDIA H100 GPU, taking approximately 5 hours per scene.

356 4.2 DATASETS

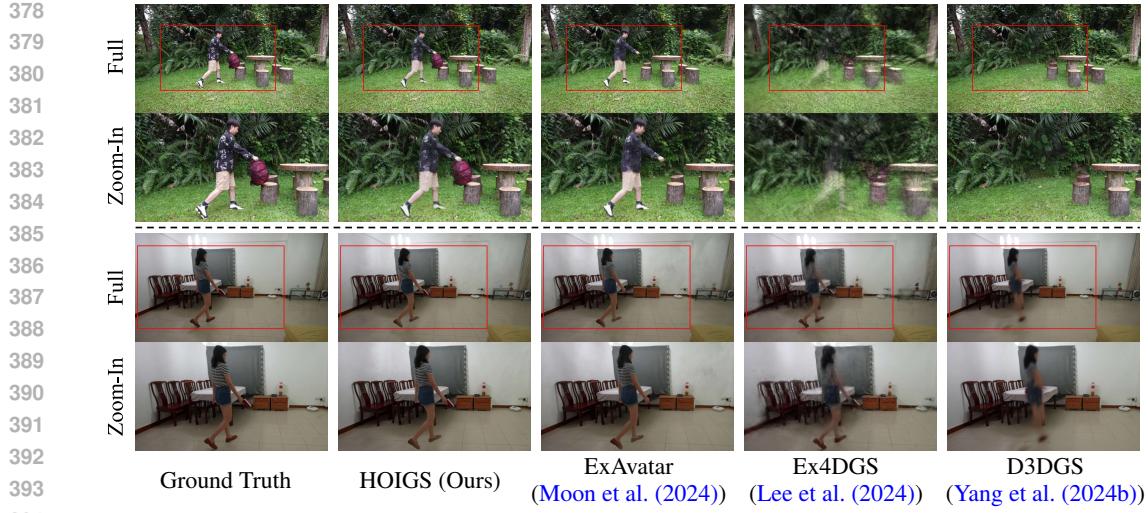
358 **HOSNeRF dataset (Liu et al. (2023)).** We use the monocular dynamic-scene dataset HOSNeRF,
 359 which captures human–object interaction scenarios. The dataset comprises recordings in six indoor
 360 and outdoor locations with six subjects interacting with objects within a single scenario. Each
 361 sequence contains 300–400 frames. For evaluation, we uniformly select 16 frames per sequence for
 362 testing and use the remaining frames for training, following HOSNeRF.

363 **BEHAVE dataset (Bhatnagar et al. (2022)).** We use the BEHAVE multi-view RGB-D human–object
 364 interaction dataset, but adapt it to a monocular setting by selecting a single fixed camera from the
 365 four static viewpoints for each sequence. Specifically, we curate 9 sequences covering four distinct
 366 indoor environments, five subjects, and four objects. From each sequence’s raw video, we uniformly
 367 sample 300 frames. For evaluation, we uniformly select 16 frames per sequence for testing and use
 368 the remaining frames for training.

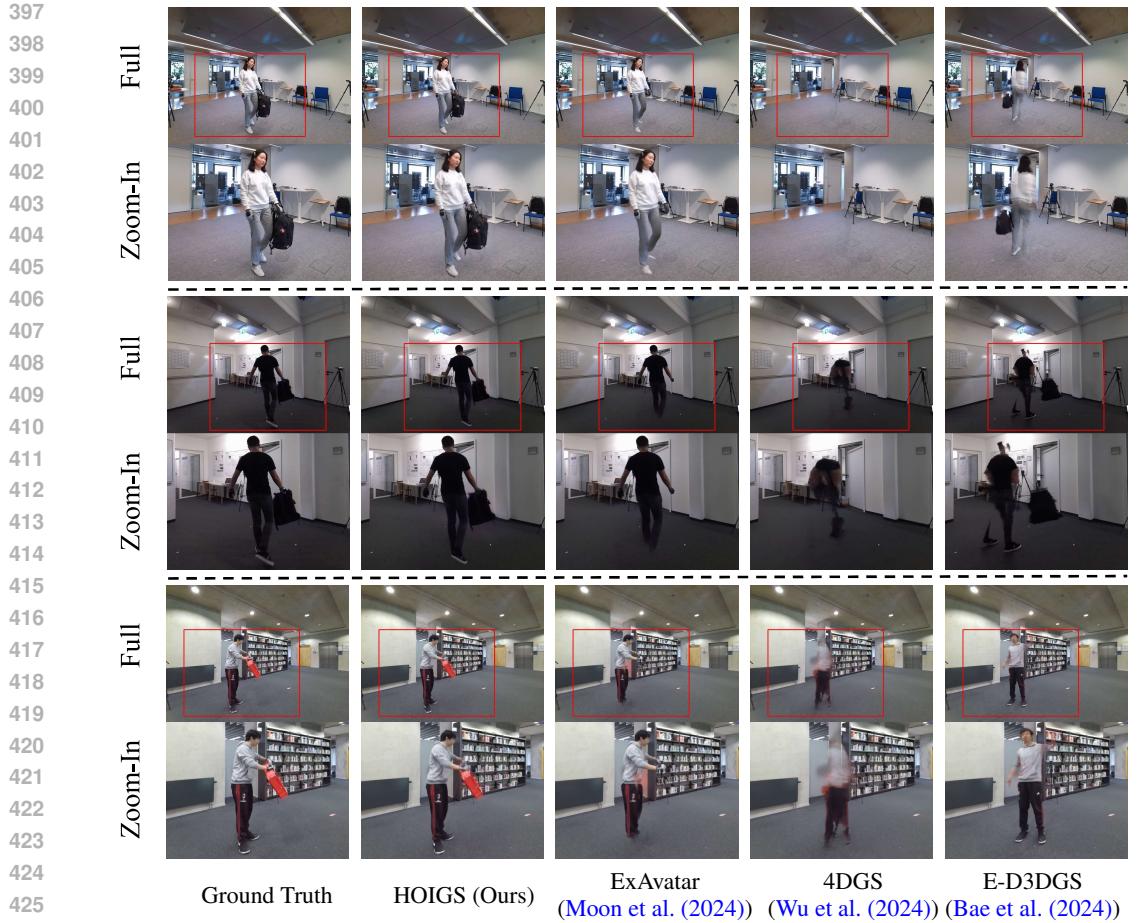
369 **ARCTIC dataset (Fan et al. (2023)).** We use the ARCTIC hand–object interaction dataset and
 370 extend comparisons to hand–object baselines. Since HOIGS is human-centric rather than hand-only,
 371 we evaluate only sequences where the full body is visible. Specifically, we use sequences of one
 372 subject interacting with four objects. Each monocular sequence (600 frames) is split by uniformly
 373 sampling 16 frames for testing and using the rest for training.

375 4.3 QUALITATIVE RESULTS

377 We compare our view-synthesis results with existing Gaussian-based models, which generally
 378 outperform NeRF-based methods in rendering quality. The experimental results are visualized



395 Figure 3: Qualitative comparison of reconstructed rendered view results on the HOSNeRF dataset.
396 We display the full-frame (top) rendering and a zoom-in (bottom) of the red Region of Interest (ROI).



427 Figure 4: Qualitative comparison of reconstructed rendered view results on the BEHAVE dataset.
428 We display the full-frame (top) rendering and a zoom-in (bottom) of the red Region of Interest (ROI).

429
430
431 in Fig. 3. The dynamic-scene models D3DGS (Yang et al. (2024b)) and Ex4DGS (Lee et al. (2024))
432 yield ghosting artifacts for both human and dynamic objects because they fail to disentangle human
433 and object motions within complex interactions. ExAvatar (Moon et al. (2024)) reconstructs humans



Figure 5: Qualitative comparison of reconstructed rendered view results on the ARCTIC dataset.

Methods	Backpack		Tennis		Suitcase		Playground		Dance		Lounge	
	PSNR↑	LPIPS↓										
K-Planes Fridovich-Keil et al. (2023)	19.05	0.557	19.31	0.536	18.64	0.602	17.92	0.635	18.17	0.623	24.21	0.453
D ² NeRF Wu et al. (2022)	20.52	0.608	23.97	0.540	20.99	0.645	21.23	0.616	19.92	0.647	27.13	0.509
Nerfies Park et al. (2021a)	19.56	0.559	22.12	0.443	19.01	0.555	21.14	0.533	19.37	0.524	25.90	0.342
HyperNeRF Park et al. (2021b)	19.62	0.587	21.26	0.510	19.41	0.607	21.67	0.578	19.30	0.601	27.25	0.332
NeuMan Jiang et al. (2022)	21.21	0.478	23.17	0.442	20.84	0.551	21.46	0.551	21.19	0.490	28.40	0.341
4DGS Wu et al. (2024)	24.49	0.192	26.57	0.162	17.98	0.460	24.34	0.222	21.34	0.212	30.50	0.067
D3DGS Yang et al. (2024b)	24.06	0.099	25.09	0.125	17.85	0.453	23.93	0.141	21.07	0.117	26.90	0.072
E-D3DGS Bae et al. (2024)	24.78	0.146	26.53	0.161	18.05	0.461	24.37	0.206	23.87	0.159	30.04	0.086
Ex4DGS Lee et al. (2024)	18.07	0.433	17.90	0.399	15.25	0.557	16.36	0.535	17.08	0.529	23.15	0.310
ExAvatar Moon et al. (2024)	24.15	0.107	23.57	0.160	20.32	0.260	25.30	0.129	23.32	0.170	29.43	0.048
HOSNeRF Liu et al. (2023)	22.56	0.243	24.15	0.320	21.74	0.382	22.67	0.336	22.63	0.248	27.74	0.227
HOIGS (Ours)	25.78	0.082	27.12	0.108	22.09	0.246	25.23	0.103	24.17	0.098	30.97	0.048

Table 1: Per-scene quantitative evaluation on the HOSNeRF dataset against baselines of our method. We color each cell as **best** and **second best**.

but does not handle dynamic objects. Our method accurately reconstructs humans and objects with temporally coherent motion, using CHS object trajectories with velocity vectors and the human backbone based on hexplane and LBS, while the HOI module further ensures contact consistency. On the ARCTIC dataset, as shown in Fig. 5, HOLD (Fan et al. (2024)) shows limited performance in full-body–object interactions, whereas HOIGS successfully reconstructs them. This is because HOLD reconstructs only hands, while HOIGS reconstructs the entire human body including the hands. On the BEHAVE dataset, as shown in Fig. 4, whereas ExAvatar suffers body–background overlap due to human misalignment in world space, our depth-based alignment ensures accurate human placement. Through qualitative results, we further confirm that our method effectively reconstructs complex human–object interactions with visually consistent outcomes.

4.4 QUANTITATIVE RESULTS

As shown in Tab. 1, HOIGS achieves the highest PSNR and the lowest LPIPS on the Backpack, Tennis, Suitcase, Dance, and Lounge scenarios of the HOSNeRF dataset, surpassing prior 3D Gaussian-based models in visual quality. Tab. 2 shows that on the BEHAVE dataset, it likewise attains the highest PSNR and lowest LPIPS, demonstrating effective reconstruction of complex human–object interactions from single-view input. Tab. 3 shows that on the ARCTIC dataset, our method outperforms the hand–object model HOLD (Fan et al. (2024)). Unlike HOLD, our model reconstructs complex full-body geometry while simultaneously capturing interactions with dynamic objects.

Methods	Backpack.1		Plasticcontainer.1		Plasticcontainer.2		Suitcase.1	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
4DGs Wu et al. (2024)	21.81	0.076	22.92	0.072	26.37	0.081	26.66	0.071
E-D3DGS Bae et al. (2024)	19.99	0.086	20.15	0.086	24.75	0.078	25.85	0.058
ExAvatar Moon et al. (2024)	27.86	0.041	29.96	0.042	30.11	0.038	30.86	0.032
HOIGS (Ours)	31.79	0.031	33.10	0.032	32.39	0.034	34.58	0.028
Methods	Backpack.2		Plasticcontainer.3		Backpack.3		Trashbin	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
4DGs Wu et al. (2024)	24.59	0.085	24.60	0.087	23.43	0.090	26.07	0.082
E-D3DGS Bae et al. (2024)	23.72	0.074	23.81	0.070	22.07	0.079	25.56	0.062
ExAvatar Moon et al. (2024)	26.47	0.054	26.71	0.056	25.78	0.038	29.81	0.029
HOIGS (Ours)	30.17	0.044	29.38	0.046	29.05	0.030	31.62	0.023

Table 2: Per-scene quantitative evaluation on the BEHAVE dataset against baselines. We color each cell as **best** and **second best**.

Methods	Capsulemachine		Box		Espressomachine		Mixer	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
4DGs Wu et al. (2024)	26.15	0.124	22.22	0.182	21.80	0.196	23.21	0.166
E-D3DGS Bae et al. (2024)	25.10	0.089	20.60	0.153	19.50	0.227	22.14	0.139
HOLD Fan et al. (2024)	25.52	0.522	24.72	0.494	23.52	0.547	23.35	0.540
HOIGS (Ours)	27.05	0.069	23.50	0.124	25.29	0.079	24.59	0.095

Table 3: Per-scene quantitative evaluation on the ARCTIC dataset against baselines of our method. We color each cell as **best** and **second best**.

	Avg (6 scenes)	
	PSNR↑	LPIPS↓
w/o CHS deformation (using MLP)	24.52	0.154
Baseline deformation	25.01	0.130
w/o human feature	25.67	0.119
w/o HOI module	25.24	0.128
HOIGS (Ours)	25.89	0.114

Table 4: Ablation studies on the HOSNeRF dataset using our method. The **best** results are highlighted.

4.5 ABLATION STUDY

We conduct ablation studies to validate the effectiveness of the proposed method. As shown in Tab. 4, modeling object deformation with a simple MLP yields the lowest performance, while our CHS-based baseline deformation improves PSNR by 0.5, demonstrating its superiority. Removing the HOI module and applying only velocity further results in a 0.6 drop in PSNR compared to the full model, confirming the necessity of explicitly modeling human–object interactions. Finally, replacing the time-varying hexplane features with simple parameter embeddings for the human features leads to a 0.2 decrease in PSNR, highlighting the effectiveness of our human feature design.

5 CONCLUSION

We presented HOIGS, a novel framework for reconstructing dynamic scenes with explicit modeling of human–object interactions from monocular videos. By combining hexplane-based human deformation, spline-based object motion, and an interaction-aware HOI module, our method achieves stable and accurate reconstruction even in challenging scenarios with contact and manipulation. In particular, the explicit treatment of human–object interactions enables our framework not only to recover realistic human geometry but also to faithfully capture object dynamics and their mutual influences, which have been largely overlooked in prior works. Extensive experiments on HOSNeRF, BEHAVE, and ARCTIC datasets demonstrate that HOIGS outperforms state-of-the-art human-centric and 4D Gaussian approaches in both visual quality and consistency, highlighting its effectiveness in advancing realistic modeling of complex human–object interactions.

Limitations and future works. While our framework handles typical dynamic motions well, it struggles under minimal camera movement, where COLMAP-based pose and point cloud estimation becomes unreliable. This often leads to rendering artifacts. Future work may improve robustness in such low-baseline settings by jointly optimizing camera poses during training.

540 REFERENCES
541

542 J Harold Ahlberg, Edwin Norman Nilson, and Joseph Leonard Walsh. *The Theory of Splines and Their*
543 *Applications: Mathematics in Science and Engineering: A Series of Monographs and Textbooks*,
544 Vol. 38, volume 38. Elsevier, 2016.

545 Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruc-
546 tion of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision*
547 *and Pattern Recognition*, pp. 1506–1515, 2022.

548 Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Per-gaussian
549 embedding-based deformation for deformable 3d gaussian splatting. In *European Conference on*
550 *Computer Vision*, pp. 321–335. Springer, 2024.

551 Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt,
552 and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions.
553 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
554 15935–15946, 2022.

555 Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang.
556 Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025.

557 Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene
558 generation from monocular videos. *NeurIPS*, 2024.

559 Carl De Boor and Carl De Boor. *A practical guide to splines*, volume 27. Springer New York, 1978.

560 Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and
561 Jerome Revaud. MAST3r-sfm: a fully-integrated solution for unconstrained structure-from-motion.
562 In *International Conference on 3D Vision 2025*, 2025. URL [https://openreview.net/](https://openreview.net/forum?id=5uw1GRBFoT)
563 forum?id=5uw1GRBFoT.

564 Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J
565 Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation.
566 In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.
567 12943–12954, 2023.

568 Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black,
569 and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects
570 from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*
571 *Recognition*, pp. 494–504, 2024.

572 Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo
573 Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of*
574 *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12479–12488, 2023.

575 Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar recon-
576 struction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the*
577 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12858–12868, 2023.

578 Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and
579 Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via
580 animatable 3d gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and*
581 *pattern recognition*, pp. 634–644, 2024a.

582 Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu,
583 Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model
584 for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis*
585 *and Machine Intelligence*, 2024b.

586 Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular
587 human videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*
588 *recognition*, pp. 20418–20431, 2024c.

594 Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying
 595 Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*,
 596 2025.

597

598 Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction
 599 of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*
 600 *Recognition*, pp. 3093–3102, 2020.

601 Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural
 602 human radiance field from a single video. In *European Conference on Computer Vision*, pp.
 603 402–418. Springer, 2022.

604

605 HyunJun Jung, Nikolas Brasch, Jifei Song, Eduardo Perez-Pellitero, Yiren Zhou, Zhihao Li, Nassir
 606 Navab, and Benjamin Busam. Deformable 3d gaussian splatting for animatable human avatars.
 607 *arXiv preprint arXiv:2312.15059*, 2023.

608

609 Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of
 610 human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern*
 611 *recognition*, pp. 7122–7131, 2018.

612

613 Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting
 614 for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

615

616 Sangmin Kim, Seunguk Do, and Jaesik Park. Showmak3r: Compositional tv show reconstruction. In
 617 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 864–874, 2025.

618

619 Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs:
 620 Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and*
 621 *pattern recognition*, pp. 505–515, 2024.

622

623 Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully explicit dynamic
 624 gaussian splatting. *Advances in Neural Information Processing Systems*, 37:5384–5409, 2024.

625

626 Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time
 627 dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
 628 *Pattern Recognition*, pp. 8508–8520, 2024a.

629

630 Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent
 631 gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF confer-*
 632 *ence on computer vision and pattern recognition*, pp. 19711–19722, 2024b.

633

634 Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang,
 635 Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single
 636 image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
 637 pp. 8662–8672, 2023.

638

639 Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie,
 640 and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a
 641 single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.
 642 18483–18494, 2023.

643

644 Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian:
 645 Fast and high-quality reconstruction of multiple human avatars. In *Proceedings of the 32nd ACM*
 646 *International Conference on Multimedia*, pp. 1120–1129, 2024.

647

648 Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl:
 649 A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries,*
 650 *Volume 2*, pp. 851–866. 2023.

651

652 Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementa-
 653 tion of Instance Segmentation and Object Detection algorithms in PyTorch. <https://github.com/facebookresearch/maskrcnn-benchmark>, 2018.

648 Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and
 649 Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications*
 650 *of the ACM*, 65(1):99–106, 2021.

651 Aymen Mir, Arthur Moreau, Helisa Dhamo, Zhensong Zhang, and Eduardo Pérez-Pellitero. Gaspacho:
 652 Gaussian splatting for controllable humans and objects. *arXiv preprint arXiv:2503.09342*, 2025.

653 Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for
 654 whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF conference on computer*
 655 *vision and pattern recognition workshops*, pp. 2308–2317, 2022.

656 Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar.
 657 In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

658 Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz,
 659 and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the*
 660 *IEEE/CVF international conference on computer vision*, pp. 5865–5874, 2021a.

661 Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman,
 662 Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for
 663 topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021b.

664 Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios
 665 Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single
 666 image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
 667 pp. 10975–10985, 2019.

668 Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei
 669 Zhou. Neural body: Implicit neural representations with structured latent codes for novel view
 670 synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision*
 671 and pattern recognition, pp. 9054–9063, 2021.

672 Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar:
 673 Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF*
 674 *conference on computer vision and pattern recognition*, pp. 5020–5030, 2024.

675 Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham
 676 Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev
 677 Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer.
 678 Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL
 679 <https://arxiv.org/abs/2408.00714>.

680 Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li.
 681 Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings*
 682 *of the IEEE/CVF international conference on computer vision*, pp. 2304–2314, 2019.

683 Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned
 684 implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF*
 685 *conference on computer vision and pattern recognition*, pp. 84–93, 2020.

686 Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar:
 687 Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings*
 688 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2059–2069,
 689 2024.

690 Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-
 691 Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video.
 692 In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pp.
 693 16210–16220, 2022.

694 Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian,
 695 and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings*
 696 *of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20310–20320, 2024.

702 Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D[^] 2nerf:
703 Self-supervised decoupling of dynamic and static objects from a monocular video. *Advances in*
704 *neural information processing systems*, 35:32653–32666, 2022.

706 Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed hu-
707 mans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern*
708 *Recognition (CVPR)*, pp. 13286–13296. IEEE, 2022.

710 Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed
711 humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer*
712 *vision and pattern recognition*, pp. 512–523, 2023.

714 Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai:
715 Adapting segment anything model for zero-shot visual tracking with motion-aware memory, 2024a.
716 URL <https://arxiv.org/abs/2411.11922>.

718 Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything:
719 Segment anything meets videos, 2023a.

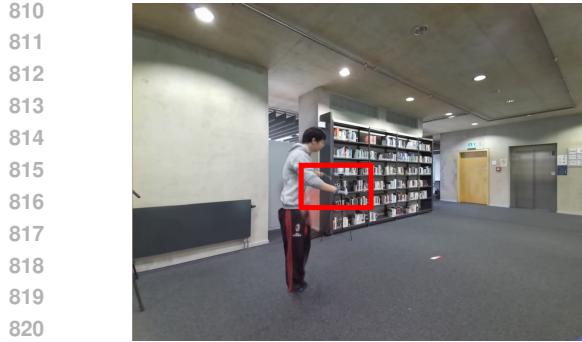
721 Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene represen-
722 tation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023b.

725 Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable
726 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the*
727 *IEEE/CVF conference on computer vision and pattern recognition*, pp. 20331–20341, 2024b.

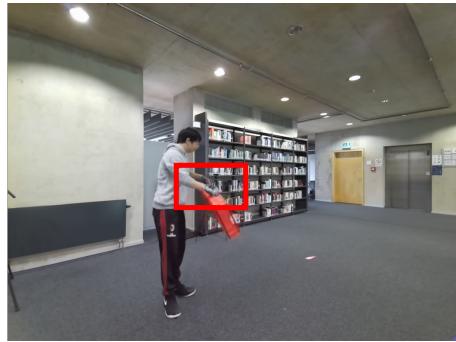
756 **6 APPENDIX**
757758 **STATEMENT ON THE USE OF LARGE LANGUAGE MODELS**
759760 In the interest of transparency and in compliance with the ICLR 2026 guidelines, we report that a
761 large language model (LLM) was used to assist in the refinement of this paper’s text.
762763 **Scope of Use.** The model’s role was strictly limited to that of a writing assistant. Its contributions
764 include:
765766

- 767 • Correcting grammatical errors, spelling, and punctuation.
- 768 • Improving sentence structure and flow for enhanced clarity.
- 769 • Refining word choices for greater precision and conciseness.

792 **Figure 6: Decomposed scene reconstruction.** Visualization of individual scene components demon-
793 strating geometric integrity in occluded regions. From left to right: human-only rendering, object-only
794 rendering, full scene rendering, and ground truth. Each row shows a different frame from the sequence,
795 highlighting that both human and object maintain coherent geometry even during close contact and
796 occlusion.
797798 **6.1 DECOMPOSED VISUALIZATION**
799800 **Component-level Reconstruction Quality.** To address concerns about reconstruction quality in
801 occluded regions, we provide decomposed visualizations that isolate individual scene components.
802 As shown in Figure 6, we render the human and object separately to demonstrate that each component
803 maintains geometric integrity even in regions with heavy occlusion or contact.
804805 **Human-only Rendering (Column 1):** The isolated human reconstruction shows coherent body
806 geometry throughout the sequence, including regions that were occluded by the object (backpack)
807 in the original footage. This demonstrates that our hexplane-based human deformation successfully
808 captures the complete body structure without artifacts from the interacting object.
809810 **Object-only Rendering (Column 2):** The object is reconstructed as a distinct, stable entity with
811 well-defined geometry. Unlike single-field approaches that often produce fused or melted geometry
812



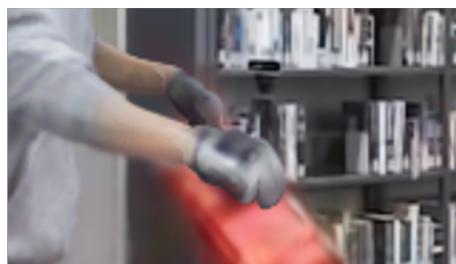
ExAvatar [ECCV 24]



Ours



ExAvatar [ECCV 24]



Ours

831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

Figure 7: Qualitative comparison of human pose refinement on the BEHAVE dataset. Visual comparison between ExAvatar and our method (HOIGS). The red boxes highlight the interaction regions (hands and objects). Our HOI module explicitly refines the hand and forearm poses by leveraging object motion features, leading to accurate contact modeling, whereas ExAvatar exhibits misalignment in these interaction-rich regions.

at contact points, our CHS-based object deformation maintains clear boundaries and structural consistency throughout the interaction.

Full Scene Rendering (Column 3): The combined rendering seamlessly integrates both components and closely matches the ground truth, confirming that our explicit modeling of human-object interactions through the HOI module enables accurate disentanglement while preserving realistic appearance.

These results validate that HOIGS does not simply overfit the combined RGB appearance but genuinely learns separate geometric representations for humans and objects. The clean separation at contact boundaries and the preservation of geometry in occluded regions demonstrate the effectiveness of our approach in handling complex interaction scenarios.

6.2 QUANTITATIVE EVALUATION OF HUMAN POSE ACCURACY

Geometric Fidelity Analysis. We acknowledge that rendering metrics alone are insufficient to fully validate the geometric fidelity of complex human–object interactions. To address this, we conducted additional evaluations on human pose accuracy using the BEHAVE dataset. We compare our method against ExAvatar using PA-MPJPE (Procrustes Aligned Mean Per Joint Position Error) and PA-PVE (Procrustes-Aligned Per Vertex Error).

Effect of the HOI Module on Pose Refinement. Our method explicitly models the mutual dependency between the human and the object. The HOI module leverages a cross-attention mechanism to use object motion features as contextual cues to refine human features. Specifically, as described in Eq. 14 of the main paper, the module predicts refinement offsets $\Delta\text{SMPL-X}$ for the body and hands. This capability allows the network to correct the human pose—even under partial occlusion—by inferring the likely body configuration from the object’s trajectory.

864	Model	Backpack1	Plasticcontainer1	Plasticcontainer2	Suitcase1
865	ExAvatar	0.4196 / 0.4628 / 0.3687	0.3875 / 0.4282 / 0.3563	0.3094 / 0.3145 / 0.2897	0.2654 / 0.2957 / 0.2505
866	HOIGS (Ours)	0.4177 / 0.4539 / 0.3656	0.2964 / 0.3344 / 0.2863	0.2973 / 0.3120 / 0.2837	0.2438 / 0.2649 / 0.2352
867	Model	Backpack2	Plasticcontainer3	Backpack3	Trashbin
868	ExAvatar	0.2690 / 0.3135 / 0.2488	0.3293 / 0.3298 / 0.2970	0.2177 / 0.2494 / 0.2092	0.2294 / 0.2597 / 0.2156
869	HOIGS (Ours)	0.2629 / 0.3068 / 0.2438	0.3270 / 0.3265 / 0.2948	0.2110 / 0.2380 / 0.2020	0.2263 / 0.2550 / 0.2126

Table 5: Unified quantitative evaluation on the BEHAVE dataset. The values in each cell correspond to **PA-MPJPE / PA-MPJPE (Hand/Forearm) / PA-PVE**. HOIGS consistently outperforms the baseline across these metrics, particularly in interaction-critical regions.

874	Method Combinations	Backpack	Tennis	Suitcase	Playground	Dance	Lounge	Average							
875		PSNR \uparrow LPIPS \downarrow													
876	Samurai + MetricV2	25.78	0.082	27.12	0.108	22.09	0.246	25.23	0.103	24.17	0.098	30.97	0.048	25.89	0.114
877	Samurai + Video Depth Anything	25.85	0.080	27.18	0.106	22.15	0.241	25.28	0.102	24.22	0.096	31.05	0.046	25.96	0.112
878	Samurai + DepthCrafter	25.72	0.088	27.08	0.108	22.06	0.249	25.20	0.109	24.08	0.099	30.93	0.048	25.85	0.117
879	TrackAnything + MetricV2	25.72	0.086	27.05	0.109	22.03	0.246	25.18	0.106	24.15	0.100	30.93	0.052	25.84	0.116
880	SAMv2 + Video Depth Anything	26.01	0.076	27.38	0.103	22.33	0.241	25.47	0.100	24.42	0.095	31.20	0.041	26.14	0.109
881	MaskRCNN + MetricV2	25.33	0.099	26.67	0.125	21.66	0.257	24.78	0.117	23.69	0.110	30.52	0.065	25.44	0.129

Table 6: Sensitivity analysis of HOIGS on the HOSNeRF dataset using different combinations of segmentation and depth estimation priors. The results demonstrate the robustness of our method, with consistent performance across various modern priors and strong performance even with older baselines (MaskRCNN).

Quantitative Results. Table 5 summarizes the evaluation on the BEHAVE dataset. We report the average PA-MPJPE and PA-PVE across all test frames. Additionally, we provide a specific analysis for Hand and Forearm joints, which are the most critical regions for interaction tasks. As shown in Table 5, HOIGS consistently outperforms the baseline. Notably, we observe a larger performance gain in the **PA-MPJPE (Hand/Forearm joints)**. This indicates that our HOI module effectively refines the poses of interaction-related body parts, resulting in physically more accurate reconstructions compared to ExAvatar, which lacks mutual feedback between the human and the object. Please refer to the per-sequence detailed tables at the bottom of the appendix.

6.3 SENSITIVITY ANALYSIS ON EXTERNAL MODULES

Robustness to External Priors. To address concerns regarding the reliance on external modules, we conducted a sensitivity analysis on the HOSNeRF dataset by evaluating our framework with various combinations of segmentation (e.g., Samurai Yang et al. (2024a), SAMv2 Ravi et al. (2024), MaskRCNN Massa & Girshick (2018), TrackAnything Yang et al. (2023a)) and depth estimation (e.g., Video Depth Anything Chen et al. (2025), MetricV2 Hu et al. (2024b), DepthCrafter Hu et al. (2025)) models. As shown in Table 6, HOIGS maintains highly consistent performance (Avg PSNR 25.8–26.1) across different modern priors, demonstrating that our method is robust to variations in preprocessing quality. Notably, even when employing the standard, older baseline of MaskRCNN combined with MetricV2, our model achieves an average PSNR of 25.44. This performance remains significantly higher than the state-of-the-art human-centric baseline, ExAvatar (Avg PSNR 24.35), and the 4DGS baseline, Ex4DGS (Avg PSNR 17.97).

6.4 COMPUTATIONAL COMPLEXITY AND RUNTIME ANALYSIS

Runtime Performance. We evaluate the computational efficiency of our method on the HOSNeRF dataset using a single NVIDIA H100 GPU. As shown in Table 7, our method achieves an inference speed of **44.27 FPS**. While this is slightly lower than 4DGS Wu et al. (2024) (61.04 FPS), it remains comparable to Ex4DGS Lee et al. (2024) (46.38 FPS) and outperforms D3DGS Yang et al. (2024b) (37.79 FPS). This result confirms that the inclusion of the HOI attention mechanism does not create a significant bottleneck, allowing our method to comfortably support real-time applications.

Complexity Analysis. The efficiency of our HOI module stems from the token-based architectural design. The cross-attention is computed between M human part tokens (where $M = 16$ is fixed) and N object Gaussian tokens. Unlike standard self-attention which scales quadratically ($O(N^2)$), our cross-attention scales linearly ($O(M \cdot N)$) with respect to the number of object Gaussians.

Methods	Training Time	Inference Speed (FPS)
4DGs Wu et al. (2024)	40 min	61.04
Ex4DGs Lee et al. (2024)	2 hr 30 min	46.38
D3DGS Yang et al. (2024b)	3 hr	37.79
E-D3DGS Bae et al. (2024)	2 hr	54.71
HOIGS (Ours)	5 hr	44.27

Table 7: Runtime performance comparison on the HOSNeRF dataset. We report the approximate training time per scene and the inference speed in Frames Per Second (FPS). Our method maintains real-time performance (>30 FPS) despite the added complexity of interaction modeling.

Methods	Backpack PSNR↑ LPIPS↓	Tennis PSNR↑ LPIPS↓	Suitcase PSNR↑ LPIPS↓	Playground PSNR↑ LPIPS↓	Dance PSNR↑ LPIPS↓	Lounge PSNR↑ LPIPS↓	Average PSNR↑ LPIPS↓
MAS3R Prior	23.51 0.135	25.25 0.121	22.40 0.197	24.65 0.074	23.63 0.115	28.99 0.057	24.59 0.128
Depth Recon Prior	21.63 0.142	25.65 0.122	22.13 0.230	25.24 0.103	24.08 0.123	28.95 0.095	25.36 0.136
Diffusion Prior (Ours)	23.70 0.082	27.13 0.112	22.96 0.235	25.63 0.123	24.17 0.093	29.97 0.043	25.89 0.114

Table 8: Quantitative ablation study on Object Priors using the HOSNeRF dataset. We evaluate the effectiveness of our Diffusion Prior against MAS3R and Depth Reconstruction priors.

Furthermore, we utilize compact 32-dimensional embeddings for object motion features, which minimizes the memory footprint and matrix multiplication overhead during the forward pass.

Training Cost Justification. We acknowledge that our training time (~ 5 hours) is longer than the baselines. This is a deliberate trade-off to prioritize physical plausibility and interaction accuracy. Explicitly modeling mutual dependencies and backpropagating gradients through the attention mechanism requires more iterations. However, this cost is strictly confined to the offline training phase, ensuring that the final online user experience remains real-time.

6.5 ADDITIONAL ABLATION STUDIES

Impact of Object Diffusion Prior. To validate the effectiveness of our design choice, we investigate the impact of different geometric priors on the final reconstruction quality. We compare our proposed method, which utilizes a generative Diffusion Prior, against two alternative initialization strategies: (1) MAS3R Prior: Initialization using MAS3RDuisterhof et al. (2025), a state-of-the-art dense matching and reconstruction model. (2) Depth Reconstruction Prior: Initialization using standard monocular metric depth estimation.

Table 8 presents the quantitative comparison on the HOSNeRF dataset. Our method equipped with the Diffusion Prior achieves the highest average reconstruction quality (25.89 PSNR), outperforming the MAS3R prior (24.59 PSNR) and the Depth prior (25.36 PSNR). While discriminative approaches like MAS3R or metric depth estimation rely heavily on visible cues, they often struggle to reconstruct accurate geometry in the presence of heavy occlusions, a common occurrence in human-object interaction scenarios (e.g., hands covering objects). In contrast, the Diffusion Prior leverages generative knowledge to plausibly complete 3D geometry even in occluded or unseen regions. This holistic geometric initialization provides a more robust starting point for our Cubic Hermite Spline (CHS) deformation, leading to sharper rendering and more stable tracking throughout the dynamic sequence.

6.6 FEATURE EXTRACTION

Object feature. As shown in Fig. 8(a), we extract object features by leveraging the velocity vectors and embedding parameters of Gaussians at key frames. As shown in Fig. 8(b), each key frame’s velocity vector is applied to the CHS and jointly optimized with the baseline deformation as input features for the HOI module. In addition, a 29-dimensional learnable parameter is embedded for each key frame Gaussian, which is concatenated with the velocity vector to form the feature representation. The interpolated Gaussian features produced by CHS are then combined with the concatenated feature and time information, and projected through a shallow MLP, resulting in a 32-dimensional feature vector.

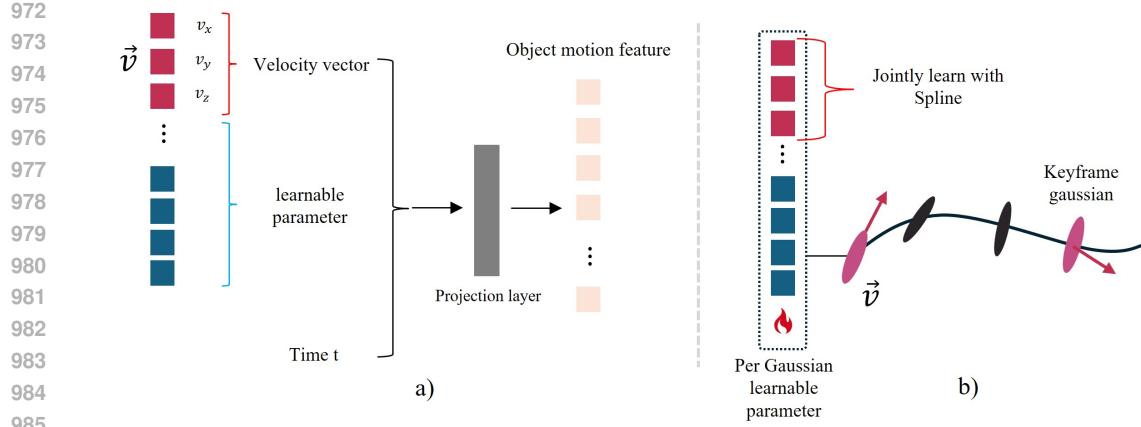


Figure 8: **Object feature extraction.** Extraction of object motion features using the embedded parameters and velocity vectors of each key frame.

Human feature. Fig. 9 illustrates the process of human feature extraction. We divide the SMPL-X model into 16 body parts and learn features corresponding to each part. Temporal features are sampled from the hexplane at SMPL-X vertices, where each feature at time t is obtained based on the coordinates (x_t, y_t, z_t) . For each body part, the features of its associated vertices are averaged to form the part-specific representation F_{human} :

$$F_{\text{part}} = \frac{1}{N} \sum_{i \in \text{part}} f_i(x_t, y_t, z_t), \quad (10)$$

where N denotes the number of vertices belonging to the part. As a result, 16 part features, including head, torso, arms, and legs, are obtained and used as inputs to the HOI module. This design captures temporally varying dynamic representations while preserving semantically meaningful features for individual body parts.

6.7 HOI MODULE NETWORK DETAIL

As shown in Fig. 10, the proposed HOI module takes the time-varying features of humans and objects as inputs and explicitly models their interactions. Let the human feature be denoted as $F_{\text{Human}} \in \mathbb{R}^{N_h \times d}$ and the object feature as $F_{\text{Object}} \in \mathbb{R}^{N_o \times d}$, where N_h and N_o are the numbers of feature tokens for human and object respectively, and d is the feature dimension.

To capture interdependencies between the two modalities, we apply a *mutual-attention* mechanism. Specifically, queries (Q), keys (K), and values (V) are obtained via learnable linear projections:

$$Q_h = F_{\text{Human}} W_h^Q, \quad K_o = F_{\text{Object}} W_o^K, \quad V_o = F_{\text{Object}} W_o^V, \quad (11)$$

$$Q_o = F_{\text{Object}} W_o^Q, \quad K_h = F_{\text{Human}} W_h^K, \quad V_h = F_{\text{Human}} W_h^V, \quad (12)$$

where $W_h^Q, W_h^K, W_h^V, W_o^Q, W_o^K, W_o^V \in \mathbb{R}^{d \times d}$ are learnable projection matrices.

Cross-attention is then computed in both directions: from human to object and from object to human. To enforce spatial priors, a distance mask $B \in \mathbb{R}^{N_h \times N_o}$ is added to the attention logits, where B_{ij} encodes the relative distance between the i -th human token and the j -th object token. The resulting attention maps are defined as:

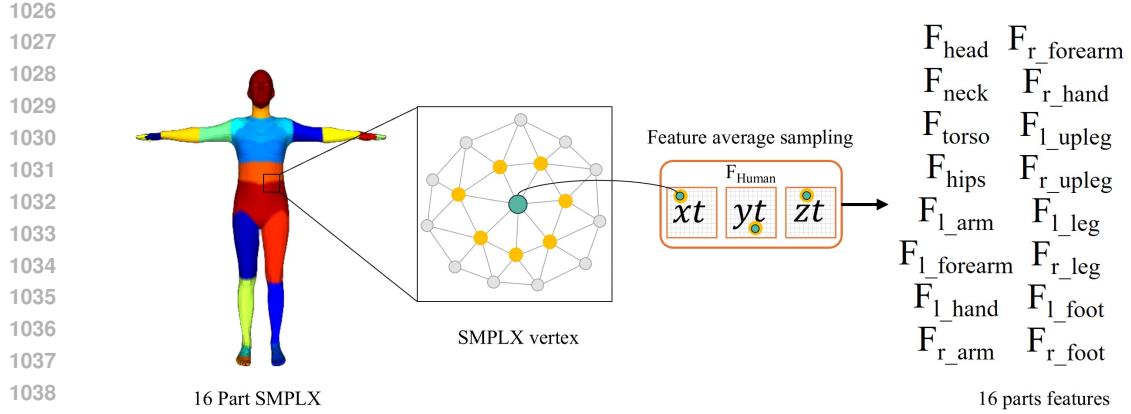
$$A_{h \leftarrow o} = \text{softmax} \left(\frac{Q_h K_o^\top}{\sqrt{d}} + B \right), \quad A_{o \leftarrow h} = \text{softmax} \left(\frac{Q_o K_h^\top}{\sqrt{d}} + B^\top \right). \quad (13)$$

Using these attention weights, the updated features are obtained as:

$$F'_{\text{Human}} = A_{h \leftarrow o} V_h, \quad F'_{\text{Object}} = A_{o \leftarrow h} V_o. \quad (14)$$

The updated human feature F'_{Human} is then fed into a small MLP head to regress the refinement terms of SMPL-X parameters:

$$\Delta \text{SMPL-X} = \{\Delta \theta_{\text{body}}, \Delta \theta_{\text{hand}}\}, \quad (15)$$

Figure 9: **Human feature extraction.**

where $\Delta\theta_{\text{body}}$ and $\Delta\theta_{\text{hand}}$ denote pose corrections for body and hands. Similarly, the updated object feature F'_{Object} is used to regress Gaussian-based object motion corrections:

$$\Delta G_{\text{object}} \in \mathbb{R}^{N_o \times 3}, \quad (16)$$

which represent displacement vectors applied to object Gaussians.

In this way, the HOI module augments the baseline deformations (hexplane+LBS for humans and CHS for objects) with interaction-aware refinements, enabling accurate reconstruction of complex human–object interaction scenes.

6.8 OBJECTIVE FUNCTION DETAILS

The overall loss function of our model is defined as follows:

$$L = \gamma L_{\text{object motion}} + \beta L_{\text{human}} + \sigma L_{\text{scene}} + L_{\text{depth}}, \quad (17)$$

where $L_{\text{object motion}}$, L_{human} , and L_{scene} correspond to losses for object motion, human modeling, and scene context, respectively. The weights γ , β , and σ control the relative importance of each loss term and are specifically set to 1.0, 0.5, and 0.25, respectively. In our approach, these three terms are optimized simultaneously to consistently model the interactions between humans and objects.

Human Loss details

The L_{human} term consists of losses related to human representation using the SMPL-X (Pavlakos et al. (2019)) model. Specifically, it includes the reprojection error between the 3D human joint positions and detected 2D keypoints in images, a mesh-based face loss enhancing the consistency of facial geometry and texture, and a Laplacian regularization term. Additionally, there is an L1 loss (L_{smp}) between the optimized SMPL-X parameters and the frame-wise initial SMPL-X parameters obtained by a regressor. These loss terms are directly adopted from previous methods such as ExAvatar (Moon et al. (2024)), without modifications. For example, the face loss optimizes the consistency between rendered facial images and actual facial images, ensuring geometry-texture coherence. Laplacian regularization is applied to enhance the stability of human body shape. Further details can be found in the referenced research.

Formally, the human loss is given by:

$$L_{\text{human}} = L_{\text{kpt}} + L_{\text{face}} + L_{\text{reg}} + 0.1 \times L_{\text{smp}}, \quad (18)$$

Scene Loss details

The L_{scene} term is a photometric loss focusing on the background regions of the entire scene, following the image similarity-based loss used in existing 3D Gaussian Splatting (Kerbl et al. (2023)) (3DGS) methods. Specifically, a pre-trained human/object segmentation model is employed to mask out human and object regions in the images, optimizing the background Gaussians for the remaining pixels only. This involves minimizing the difference between the rendered result and the background pixels

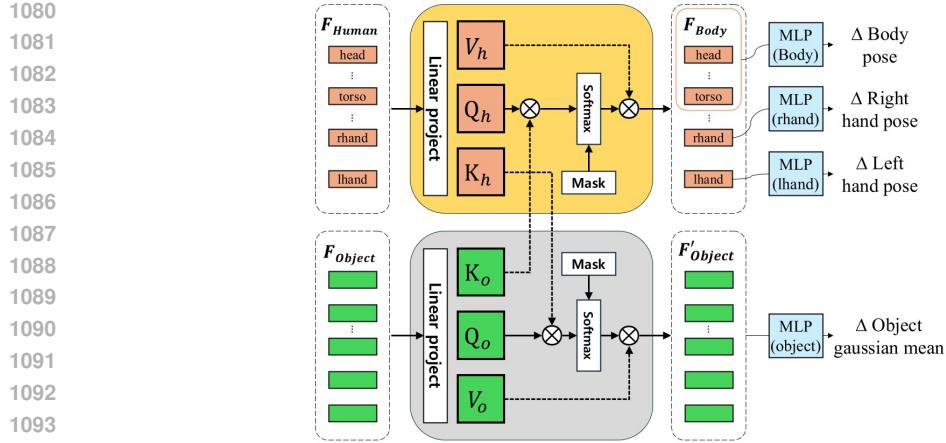


Figure 10: **Detailed HOI network.** The proposed architecture for estimating human-object interactions, leveraging features from human body parts and object Gaussian representations. The model takes as input human part features and per-Gaussian object features, processes them through bidirectional attention mechanisms to incorporate mutual contextual information, and outputs predictions for SMPL-X parameters per body part along with offset adjustments for object Gaussian centers.

excluding the segmented human and object areas. Occlusions frequently occur during interactions between human hands and objects, causing inconsistencies in masks. By optimizing humans, objects, and backgrounds simultaneously, our method effectively mitigates these boundary inconsistencies.

The scene loss is explicitly defined as:

$$L_{\text{scene}} = 0.8 \times L_1(I_{\text{gt}}, I_{\text{render}}) + 0.2 \times L_{\text{D-SSIM}}(I_{\text{gt}}, I_{\text{render}}), \quad (19)$$

Object Loss details

The L_{object} term is a photometric loss that focuses exclusively on the object regions within the scene. We render only the segmented object areas and compute the loss solely on these regions. A pre-trained object segmentation model is employed to isolate object masks in the input images. The object loss encourages accurate reconstruction and appearance consistency for moving objects, which often undergo significant deformation and motion. By supervising only the object regions, this loss helps to refine the geometry and texture of the object-specific Gaussians without being influenced by background or human-related elements.

The object loss is defined as:

$$L_{\text{object motion}} = 0.8 \times L_1(I_{\text{gt}}, I_{\text{obj}}) + 0.2 \times L_{\text{D-SSIM}}(I_{\text{gt}}, I_{\text{obj}}). \quad (20)$$