HOIGS:HUMAN-OBJECT INTERACTION GAUSSIAN SPLATTING FROM MONOCULAR VIDEOS

Anonymous authors
Paper under double-blind review

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

ABSTRACT

Reconstructing dynamic scenes with complex human–object interactions is a fundamental challenge in computer vision and graphics. Existing Gaussian Splatting methods either rely on human pose priors, neglecting dynamic objects, or approximate all motions within a single field, limiting their ability to capture interaction-rich dynamics. To address this gap, we propose Human-Object Interaction Gaussian Splatting (HOIGS), which explicitly models interaction-induced deformation between humans and objects through a cross-attention based HOI module. Distinct deformation baselines are employed to extract complementary motion features: hexplane for humans and Cubic Hermite Spline (CHS) for objects. By integrating these heterogeneous features, HOIGS effectively captures interdependent motions and improves deformation estimation in scenarios involving occlusion, contact, and object manipulation. Comprehensive experiments on multiple datasets demonstrate that our method consistently outperforms state-of-the-art human-centric and 4D Gaussian approaches, highlighting the importance of explicitly modeling human–object interactions for high-fidelity reconstruction. The video results of HOIGS are available at: `https://anonymous.4open.science/w/HOI-GS/`

1 INTRODUCTION

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

Recent approaches on human-centric video scene reconstruction (Kocabas et al. (2024); Moon et al. (2024); Hu et al. (2024b); Qian et al. (2024); Liu et al. (2024); Hu et al. (2024a); Wen et al. (2024); 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

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

Recent studies on 4D Gaussian Splatting extend beyond humans to encompass arbitrary moving objects, offering the advantage of general applicability. However, they generally exhibit lower reconstruction performance for humans compared to human-centric models. These approaches typically either define a canonical space and learn an implicit function that deforms it into the world coordinate system (Wu et al. (2024); Jung et al. (2023); Bae et al. (2024)), or explicitly parameterize object motions and optimize the corresponding parameters (Yang et al. (2023); Li et al. (2024); Lee et al. (2024)). Nevertheless, they do not explicitly model interactions between objects and instead treat all moving entities within a single motion field, which limits their ability to capture complex interactions. As a result, implicit methods struggle to represent long-term or highly non-linear 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.

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

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

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

In summary, our main contributions are as follows:

- We propose an entity-aware cross-attention based HOI module that explicitly enforces motion consistency between humans and objects. By attending across their motion features, the module captures interdependent dynamics and improves reconstruction in scenarios such as contact and object manipulation.

- We design distinct strategies for humans and objects using tailored deformation baselines. Hexplane encodes temporal and spatial features for human motion, while Cubic Hermite

Splines (CHS) embed velocity vectors and learnable parameters for objects. This separation enables accurate and expressive motion representations for both entities.

- We conduct extensive experiments on diverse human–object interaction scenes and demonstrate that our method achieves more accurate reconstruction compared to existing human-centric and 4D Gaussian approaches.

2 Related Works

2.1 Human Modeling

Research on realistic human modeling has long been pursued. Early parametric models enabled efficient estimation of human pose, exemplified by HMR (Kanazawa et al. (2018)), but struggled to capture clothing and accessories. To address this, implicit function-based methods (Huang et al. (2020); Saito et al. (2019; 2020); Xiu et al. (2022; 2023)) were proposed, which recover fine details such as hair and clothing but remain limited in global consistency and rendering efficiency. These methods mainly focused on human geometry with little attention to human-object interactions. With Neural Radiance Fields (NeRF) (Mildenhall et al. (2021)), several works applied it to human modeling (Peng et al. (2021); Jiang et al. (2022); Weng et al. (2022); Alldieck et al. (2022); Liao et al. (2023); Guo et al. (2023)), achieving realistic appearance and view consistency but still suffering from high training cost and slow rendering. In terms of human-object interactions, some attempts (Fan et al. (2024)) introduced objects, yet dynamic interactions were not fully captured. Recently, 3D Gaussian Splatting (3DGS) (Kerbl et al. (2023)) emerged as a new representation and has been applied to human reconstruction (Kocabas et al. (2024); Moon et al. (2024); Hu et al. (2024b); Liu et al. (2024); Hu et al. (2024a)). However, most efforts still regard objects as static. To overcome this, we propose HOIGS, a model for stable human reconstruction in dynamic scenes that explicitly captures human–object interactions.

2.2 Dynamic Scene Modeling

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

3 Method

As shown in Fig. 2, we reconstruct the scene by independently modeling the deformations of humans and objects, and then incorporating interaction-aware transformations through the HOI module. Object deformations are estimated using a Cubic Hermite Spline (CHS). Human deformations are based on hexplane features, where time-invariant spatial features are used to learn the texture of the 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 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,

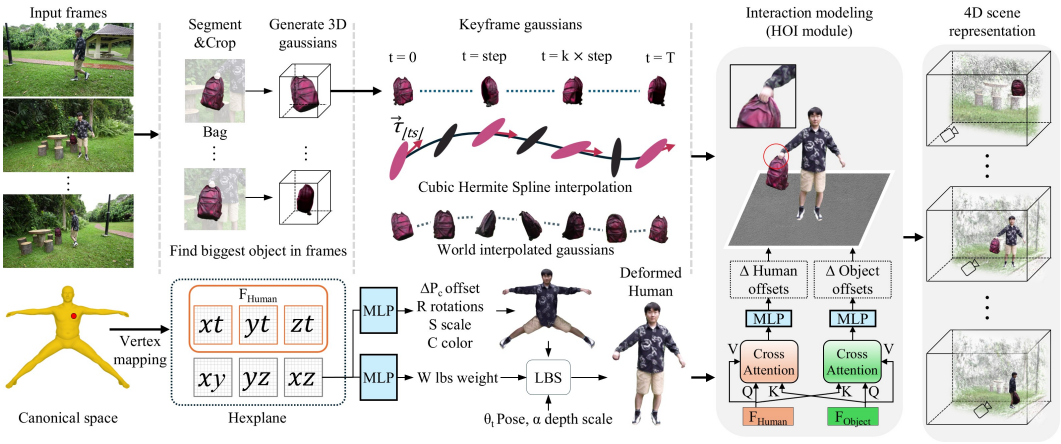

Figure 2: **Overview of the Proposed Framework.** Given an input video sequence, we first extract object-specific information, which is then used to reconstruct the 3D object shape via a diffusion prior. Based on the reconstructed shape, we initialize 3D Gaussians for each keyframe and use spline-based 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.

which accounts for interaction-driven transformations and determines the final positions of humans and objects in the reconstructed interaction scene.

3.1 Object Deformation

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:

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

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

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

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 $\tau_{\lfloor t_s \rfloor} = \frac{1}{2}\left(m_{\lfloor t_s \rfloor+1} - m_{\lfloor t_s \rfloor-1}\right)$. 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

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:

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

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

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.

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

$$\mathcal{L}_{depth} = \|D_{render} - D\|_1 , \tag{5}$$

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

3.3 HOI MODULE

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

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

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

$$Q_h = F_{\text{Human}} W_h^Q, \;\; K_o = F_{\text{Object}} W_o^K, \;\; V_o = F_{\text{Object}} W_o^V, \tag{6}$$

$$Q_o = F_{\text{Object}} W_o^Q, \;\; K_h = F_{\text{Human}} W_h^K, \;\; V_h = F_{\text{Human}} W_h^V . \tag{7}$$

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

	Backpack		Tennis		Suitcase		Playground		Dance		Lounge		Avg.	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	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	19.55	0.568
D^2NeRF (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	22.29	0.594
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	21.18	0.493
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	21.42	0.536
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	22.71	0.476
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	24.20	0.219
D3DGS (Yang et al. (2024))	24.06	0.099	25.09	0.125	17.85	0.453	23.93	0.141	21.07	0.117	26.90	0.072	23.15	0.168
ED3DGS (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	24.61	0.203
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	17.97	0.461
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	24.35	0.146
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	23.58	0.293
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	25.89	0.114

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

$$A_{h\leftarrow o} = \text{softmax}\left(\frac{Q_h K_o^\top}{\sqrt{d}} + B\right), \quad A_{o\leftarrow h} = \text{softmax}\left(\frac{Q_o K_h^\top}{\sqrt{d}} + B^\top\right). \tag{8}$$

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

3.4 Optimization

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 Gaussian background to be optimized using a photometric loss. For human modeling, we regress the SMPL parameters (Loper et al. (2023)), and incorporate an SMPL-X-based avatar model to ensure natural interaction with the object. For each frame, we extract the SMPL-X parameters and define a canonical T-pose human avatar. This canonical avatar is then deformed to match each frame using LBS. During training, image-based loss metrics such as SSIM, LPIPS, and L1-norm were utilized to compare the Gaussian renderer's output with the human region in the image.

Object Motion Optimization

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

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

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

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-related factors, and the scene context, respectively. Here, γ, β, and σ are hyperparameters that control the relative weight of each loss term during training. By tuning these hyperparameters, we balance the influence of each component on the training objective. This integrated optimization approach ensures that the model simultaneously accounts for object motion accuracy, human interaction plausibility, and scene consistency during learning.

Figure 3: **Qualitative comparison of reconstructed rendered view results on the HOSNeRF dataset.**

4 Experiments

4.1 Implementation Details

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

4.2 Datasets

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

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

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

4.3 Qualitative results

We compare our view-synthesis results with existing Gaussian-based models, which generally outperform NeRF-based methods in rendering quality. The experimental results are visualized in Fig. 3. The dynamic-scene models D3DGS (Yang et al. (2024)) and Ex4DGS (Lee et al. (2024)) yield ghosting artifacts for both human and dynamic objects because they fail to disentangle human and object motions within complex interactions. ExAvatar (Moon et al. (2024)) reconstructs humans 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. 4, 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. 5, whereas ExAvatar suffers body–background overlap due to

Table 2: Per-scene quantitative evaluation on the ARCTIC dataset against baselines of our method.

	Box		Capsulemachine		Espressomachine		Mixer		Avg	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
4DGS (Wu et al. (2024))	22.22	0.182	26.15	0.124	21.80	0.196	23.21	0.166	23.35	0.167
ED3DGS (Bae et al. (2024))	20.60	0.153	25.10	0.089	19.50	0.227	22.14	0.139	21.84	0.152
HOLD (Fan et al. (2024))	**24.72**	0.494	25.52	0.522	23.52	0.547	23.35	0.540	24.28	0.526
Ours	23.50	**0.124**	**27.05**	**0.069**	**25.29**	**0.079**	**24.59**	**0.095**	**25.11**	**0.092**

Figure 4: **Qualitative comparison of reconstructed rendered view results on the ARCTIC dataset.**

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 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. Tab. 3 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.

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

Table 3: Per-scene quantitative evaluation on the BEHAVE dataset against baselines of our method.

	BACKPACK$_1$		PLASTICCONTAINER$_1$		PLASTICCONTAINER$_2$		SUITCASE$_1$		BACKPACK$_2$	
	PSNR↑	LPIPS↓	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	24.59	0.085
ED3DGS (Bae et al. (2024))	19.99	0.086	20.15	0.086	24.75	0.078	25.85	0.058	23.72	0.074
ExAvatar (Moon et al. (2024))	27.86	0.041	29.96	0.042	30.11	0.038	30.86	0.032	26.47	0.054
Ours	**31.79**	**0.031**	**33.10**	**0.032**	**32.39**	**0.034**	**34.58**	**0.028**	**30.17**	**0.044**

	PLASTICCONTAINER$_3$		PLASTICCONTAINER$_4$		BACKPACK$_3$		TRASHBIN		AVG	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
4DGS (Wu et al. (2024))	24.60	0.087	24.59	0.090	23.43	0.090	26.07	0.082	24.61	0.082
ED3DGS (Bae et al. (2024))	23.81	0.070	22.98	0.083	22.07	0.079	25.56	0.062	23.31	0.075
ExAvatar (Moon et al. (2024))	26.71	0.056	27.05	**0.042**	25.78	0.038	29.81	0.029	28.29	0.041
Ours	**29.38**	**0.046**	**27.50**	0.043	**29.05**	**0.030**	**31.62**	**0.023**	**31.06**	**0.034**

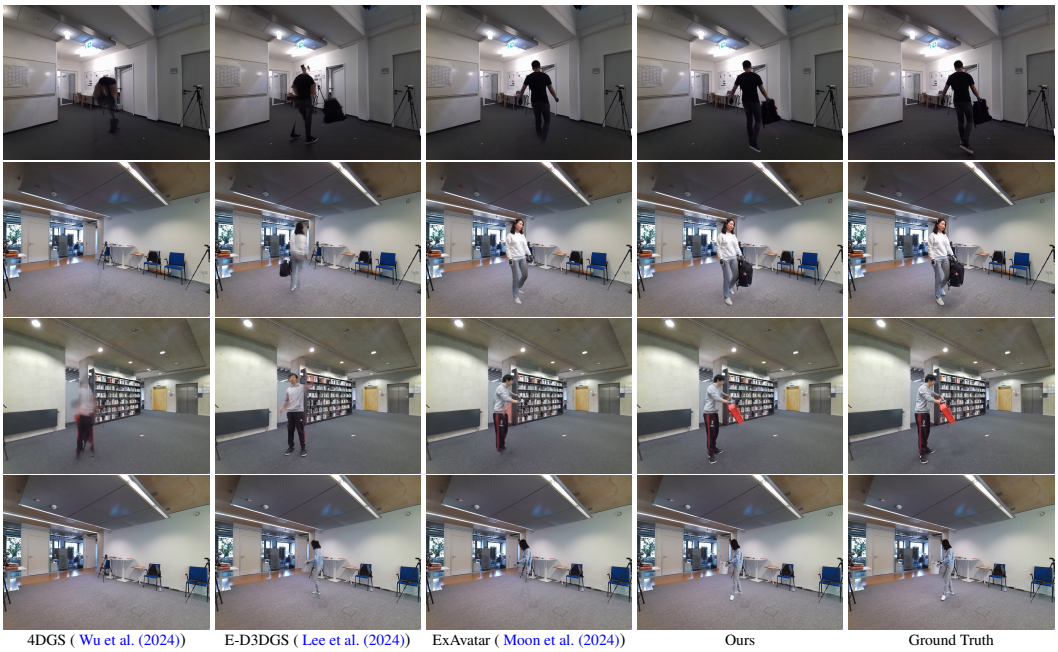

Figure 5: **Qualitative comparison of reconstructed rendered view results on the BEHAVE dataset.**

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

	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**

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.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6 Appendix

Statement on the Use of Large Language Models

In the interest of transparency and in compliance with the ICLR 2026 guidelines, we report that a large language model (LLM) was used to assist in the refinement of this paper's text.

Scope of Use. The model's role was strictly limited to that of a writing assistant. Its contributions include:

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

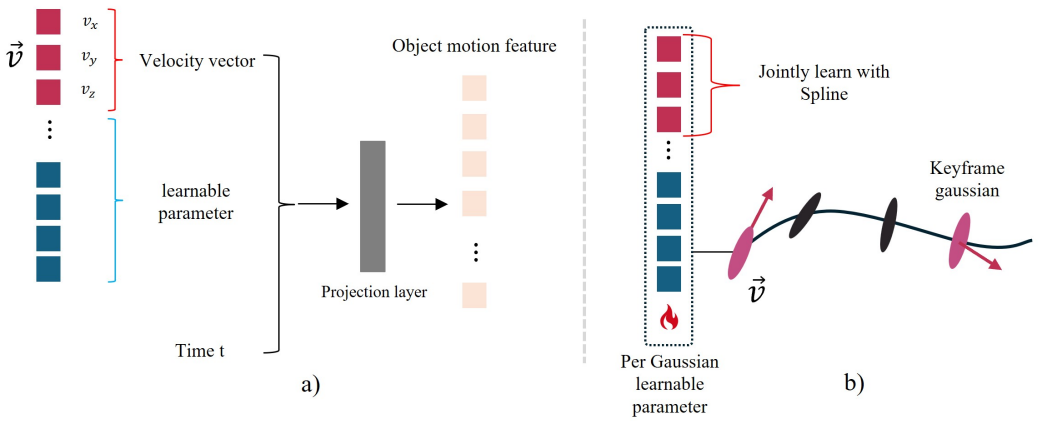

Figure 6: **Object feature extraction.** Extraction of object motion features using the embedded parameters and velocity vectors of each key frame.

6.1 Feature Extraction

Object feature As shown in Fig. 6(a), we extract object features by leveraging the velocity vectors and embedding parameters of Gaussians at key frames. As shown in Fig. 6(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.

Human feature Fig. 7 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}:

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

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.

6.2 HOI Module Network Detail

As shown in Fig. 8, 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:

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

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

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}

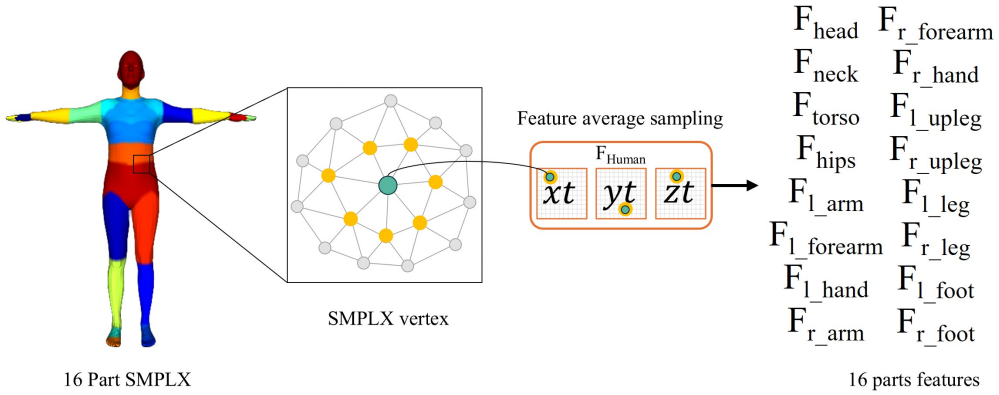

Figure 7: **Human feature extraction.**

encodes the relative distance between the i-th human token and the j-th object token. The resulting attention maps are defined as:

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

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

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

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

$$\Delta\text{SMPL-X} = \{\Delta\theta_{\text{body}},\ \Delta\theta_{\text{hand}},\ \Delta t\}, \tag{15}$$

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

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

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.

6.3 Objective Function Details

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

$$L = \gamma L_{\text{object motion}} + \beta L_{\text{human}} + \sigma L_{\text{scene}}, \tag{17}$$

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_{smplx}) 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

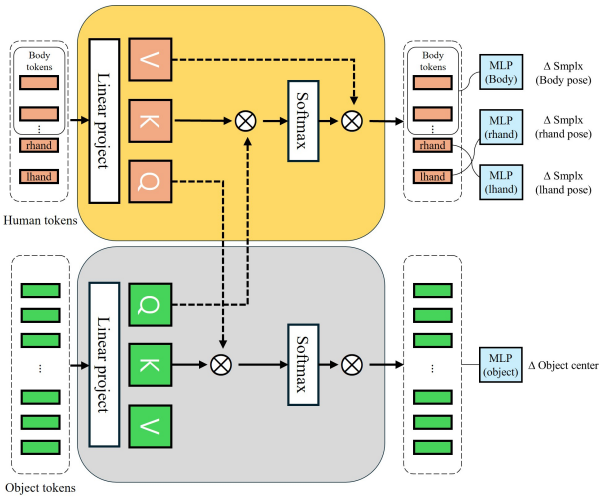

Figure 8: **Detailed HOI network.** The proposed architecture for estimating human-object interactions, leveraging features from human body parts and object Gaussian representations. The model takes as input human part features and per-Gaussian object features, processes them through bidirectional attention mechanisms to incorporate mutual contextual information, and outputs predictions for SMPL-X parameters per body part along with offset adjustments for object Gaussian centers.

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:

$$L_{\text{human}} = L_{\text{kpt}} + L_{\text{face}} + L_{\text{reg}} + 0.1 \times L_{\text{smplx}}, \tag{18}$$

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 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:

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

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:

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