2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Rendering Multi-Human and Multi-Object with 3D Gaussian Splatting

Weiquan Wang[1], Jun Xiao[1], Feifei Shao[1], Yi Yang[1], Yueting Zhuang[1], Long Chen[2]*

***Abstract*— Reconstructing dynamic scenes with multiple interacting humans and objects from sparse-view inputs is a critical yet challenging task, essential for creating high-fidelity digital twins for robotics and VR/AR. This problem, which we term Multi-Human Multi-Object (MHMO) rendering, presents two significant obstacles: achieving view-consistent representations for individual instances under severe mutual occlusion, and explicitly modeling the complex and combinatorial dependencies that arise from their interactions. To overcome these challenges, we propose MM-GS, a novel hierarchical framework built upon 3D Gaussian Splatting. Our method first employs a *Per-Instance Multi-View Fusion* module to establish a robust and consistent representation for each instance by aggregating visual information across all available views. Subsequently, a *Scene-Level Instance Interaction* module operates on a global scene graph to reason about relationships between all participants, refining their attributes to capture subtle interaction effects. Extensive experiments on challenging datasets demonstrate that our method significantly outperforms strong baselines, producing state-of-the-art results with high-fidelity details and plausible inter-instance contacts.**

I. INTRODUCTION

The development of intelligent robotic systems — capable of operating autonomously and safely in human-centric environments — is a central goal in robotics research [45], [4]. A cornerstone for these systems is the ability to perform safe and socially-aware navigation, a task that demands a deep, fine-grained understanding of the surrounding world and the complex interactions within it [3], [20], [1]. In applications ranging from assistive robotics to human-robot collaboration [6], [50], [39], navigation requires that a robot must not only avoid static obstacles but also comprehend and predict the dynamics of multiple humans interacting with objects. Creating high-fidelity, dynamic digital twins of real-world scenes serves as a critical foundation for this capability, enabling robust path planning, human-intent prediction, and advances in sim-to-real transfer [38], [16], [18]. A critical bottleneck, however, remains in faithfully capturing and rendering the underlying 3D scene, particularly in complex, dynamic environments where multiple humans and objects interact simultaneously. This challenge gives rise to the challenging task of **Multi-Human Multi-Object (MHMO) Rendering**: the high-fidelity reconstruction of dynamic scenes from sparse-view inputs, as dense camera setups are often impractical for real-world applications.

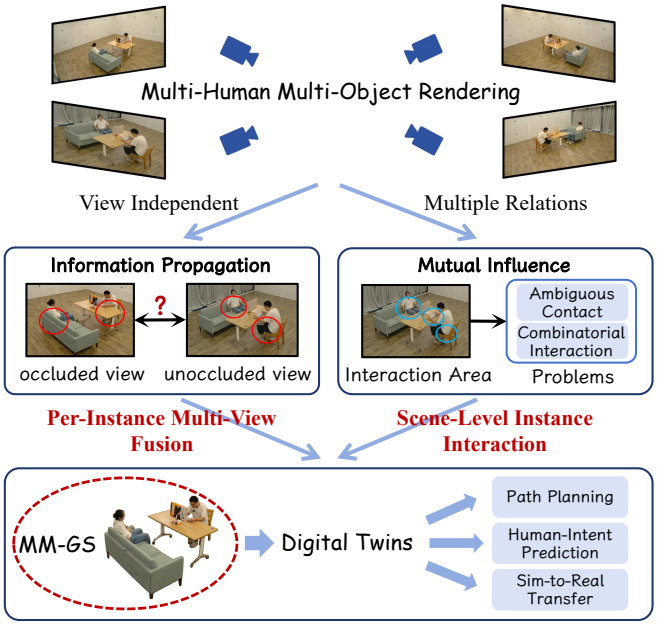

Fig. 1. **Core challenges in Multi-Human Multi-Object (MHMO) rendering.** From sparse views, rendering complex interactions involves overcoming two key challenges: ensuring cross-view consistency under severe occlusion (top) and modeling the mutual influence between instances at contact regions (bottom). Our MM-GS is designed to address both.

Tackling the MHMO rendering task presents two significant and coupled challenges, as illustrated in Fig. 1. **The first challenge** lies in the difficulty of *achieving a complete and view-consistent representation for each instance from sparse inputs*. In a typical MHMO scene, individuals and objects frequently occlude each other, leading to severe ambiguity [56]. The optimization pipelines in prevalent methods like NeRF [33] and 3DGS [19], typically process the information from each camera view independently. Such a common paradigm lacks an explicit mechanism to let informative, unoccluded views guide the reconstruction of occluded ones, *e.g.*, where geometry and appearance are poorly observed or fully occluded from other angles. This limitation often results in artifacts or geometric inconsistencies for a single instance. For example, appearing complete from one viewpoint but blurry or fragmented from another. Establishing a robust and

This work was supported by Key R&D Program of Zhejiang (2025C01128), the National Natural Science Foundation of China Young Scholar Fund Category B (62522216), Young Scholar Fund Category C (62402408), the National Natural Science Foundation of China (62441617, 62506333), Zhejiang Provincial Natural Science Foundation of China (No. LD25F020001), Fundamental Research Funds for the Central Universities (226-2025-00057), the Hong Kong SAR RGC General Research Fund (16219025), and Early Career Scheme (26208924), the China Postdoctoral Science Foundation (2025M781525), Postdoctoral Fellowship Program of CPSF (GZC20251077) and Zhejiang Province Postdoctoral Research Excellence Funding Project (ZJ2025065). (Corresponding author: Long Chen)

[1]Weiquan Wang, Jun Xiao, Feifei Shao, Yi Yang, and Yueting Zhuang are with the State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: wqwangcs@zju.edu.cn; junx@cs.zju.edu.cn; sff@zju.edu.cn; yangyics@zju.edu.cn; yzhuang@zju.edu.cn).

[2]Long Chen is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong (e-mail: longchen@ust.hk).

coherent per-instance representation is a critical prerequisite before their complex inter-dependencies can be modeled.

The second challenge, involves *explicitly modeling the subtle yet crucial dependencies between interacting instances*. An MHMO scene is far more than a simple collection of independent entities. Its realism hinges on capturing their mutual influence [56], [32]. These dependencies manifest primarily in resolving ambiguity at contact regions and creating subtle appearance adjustments due to proximity. A dedicated mechanism is required to refine the geometry and appearance of Gaussians at these contact surfaces to produce sharp, clear boundaries. Furthermore, close proximity between instances can introduce subtle visual effects like localized shadowing or slight color bleeding that must be captured. Existing works all focus on single human-object pairs [13], [43], [55], and they can only model one such relational link. They lack a mechanism to resolve the combinatorial complexity of MHMO scenes, where multiple humans and objects form a dense graph of potential relationships that must be reasoned about collectively to produce a coherent and realistic final rendering.

To overcome these two challenges, we propose **MM-GS**, a novel hierarchical framework built upon the efficiency and quality of 3DGS. Our pipeline is specifically designed to address the aforementioned problems in a coarse-to-fine manner, starting from strong human geometric priors (*e.g.*, SMPL [31]) and object templates. Specifically, we first align these canonical models to their target poses within each frame and then initialize a 3D Gaussian at each vertex of the resulting meshes. This process provides a robust initial representation for the entire scene. From this starting point, we tackle the first challenge of achieving view-consistency by introducing a ***Per-Instance Multi-View Fusion*** module. Unlike standard pipelines that treat views independently, this module employs a graph attention network to create an explicit information flow between different camera views of the same instance. This allows well-observed viewpoints to inform the reconstruction of occluded or poorly-defined regions, ensuring a complete and coherent representation for each individual participant. Subsequently, to address the second challenge of modeling inter-instance dependencies, we introduce the ***Scene-Level Instance Interaction*** module. This component constructs a global scene graph that connects all interacting instances, allowing our model to collectively reason about their dense web of relationships. By propagating information across this graph, MM-GS refines the Gaussian attributes at contact surfaces to delineate sharp boundaries and render subtle, proximity-based appearance effects, yielding a more realistic depiction of the interaction.

In summary, our main contributions are as follows:

- We tackle the novel and challenging problem of MHMO rendering from sparse-view inputs.
- We propose MM-GS, a hierarchical graph-based refinement framework that models complex MHMO scenes by decoupling per-instance view-consistent representation learning from scene-level interaction modeling.
- We design two specialized modules to explicitly address the key challenges in MHMO rendering: multi-view consistency and inter-instance dependencies.
- Our method achieves state-of-the-art performance on complex MHMO datasets, producing realistic and coherent digital twins of human-centric environments.

II. RELATED WORKS

Free Viewpoint Rendering. Free Viewpoint Rendering (FVR), the task of synthesizing novel views from a set of input images [49], [48], is a pivotal capability for robotics applications such as creating digital twins for simulation and policy learning [58], [22]. This field has been revolutionized by differentiable rendering, with 3DGS-based methods [19], [60], [29] offering a compelling balance of real-time speed and photorealism over prior methods like NeRF-based techniques [33], [59], [24]. However, existing 3DGS-based methods lack a dedicated mechanism for the fine-grained refinement of Gaussian primitives. This makes them struggle to resolve ambiguities from severe occlusions and inter-instance dependencies, a challenge severely exacerbated in the sparse-view setting [54]. We are the first to address this gap, proposing a hierarchical framework that explicitly models view-consistency and inter-instance interactions to handle complex MHMO scenes from sparse views.

Reconstruction of Dynamic and Human-Centric Scenes. Extending differentiable rendering techniques to dynamic environments is a significant area of research [36], [52], [17], [41]. A prevalent approach is to learn a deformation field that maps a static canonical representation to the observed dynamic state [34], [25], [23], [5]. For human subjects, this strategy is commonly combined with parametric body models like SMPL [31] to handle complex, non-rigid motion. This has enabled impressive reconstructions of single and multi-person scenes across various neural representations [48], [11], [9], [37], [28], [14]. Despite these advances, a crucial gap remains: the vast majority of existing methods treat each participant as a standalone, independent entity. They lack a mechanism to explicitly model the subtle yet crucial dependencies that arise from inter-person and person-object interactions. Our work is the first to address this gap by focusing on the high-fidelity reconstruction of complex scenes with a dense graph of inter-instance interactions.

Modeling Human and Object Interactions. For a scene to be realistic, it is crucial to accurately model complex interactions between instances [7], [40], [8], [43], [42], [53], [51], [2]. While a line of research on human-object interaction rendering has enabled high-quality results for isolated human-object pairs, these methods are fundamentally limited in scope [13], [43], [15], [46], [10], [27]. A prime example is NeuralDome [55], which relies on dense inputs and expensive volumetric rendering. Even recent advancements leveraging 3DGS and physical priors have focused exclusively on improving the realism of these single human-object pairs [46]. A fundamental limitation of all these methods is their singular focus on a single relationship. They are not designed to handle the combinatorial complexity of realistic scenes where a dense graph of relationships

exists between multiple humans and objects. In contrast, our work tackles this more general and challenging MHMO rendering problem, where a collective reasoning process over all participants is required to produce a coherent result.

III. PRELIMINARY

Our method is built upon the 3DGS [19] representation, which models a 3D scene as a collection of differentiable Gaussian primitives. This representation achieves state-of-the-art visual quality and real-time rendering speeds. Each Gaussian primitive is defined by a set of optimizable attributes $\mathcal{G} = \{\boldsymbol{\mu}, \mathbf{c}, \alpha, \mathbf{r}, \mathbf{s}\}$, which include its center $\boldsymbol{\mu} \in \mathbb{R}^3$, Spherical Harmonics (SH) coefficients for view-dependent color $\mathbf{c}$, opacity $\alpha \in \mathbb{R}$, a rotation representation $\mathbf{r}$ (e.g., a quaternion), and a scaling vector $\mathbf{s} \in \mathbb{R}^3$. The rotation and scale are used to form a covariance matrix $\boldsymbol{\Sigma}$.

To render an image, all 3D Gaussians are projected onto the 2D image plane. The color of a pixel $\mathbf{p}$ is then computed by blending the projected 2D Gaussians sorted by depth:

$$\hat{C}(\mathbf{p}) = \sum_{k \in \mathcal{N}} \mathbf{c}_k \sigma_k \prod_{l=1}^{k-1} (1 - \sigma_l), \quad \sigma_k = \alpha_k G'_k(\mathbf{p}), \tag{1}$$

where $\mathcal{N}$ is the set of sorted Gaussians overlapping pixel $\mathbf{p}$, and G'_k is the projected 2D Gaussian. In this work, we manipulate the attributes of Gaussian sets belonging to different instances. We will use the notation $\mathcal{G}_i^s$ to denote the full attribute set for instance i at stage s of our pipeline.

IV. METHODOLOGY

Given N sparse-view images $\{I_j\}_{j=1}^N$ of an MHMO scene with M instances, along with their camera parameters and instance masks, our objective is to reconstruct and render the scene with high fidelity. Our method aims to achieve both per-instance view-consistency and realistic inter-instance interactions. To this end, our MM-GS framework, illustrated in Fig. 2, introduces a hierarchical refinement process. First, we establish a robust initial representation by deforming canonical models to their target poses and initializing them with 3D Gaussians (Sec. IV-A). Furthermore, the **Per-Instance Multi-View Fusion** stage (Sec. IV-B) explicitly aggregates visual information from multiple viewpoints to ensure the completeness and consistency of each individual instance. Finally, the **Scene-Level Instance Interaction** stage (Sec. IV-C) employs a global scene graph to model the dependencies between all instances, which is crucial for resolving ambiguities at contact boundaries and capturing subtle interaction-driven appearance effects.

A. Human-Object Deformation

The first stage of our pipeline establishes a robust initial geometric representation for the MHMO scene. This involves transforming all M humans and objects from their canonical template spaces into their target poses, yielding the initial state (stage 0) of their Gaussian attributes, denoted as $\mathcal{G}_i^0$.

Human Deformation. For each human instance i, we leverage the SMPL [31] parametric body model. We initialize 3D Gaussians on the vertices of the canonical Da-pose SMPL mesh and then deform them to the target pose via a modulated Linear Blending Skinning (LBS) process [12], [35], [26], [27]. Specifically, the posed center $\boldsymbol{\mu}_i^0$ is computed from its canonical center $\boldsymbol{\mu}_i^c$ by applying a weighted average of each joint's transformation:

$$\boldsymbol{\mu}_i^0 = \sum_{k=1}^{K} w_k (\mathbf{R}_k \boldsymbol{\mu}_i^c + \mathbf{t}_k) + \mathbf{b}, \tag{2}$$

where w_k is the skinning weight for the k-th joint and $(\mathbf{R}_k, \mathbf{t}_k)$ is the joint's transformation. Moreover, the covariance matrix $\boldsymbol{\Sigma}^c$ is transformed by first computing an effective blended rotation matrix, $\mathbf{r}^0$, and then applying it to the canonical covariance:

$$\boldsymbol{\Sigma}^0 = \mathbf{r}^0 \boldsymbol{\Sigma}^c (\mathbf{r}^0)^{\mathrm{T}}, \quad \mathbf{r}^0 = \sum_{k=1}^{K} w_k \mathbf{R}_k. \tag{3}$$

While standard LBS effectively captures the overall pose, it often struggles with fine-grained details. To address this, we employ a small MLP, Φ_{lbs}, to predict a modulation vector $\mathbf{m}$ for the initial SMPL weights w_k^{SMPL}. The final, modulated skinning weight used in Eq. (2) is then computed as:

$$w_k = \texttt{softmax}(w_k^{\text{SMPL}} + m_k). \tag{4}$$

This approach yields a more detailed initial pose for each human, with the full attribute set denoted as $\mathcal{G}_i^0 = \{\boldsymbol{\mu}_i^0, \mathbf{c}_i^0, \alpha_i^0, \mathbf{r}_i^0, \mathbf{s}_i^0\}$, where attributes other than the center are initialized to generic values.

Object Deformation. For each object instance i in the MHMO scene, we assume it behaves as a rigid body. Our method directly utilizes the ground-truth pose information provided by the datasets [56], [30]. Specifically, we take the given rigid transformation, defined by a rotation matrix $\mathbf{R}_{\text{obj}}$ and a translation vector $\mathbf{T}_{\text{obj}}$, and apply it to the vertices of the object's canonical template mesh. This process accurately positions the object in the scene and defines its initial posed Gaussian centers $\boldsymbol{\mu}_i^0$. The remaining attributes are initialized to generic values to form the complete initial set $\mathcal{G}_i^0$.

B. Per-Instance Multi-View Fusion

The deformation stage provides a robust geometric foundation $\mathcal{G}_i^0$. However, the initial attributes $\{\mathbf{c}_i^0, \alpha_i^0, \mathbf{r}_i^0, \mathbf{s}_i^0\}$ are generic and non-expressive. As established in Sec. I, a key challenge in MHMO rendering from sparse views is the severe ambiguity caused by inter-instance occlusions. Consequently, our primary objective in this stage is to produce a complete and view-consistent representation for each instance. To achieve this, our ***Per-Instance Multi-View Fusion*** module intelligently aggregates visual information from all available viewpoints. This process operates independently on each instance i and updates its state from $\mathcal{G}_i^0$ to $\mathcal{G}_i^1$.

View-dependent Feature Construction. To enable context-aware fusion, we first construct a comprehensive feature representation for the instance's Gaussian set that encodes both its current 3D state and the visual evidence from each specific viewpoint. We begin by employing a 2D CNN-based image encoder Φ_{2D} to extract a multi-channel feature map

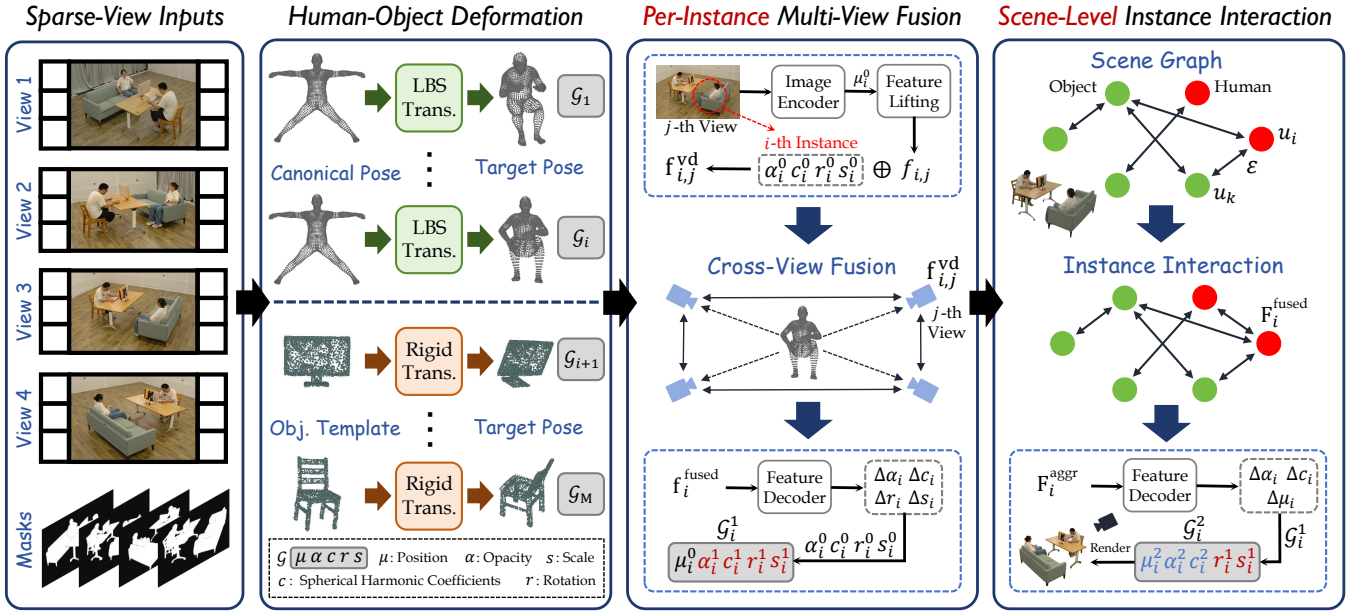

Fig. 2. **Overview of MM-GS pipeline.** Our method refines initial 3D Gaussian representations through three main stages. ***(a) Human-Object Deformation:*** We initialize the scene by deforming canonical human and object models to their target poses and representing them as collections of 3D Gaussians. ***(b) Per-Instance Multi-View Fusion:*** A Cross-View Fusion network refines each instance's appearance and local geometry by aggregating visual features from all its visible viewpoints, ensuring a view-consistent representation. ***(c) Scene-Level Instance Interaction:*** Finally, an Instance Interaction network operates on a global scene graph to model the dependencies between all participants, enabling a final refinement to capture interaction-driven effects.

$\mathbf{F}_j$ from each input image I_j. Then, we perform a *Point-wise Feature Lifting* operation, which projects the posed Gaussian centers $\boldsymbol{\mu}_i^0$ onto the image plane of view j to sample corresponding features, as formulated below:

$$\mathbf{f}_{i,j}^{\text{vis}} = \mathcal{S}(\mathbf{F}_j, \Pi(\boldsymbol{\mu}_i^0, \mathbf{K}_j, \mathbf{W}_j)), \tag{5}$$

where Π is the projection function and $\mathcal{S}$ is the bilinear sampling function. Finally, this collection of visual features is concatenated with the initial optimizable attributes to form a rich and view-dependent feature representation:

$$\mathbf{f}_{i,j}^{\text{vd}} = \text{Concat}(\mathbf{f}_{i,j}^{\text{vis}}, \{\mathbf{c}_i^0, \alpha_i^0, \mathbf{r}_i^0, \mathbf{s}_i^0\}). \tag{6}$$

Cross-View Fusion. The features $\{\mathbf{f}_{i,j}^{\text{vd}}\}_{j=1}^{N}$ constructed from individual views are merely isolated hypotheses. The core of our fusion process is to build a robust consensus by allowing these features to communicate and mutually refine one another. Specifically, for a Gaussian occluded in view j, this mechanism aggregates complementary appearance and geometric cues from visible context views $p \in \mathcal{V}_i$. We formulate this fusion as a single operation that transforms the set of per-view features $\{\mathbf{f}_{i,j}^{\text{vd}}\}_{j=1}^{N_{\text{ctx}}}$ into a unified, view-consistent feature representation $\mathbf{f}_i^{\text{fused}}$:

$$\mathbf{f}_i^{\text{fused}} = \frac{1}{N_{\text{ctx}}} \sum_{j=1}^{N_{\text{ctx}}} \left(\frac{1}{Z_{i,j}} \left(\mathbf{f}_{i,j}^{\text{vd}} + \gamma \sum_{p \in \mathcal{V}_i, p \neq j} \mathbf{f}_{i,p}^{\text{vd}} \right) \right), \tag{7}$$

where γ is a fusion factor, and $Z_{i,j}$ is a normalization term. $\mathbf{f}_i^{\text{fused}}$ serves as a robust target for the subsequent decoding.

Fused Feature Decoding. The feature $\mathbf{f}_i^{\text{fused}}$ now encodes robust, multi-view consistent information. The final step is to decode this abstract representation into concrete updates for the Gaussian attributes. To this end, an MLP-based decoder, Ψ_V, maps the fused feature to the final attribute updates:

$$(\Delta\mathbf{c}_i, \Delta\alpha_i, \Delta\mathbf{r}_i, \Delta\mathbf{s}_i) = \Psi_V(\mathbf{f}_i^{\text{fused}}). \tag{8}$$

The attributes of instance i are residually updated, yielding the new Gaussian set $\mathcal{G}_i^1 = \{\boldsymbol{\mu}_i^0, \mathbf{c}_i^1, \alpha_i^1, \mathbf{r}_i^1, \mathbf{s}_i^1\}$.

C. Scene-Level Instance Interaction

The multi-view fusion stage yields a set of individually refined and view-consistent instances, denoted as $\{\mathcal{G}_i^1\}_{i=1}^{M}$. However, these instances are modeled in isolation and thus lack awareness of each other. As established in Sec. I, another fundamental challenge in MHMO rendering is explicitly modeling the subtle yet crucial dependencies between interacting entities. These dependencies are key to resolving ambiguities at contact boundaries and capturing proximity-based visual effects. To address this, our ***Scene-Level Instance Interaction*** module operates on a dynamically constructed scene graph. This final stage updates the instance attributes from state $\mathcal{G}_i^1$ to their final state $\mathcal{G}_i^2$.

Scene Graph Construction. To facilitate reasoning about inter-instance relationships, we first construct a scene graph $G = (\mathcal{U}, \mathcal{E})$ for each frame. The nodes $\mathcal{U} = \{u_1, ..., u_M\}$ represent the M scene instances. Each node u_i is represented by the instance-level feature $\mathbf{f}_i^{\text{fused}}$ produced by the previous fusion stage. The edge set $\mathcal{E}$ defines the connectivity based on spatial proximity. For each instance i, we compute its axis-aligned bounding box $\mathcal{B}_i = [\min(\boldsymbol{\mu}_i^0), \max(\boldsymbol{\mu}_i^0)]$, which is defined by the component-wise minimum and maximum coordinates over the set of points. An undirected edge (u_i, u_p) is added to the edge set $\mathcal{E}$ if and only if the bounding

boxes of the two instances intersect, resulting in a graph that connects all potentially interacting instances:

$$(u_i, u_p) \in \mathcal{E} \iff \mathcal{B}_i \cap \mathcal{B}_p \neq \emptyset. \tag{9}$$

Interaction-Aware Feature Aggregation. With the scene graph established, we employ a graph attention network (GAT) [44] to propagate information between interacting instances. The GAT takes the set of node features $\{\mathbf{f}_i^{\text{fused}}\}_{i=1}^{M}$ as input and performs message passing along the graph edges. The network adaptively weighs the influence of neighboring instances via the attention mechanism, which learns to assign higher importance to neighbors that are physically close in contact, thereby explicitly encoding interaction dependencies. This process transforms the initial node features into a set of aggregated, interaction-aware features $\{\mathbf{f}_i^{\text{aggr}}\}_{i=1}^{M}$:

$$\{\mathbf{f}_i^{\text{aggr}}\}_{i=1}^{M} = \text{GAT}(\{\mathbf{f}_i^{\text{fused}}\}_{i=1}^{M}, G). \tag{10}$$

To improve efficiency, the subsequent decoding is only applied to a subset of actively interacting instances $\mathcal{U}_{\text{active}} = \{u_i \in \mathcal{U} \mid \deg(u_i) >= \tau_{\text{deg}}\}$, where $\deg(u_i)$ represents the degree of i-th node, and τ_{deg} is a pre-defined threshold.

Gaussian Attribute Decoding. The aggregated feature $\mathbf{f}_i^{\text{aggr}}$ encapsulates the necessary contextual information to refine the Gaussian attributes in a way that reflects the interaction. Specifically, a final MLP-based decoder, Ψ_I, takes the interaction-aware feature $\mathbf{f}_i^{\text{aggr}}$ and predicts a set of residual updates for the optimizable attributes:

$$(\Delta\boldsymbol{\mu}_i, \Delta\mathbf{c}_i, \Delta\alpha_i) = \Psi_I(\mathbf{f}_i^{\text{aggr}}). \tag{11}$$

The final Gaussian attributes for instance i are thus given by $\mathcal{G}_i^2 = \{\boldsymbol{\mu}_i^2, \mathbf{c}_i^2, \alpha_i^2, \mathbf{r}_i^1, \mathbf{s}_i^1\}$, where $\boldsymbol{\mu}_i^2 = \boldsymbol{\mu}_i^0 + \Delta\boldsymbol{\mu}_i$, $\mathbf{c}_i^2 = \mathbf{c}_i^1 + \Delta\mathbf{c}_i$, and $\alpha_i^2 = \alpha_i^1 + \Delta\alpha_i$. Note that the position update $\Delta\boldsymbol{\mu}_i$ is applied to the initial posed centers $\boldsymbol{\mu}_i^0$ (as they were unaltered in the fusion stage), whereas the appearance updates are applied to the already-refined $\mathbf{c}_i^1$ and α_i^1.

D. Hierarchical Refinement Strategy and Loss Function

Our framework's core design is a hierarchical strategy that decouples the optimization of different Gaussian attributes across stages. This staged refinement is key to resolving the complex MHMO rendering problem.

The ***Per-Instance Multi-View Fusion*** stage establishes a view-consistent representation for each instance in isolation by refining its appearance $(\mathbf{c}, \alpha)$ and local geometry $(\mathbf{r}, \mathbf{s})$ attributes. The Gaussian centers $(\boldsymbol{\mu})$ remain frozen during this stage for two key reasons: first, to preserve the strong geometric prior from the initial pose $\boldsymbol{\mu}^0$, and second, to provide stable spatial anchors for the feature lifting process.

The ***Scene-Level Instance Interaction*** stage then addresses the relationships between instances. Here, we introduce updates to the Gaussian centers $(\boldsymbol{\mu})$ to resolve physical ambiguities like contact and penetration, which is only possible once contextual information from neighboring instances is available. We also continue to refine appearance attributes $(\mathbf{c}, \alpha)$ to model effects like contact shadows. The local geometry attributes $(\mathbf{r}^1, \mathbf{s}^1)$ are kept fixed, preserving the high-quality shape learned in the previous stage.

The entire framework is trained end-to-end by minimizing a composite loss, $\mathcal{L}_{\text{render}}$, between the rendered image $\hat{I}_j$ (with final attributes $\mathcal{G}^2$) and the ground-truth image I_j:

$$\begin{aligned}\mathcal{L}_{\text{render}} =& \lambda_{\text{L1}}||\hat{I}_j - I_j||_1 + \lambda_{\text{SSIM}}(1 - \text{SSIM}(\hat{I}_j, I_j)) \\ &+ \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(\hat{I}_j, I_j),\end{aligned} \tag{12}$$

where the loss is a weighted combination of L1, SSIM [47], and LPIPS [57] terms.

V. Experiments

A. Experimental Setup

Datasets. We conduct a comprehensive evaluation of our method on two challenging datasets featuring complex multi-human and object interactions. ***HOI-M***^{3} [56] is a large-scale, high-quality dataset specifically designed for capturing interactions involving multiple humans and multiple objects in various indoor scenarios. Our evaluation focuses on representative sequences from three diverse daily scenarios: Livingroom, Fitnessroom, and Office. For these sequences, we select four challenging sparse views (8, 12, 17, and 20) for training our model, and evaluate the novel view synthesis quality on the remaining unseen views. ***CORE4D-Real*** [30] is a recent dataset focusing on collaborative object rearrangement, featuring two humans interacting with a certain object. The dataset was captured using a motion capture system and four allocentric cameras. Although its scenes are slightly less complex than those in HOI-M^3, it provides an effective testbed for evaluating the generalization capabilities of our method. We test our method on sequences from two distinct interaction scenarios involving a bucket and a box. To demonstrate our model's robustness in these extremely sparse settings, we train our model on three randomly selected views and test on the held-out fourth view. We report the averaged results across all four possible splits.

Implementation Details. Our framework is implemented in PyTorch and trained on a single NVIDIA A100 GPU using the Adam optimizer with a learning rate of 1×10^{-3}. ***Per-Instance Multi-View Fusion.*** We use a lightweight CNN for 2D feature extraction. For the cross-view propagation, the fusion factor γ is set to 0.1. The subsequent MLP decoder Ψ_V (2 hidden layers, [128, 64], ReLU) maps the fused features and initial Gaussian attributes to residual updates and a 64-dim instance feature for the next stage. We use 4 context views for this fusion process. ***Scene-Level Instance Interaction.*** Instance interactions are modeled by a 2-layer GAT with a hidden dimension of 64, 4 attention heads, and a dropout of 0.1. The pre-defined threshold τ_{deg} is set to 1. A final MLP decoder, Ψ_I, maps the aggregated 64-dim GAT feature to a 7-dim modulation vector for updating the color and opacity of the Gaussians.

Baselines. As our work is the first to tackle the comprehensive MHMO rendering task, no direct prior methods exist for comparison. Therefore, we establish two strong baselines by extending state-of-the-art methods from related domains to our setting, ensuring a fair comparison by providing them with the same input data. ***NeuralHOIFVV-MM.*** Our first

TABLE I

QUANTITATIVE COMPARISON FOR NOVEL VIEW SYNTHESIS ON THE HOI-M^3 DATASET. WE REPORT PSNR(↑), SSIM(↑), AND LPIPS(↓) ON THREE REPRESENTATIVE SCENES. WE USE RED AND YELLOW TO DENOTE THE BEST AND SECOND-BEST RESULTS.

Method	Livingroom			Fitnessroom			Office		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NeuralHOIFVV-MM [55]	21.33	0.8704	0.1884	22.66	0.9148	0.1575	21.65	0.8876	0.1702
GTU-MM [21]	20.82	0.8527	0.2045	21.91	0.9095	0.1683	21.71	0.8841	0.1734
Ours	22.47	0.8894	0.1722	23.24	0.9224	0.1504	22.89	0.9029	0.1633

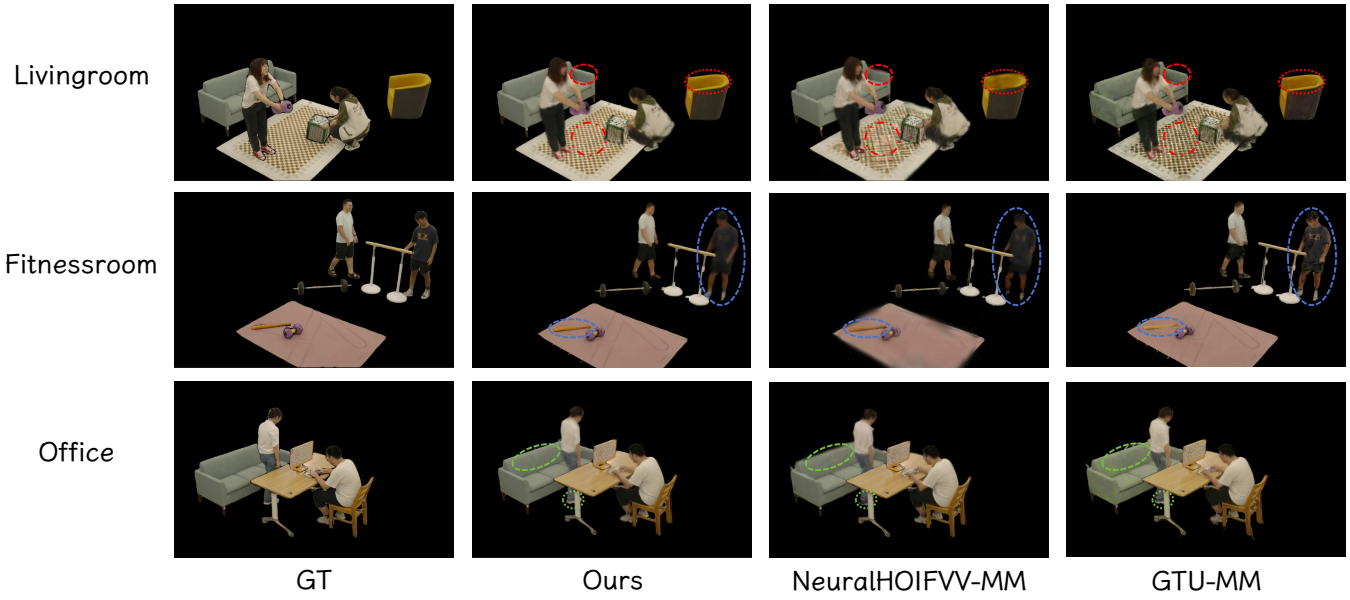

Fig. 3. **Qualitative comparison on the HOI-M^3 dataset.** We highlight specific regions with colored dashed circles to illustrate the differences. Note that our MM-GS generates significantly sharper details and more plausible contact regions. In contrast, the NeRF-based NeuralHOIFVV-MM tends to produce overly smooth or blurry results, while the 3DGS-based GTU-MM suffers from floating artifacts and geometric inconsistencies.

baseline is an extension of NeuralHOIFVV [55], a NeRF-based method for single human-object interaction rendering. We adapt its layered representation to our multi-instance setting by assigning an independent, pose-conditioned NeRF to each human and object in the scene. All layers are then rendered compositionally. ***GTU-MM.*** To provide a strong comparison based on the same underlying 3D representation, we adapt GTU [21], a SOTA 3DGS method for multi-human reconstruction. The original method focuses solely on humans. We extend its pipeline by introducing a parallel branch for objects, where we initialize their Gaussians from posed template meshes using the provided rigid transformations. The final scene, composed of all human and object Gaussians, is then jointly optimized using the same training objectives as the original method.

Metrics. To quantitatively evaluate the quality of rendered novel view and novel pose images, we evaluate the rendering quality using standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [47], and Learned Perceptual Image Patch Similarity (LPIPS) [57].

B. Rendering Evaluation

We evaluate our MM-GS against the baselines on both datasets, presenting quantitative metrics and qualitative visual comparisons.

TABLE II

GENERALIZATION PERFORMANCE ON THE CORE4D-REAL DATASET. RESULTS ARE AVERAGED OVER TWO INTERACTION SCENARIOS. NEU.-MM: NEURALHOIFVV-MM.

Method	Bucket			Box		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Neu.-MM	19.43	0.9347	0.0602	18.78	0.9218	0.0629
GTU-MM	19.07	0.9281	0.0644	18.56	0.9231	0.0637
Ours	20.08	0.9387	0.0527	19.22	0.9302	0.0598

Quantitative Results. As shown in Table I, our MM-GS consistently outperforms both baselines across all metrics on the challenging HOI-M^3 dataset. Notably, our approach achieves a significant improvement in PSNR and SSIM, while also attaining the lowest (best) LPIPS scores. The substantial margin over GTU-MM, which is also based on 3DGS, underscores the effectiveness of our hierarchical refinement process. The NeRF-based NeuralHOIFVV-MM tends to produce overly smooth results, and GTU-MM, lacking our explicit fusion and interaction modules, struggles to optimize the complex scenes, leading to lower

TABLE III

ABLATION STUDY OF CORE COMPONENTS ON THE HOI-M^3 DATASET (LIVINGROOM).

Method	PSNR↑	SSIM↑	LPIPS↓	FPS↑	Training
MM-GS (Full)	**22.47**	**0.8894**	**0.1722**	160+	∼40min
w/o View Fusion	20.98	0.8566	0.1927	160+	∼32min
w/o Interaction	21.51	0.8621	0.1893	160+	∼36min
w/o Both	19.42	0.8140	0.2031	160+	**∼30min**

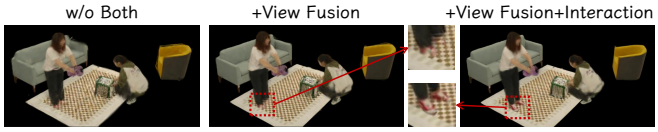

Fig. 4. **Qualitative results of our ablation study.** Removing both modules (w/o Both) leads to blurry results. Adding the View Fusion module (+ View Fusion) significantly improves sharpness. Further incorporating the Interaction network (+ View Fusion + Interaction) resolves ambiguities at contact regions, resulting in cleaner boundaries.

performance. Table II shows the generalization performance on the CORE4D-Real dataset. Even in this extremely sparse setting (training on only three views), our method continues to achieve the best results in both interaction scenarios.

Qualitative Results. The qualitative comparisons in Fig. 3 visually corroborate our quantitative superiority and highlight the specific advantages of our hierarchical design. Our method consistently generates renderings with sharp details, vibrant and consistent colors, and plausible contact regions. In contrast, NeuralHOIFVV-MM, as a NeRF-based approach, tends to produce overly smooth and blurry results, failing to capture high-frequency textures and struggling to define clear boundaries at human-object contact points. GTU-MM, while also based on 3DGS, suffers from inconsistent and mixed colors on the rendered instances. Our method avoids these issues by explicitly propagating rich visual features from the actual input views and aggregating multi-instance information, demonstrating that our proposed fusion and interaction networks are crucial for reconstructing high-fidelity MHMO scenes.

C. Ablation Study

To validate the effectiveness of our two core components, the Per-Instance Multi-View Fusion stage and the Scene-Level Instance Interaction stage, we conduct a comprehensive ablation study on the HOI-M^3 dataset. We evaluate four variants of our model: (1) our full MM-GS model; (2) our model without the Scene-Level Instance Interaction stage (denoted as 'w/o Interaction'); (3) our model without the Per-Instance Multi-View Fusion stage (denoted as 'w/o View Fusion'); and (4) a baseline version without both modules (denoted as 'w/o Both').

Quantitative Analysis. The results of our ablation study are presented in Table III. The baseline model without either of our proposed modules performs the worst across all rendering quality metrics. Our full model, which incorporates both modules, achieves the best performance by a clear margin, confirming the complementary nature of our two-stage refinement process. Either introducing the Per-Instance Multi-View Fusion or the Scene-Level Instance Interaction module brings significant gains over the baseline. Regarding computational efficiency, our full model maintains real-time rendering capabilities at over 160 FPS, demonstrating that our hierarchical design effectively balances high-fidelity interaction modeling with computational performance.

Qualitative Analysis. The qualitative results in Fig. 4 provide visual evidence for the function of each module. The baseline rendering is noticeably blurry and lacks detail. After incorporating the Per-Instance Multi-View Fusion module, the sharpness and overall clarity of the rendering are significantly improved However, ambiguities at contact points may still exist, as seen in the magnified region where the person's shoe and the carpet appear intermingled. Finally, by adding the Scene-Level Instance Interaction module, these contact boundaries become sharp and well-defined, which highlights its crucial role in reasoning about spatial relationships to produce a physically plausible and coherent scene.

VI. CONCLUSION

In this paper, we introduced and tackled the novel and challenging task of rendering complex Multi-Human Multi-Object (MHMO) interactions from sparse view inputs. We presented MM-GS, a novel hierarchical framework built upon 3D Gaussian Splatting that addresses this problem through a coarse-to-fine refinement strategy. Extensive experiments demonstrate that our approach significantly outperforms strong baselines adapted to this new task, achieving high fidelity with plausible contacts, thereby supporting high-quality digital twin creation for robotic simulation.

Limitations and Future Works. Our current scope focuses on common interaction scenarios involving articulated humans and largely rigid objects. Extending our modeling to more complex physical phenomena is a valuable avenue for future research. Moreover, our pipeline relies on available object poses and 2D masks, which pose a constraint on immediate in-the-wild robotic deployment.

REFERENCES

[1] P. Althaus, H. Ishiguro, T. Kanda, T. Miyashita, and H. I. Christensen, "Navigation for human-robot interaction tasks," in *ICRA*, 2004.

[2] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Behave: Dataset and method for tracking human object interactions," in *CVPR*, 2022.

[3] A. Bonci, P. D. Cen Cheng, M. Indri, G. Nabissi, and F. Sibona, "Human-robot perception in industrial environments: A survey," *Sensors*, 2021.

[4] H. Chen, S. Li, J. Fan, A. Duan, C. Yang, D. Navarro-Alarcon, and P. Zheng, "Human-in-the-loop robot learning for smart manufacturing: A human-centric perspective," *IEEE TASE*, 2025.

[5] Q. Chen, K. Qian, Z. Hu, Y. Tai, and Z. Yu, "H-rssg: High-fidelity robotic surgical scene generation with implicit deformable neural radiance field," *IEEE TASE*, 2025.

[6] A. Cherubini, G. Oriolo, F. Macrí, F. Aloise, F. Cincotti, and D. Mattia, "A multimode navigation system for an assistive robotics project," *Autonomous Robots*, vol. 25, no. 4, pp. 383–404, 2008.

[7] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, "High-quality streamable free-viewpoint video," *ACM TOG*, 2015.

[8] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi, "Motion2fusion: Real-time volumetric performance capture," *ACM TOG*, 2017.

[9] X. Gao, J. Yang, J. Kim, S. Peng, Z. Liu, and X. Tong, "Mps-nerf: Generalizable 3d human rendering from multiview images," *IEEE TPAMI*, 2022.

[10] A. Gavryushin, Y. Liu, D. Huang, Y.-L. Kuo, J. Valentin, L. Van Gool, O. Hilliges, and X. Wang, "Romeo: Revisiting optimization methods for reconstructing 3d human-object interaction models from images," in *ECCV*, 2024.

[11] S. Hu, F. Hong, L. Pan, H. Mei, L. Yang, and Z. Liu, "Sherf: Generalizable human nerf from a single image," in *ICCV*, 2023.

[12] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, "Arch: Animatable reconstruction of clothed humans," in *CVPR*, 2020.

[13] Y. Jiang, S. Jiang, G. Sun, Z. Su, K. Guo, M. Wu, J. Yu, and L. Xu, "Neuralhofusion: Neural volumetric rendering under human-object interactions," in *CVPR*, 2022.

[14] Y. Jiang, Z. Shen, P. Wang, Z. Su, Y. Hong, Y. Zhang, J. Yu, and L. Xu, "Hifi4g: High-fidelity human performance rendering via compact gaussian splatting," in *CVPR*, 2024.

[15] Y. Jiang, K. Yao, Z. Su, Z. Shen, H. Luo, and L. Xu, "Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream," in *CVPR*, 2023.

[16] H. Ju, R. Juan, R. Gomez, K. Nakamura, and G. Li, "Transferring policy of deep reinforcement learning from simulation to reality for robotics," *NMI*, 2022.

[17] K. Katsumata, D. M. Vo, and H. Nakayama, "A compact dynamic 3d gaussian representation for real-time dynamic view synthesis," in *ECCV*, 2024.

[18] K. Kedia, A. Bhardwaj, P. Dan, and S. Choudhury, "Interact: Transformer models for human intent prediction conditioned on robot actions," in *ICRA*, 2024.

[19] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM TOG*, 2023.

[20] C.-P. Lam, C.-T. Chou, K.-H. Chiang, and L.-C. Fu, "Human-centered robot navigation—towards a harmoniously human–robot coexisting environment," *IEEE Transactions on Robotics*, 2010.

[21] I. Lee, B. Kim, and H. Joo, "Guess the unseen: Dynamic 3d scene reconstruction from partial 2d glimpses," in *CVPR*, 2024.

[22] S. Lee, L. Chen, J. Wang, A. Liniger, S. Kumar, and F. Yu, "Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields," *IEEE RAL*, 2022.

[23] C. Li, Z. Ai, T. Wu, X. Li, W. Ding, and H. Xu, "Deformnet: Latent space modeling and dynamics prediction for deformable object manipulation," in *ICRA*, 2024.

[24] H. Li, D. Zhang, Y. Dai, N. Liu, L. Cheng, J. Li, J. Wang, and J. Han, "Gp-nerf: Generalized perception nerf for context-aware 3d scene understanding," in *CVPR*, 2024.

[25] Y. Liang, N. Khan, Z. Li, T. Nguyen-Phuoc, D. Lanman, J. Tompkin, and L. Xiao, "Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis," in *WACV*, 2025.

[26] S. Lin, H. Zhang, Z. Zheng, R. Shao, and Y. Liu, "Learning implicit templates for point-based clothed human modeling," in *ECCV*, 2022.

[27] J.-W. Liu, Y.-P. Cao, T. Yang, Z. Xu, J. Keppo, Y. Shan, X. Qie, and M. Z. Shou, "Hosnerf: Dynamic human-object-scene neural radiance fields from a single video," in *ICCV*, 2023.

[28] X. Liu, X. Zhan, J. Tang, Y. Shan, G. Zeng, D. Lin, X. Liu, and Z. Liu, "Humangaussian: Text-driven 3d human generation with gaussian splatting," in *CVPR*, 2024.

[29] Y. Liu, C. Luo, L. Fan, N. Wang, J. Peng, and Z. Zhang, "Citygaussian: Real-time high-quality large-scale scene rendering with gaussians," in *ECCV*, 2025.

[30] Y. Liu, C. Zhang, R. Xing, B. Tang, B. Yang, and L. Yi, "Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement," in *CVPR*, 2025.

[31] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: a skinned multi-person linear model," *ACM TOG*, 2015.

[32] X. Lv, L. Xu, Y. Yan, X. Jin, C. Xu, S. Wu, Y. Liu, L. Li, M. Bi, W. Zeng, *et al.*, "Himo: A new benchmark for full-body human interacting with multiple objects," in *ECCV*, 2024.

[33] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, pp. 99–106, 2021.

[34] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *ICCV*, 2021.

[35] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao, "Animatable neural radiance fields for modeling dynamic human bodies," in *ICCV*, 2021.

[36] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *CVPR*, 2021.

[37] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang, "3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting," in *CVPR*, 2024.

[38] J. R. Sánchez-Ibáñez, C. J. Pérez-del Pulgar, and A. García-Cerezo, "Path planning for autonomous mobile robots: A review," *Sensors*, 2021.

[39] S. Scheggi, M. Aggravi, and D. Prattichizzo, "Cooperative navigation for mixed human–robot teams using haptic feedback," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 462–473, 2016.

[40] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *CVPR*, 2016, pp. 4104–4113.

[41] M.-L. Shih, J.-B. Huang, C. Kim, R. Shah, J. Kopf, and C. Gao, "Modeling ambient scene dynamics for free-view synthesis," in *ACM SIGGRAPH*, 2024.

[42] Z. Su, L. Xu, D. Zhong, Z. Li, F. Deng, S. Quan, and L. Fang, "Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream," *IEEE TPAMI*, 2022.

[43] G. Sun, X. Chen, Y. Chen, A. Pang, P. Lin, Y. Jiang, L. Xu, J. Yu, and J. Wang, "Neural free-viewpoint performance rendering under complex human-object interactions," in *ACM MM*, 2021.

[44] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *ICLR*, 2018.

[45] T. Wang, P. Zheng, S. Li, and L. Wang, "Multimodal human–robot interaction for human-centric smart manufacturing: a survey," *Advanced Intelligent Systems*, 2024.

[46] W. Wang, J. Xiao, Y. Zhuang, and L. Chen, "Physics-aware human-object rendering from sparse views via 3d gaussian splatting," *arXiv preprint arXiv:2503.09640*, 2025.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, 2004.

[48] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, "Humannerf: Free-viewpoint rendering of moving people from monocular video," in *CVPR*, 2022.

[49] W. Xian, J.-B. Huang, J. Kopf, and C. Kim, "Space-time neural irradiance fields for free-viewpoint video," in *CVPR*, 2021.

[50] J. Xiao, S. L. Joseph, X. Zhang, B. Li, X. Li, and J. Zhang, "An assistive navigation framework for the visually impaired," *IEEE transactions on human-machine systems*, vol. 45, no. 5, pp. 635–640, 2015.

[51] X. Xie, B. L. Bhatnagar, and G. Pons-Moll, "Visibility aware human-object interaction tracking from single rgb camera," in *CVPR*, 2023.

[52] Z. Yan, C. Li, and G. H. Lee, "Nerf-ds: Neural radiance fields for dynamic specular objects," in *CVPR*, 2023.

[53] L. Yang, X. Zhan, K. Li, W. Xu, J. Li, and C. Lu, "Cpf: Learning a contact potential field to model the hand-object interaction," in *ICCV*, 2021.

[54] J. Zhang, J. Li, X. Yu, L. Huang, L. Gu, J. Zheng, and X. Bai, "Cor-gs: sparse-view 3d gaussian splatting via co-regularization," in *ECCV*, 2024.

[55] J. Zhang, H. Luo, H. Yang, X. Xu, Q. Wu, Y. Shi, J. Yu, L. Xu, and J. Wang, "Neuraldome: A neural modeling pipeline on multi-view human-object interactions," in *CVPR*, 2023.

[56] J. Zhang, J. Zhang, Z. Song, Z. Shi, C. Zhao, Y. Shi, J. Yu, L. Xu, and J. Wang, "Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment," in *CVPR*, 2024.

[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[58] A. Zhou, M. J. Kim, L. Wang, P. Florence, and C. Finn, "Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis," in *CVPR*, 2023.

[59] Q. Zhou, M. Maximov, O. Litany, and L. Leal-Taixé, "The nerfect match: Exploring nerf features for visual localization," in *ECCV*, 2025.

[60] Z. Zhu, Z. Fan, Y. Jiang, and Z. Wang, "Fsgs: Real-time few-shot view synthesis using gaussian splatting," in *ECCV*, 2025.