CAGE-GS: CAMERA-AWARE GEOMETRY ENHANCEMENT FOR GENERALIZABLE SPARSE-VIEW HUMAN GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse-view observations provide limited geometric evidence for generalizable human reconstruction, making reliable cross-view correspondence and view-consistent Gaussian prediction difficult. Existing methods either depend on parametric human templates or rely on depth estimates without sufficient cross-view evidence, leading to incomplete or spatially inconsistent geometry and unreliable cross-view aggregation. We present CAGE-GS, a generalizable framework that progressively transforms coarse geometry into reliable cross-view evidence for Gaussian reconstruction. Specifically, a Camera-Aware Warp Refiner establishes camera-grounded correspondence from coarse inverse depth and uses cross-view semantic consistency to guide iterative geometry refinement. The refined geometry then guides a Projection-Aware Cross-View Attention to construct spatially consistent cross-view candidates and selectively aggregate complementary multi-view features for Gaussian prediction. Together, these components turn coarse geometry into reliable camera-grounded evidence for geometry-consistent Gaussian prediction from sparse calibrated views, without relying on parametric human templates or ground-truth depth. Across three benchmarks under the four-view setting, CAGE-GS improves PSNR and SSIM by up to 3.34 dB and 0.0164, respectively, while reducing LPIPS by up to 22.8%, with consistent advantages across varying input-view configurations.

1 INTRODUCTION

Novel View Synthesis (NVS) aims to synthesize photorealistic images from unseen viewpoints given a sparse set of observations, with broad applications in digital humans, virtual reality, and immersive telepresence Li et al. (2025); Shao et al. (2024); Luiten et al. (2024). Recently, 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) has emerged as a promising paradigm for NVS owing to its explicit 3D primitives and efficient differentiable rasterization. This progress has catalyzed interest in generalizable human Gaussian splatting, which seeks to reconstruct unseen human subjects from sparse calibrated views without costly per-subject optimization. However, sparse-view human reconstruction remains particularly ill-posed: severe self-occlusion, non-rigid articulated deformation, and limited visual overlap make it difficult to establish reliable cross-view correspondence and consistently aggregate complementary observations.

The primary challenge lies in the severe **geometric ambiguity** under extreme viewpoint sparsity. To anchor human geometry, existing methods introduce parametric body models Hu et al. (2024b); Xiao et al. (2025) or pre-fitted templates Kwon et al. (2025). However, their reconstruction fidelity relies heavily on precise template fitting, which can become unreliable under loose garments, complex interactions, or non-rigid accessories. Conversely, template-free methods that directly predict pixel-wise depth Zheng et al. (2024); Kim & Kim (2026) avoid template fitting and exploit multi-view geometric constraints, yet geometry can remain uncertain in occluded or weakly constrained regions under sparse observations. Ambiguous cross-view correspondence may lead to incomplete or spatially inconsistent geometry, causing inaccurate Gaussian placement and distorted local structures. This motivates a reliable mechanism for refining coarse geometry under sparse-view settings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

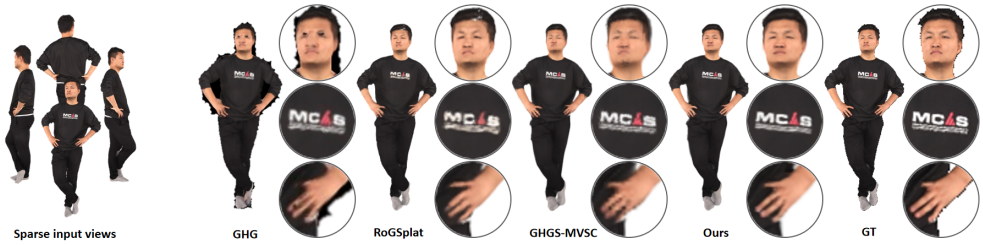

Figure 1: Qualitative comparison under the four-view setting. From left to right: input views, GHG Kwon et al. (2025), RoGSplat Xiao et al. (2025), GHGS-MVSC Kim & Kim (2026), CAGE-GS (Ours), and ground truth. The insets highlight finer facial, logo, and hand details recovered by CAGE-GS.

The second challenge arises from **cross-view feature inconsistency**. Conventional generalizable frameworks typically fuse multi-view features through heuristic pooling Lu et al. (2025), feature concatenation Zheng et al. (2024), or cross-view attention Hu et al. (2025), without explicitly enforcing spatial consistency across views. Under sparse human observations, occlusion and similar appearance across body parts make it difficult to distinguish relevant from irrelevant observations. Unconstrained aggregation can therefore mix geometrically inconsistent or occluded features, resulting in blurred details and view-inconsistent appearance. This motivates selectively aggregating complementary observations according to their spatial consistency with the target point.

To address these challenges, we exploit **camera geometry to establish reliable cross-view correspondence** and progressively build from geometric refinement to correspondence-aware feature aggregation. **First**, we leverage coarse inverse depth to establish camera-grounded correspondence across views, and use cross-view semantic consistency to assess the reliability of the resulting observations. This allows cross-view evidence to resolve geometric ambiguity and progressively refine uncertain regions, particularly where sparse observations provide limited visual support. **Second**, once reliable geometry is established, we use it to determine which observations are spatially compatible with each target point, rather than allowing cross-view features to be freely mixed. This enables selective aggregation of complementary appearance and semantic cues from geometrically consistent observations while reducing the influence of occluded or mismatched evidence.

Based on this strategy, we propose **CAGE-GS**, a template-free framework for coarse-to-fine geometry refinement and projection-aware Gaussian prediction. Given sparse calibrated views, the **Camera-Aware Warp Refiner** uses coarse inverse depth to establish camera-grounded correspondence, leverages cross-view semantic consistency to identify reliable matches, and iteratively refines geometry through a geometry–correspondence loop. The refined geometry is then leveraged by the **Projection-Aware Cross-View Attention** to identify spatially consistent observations and selectively aggregate complementary multi-view features for Gaussian prediction. CAGE-GS requires neither fitted SMPL-X models nor ground-truth 3D depth. We train the framework with a two-stage rendering-supervised strategy, where the refiner is first pretrained with an auxiliary Gaussian rendering branch and subsequently fine-tuned with the Gaussian prediction network under geometric regularization. Extensive experiments on THuman2.0 and multiple human benchmarks demonstrate the effectiveness of CAGE-GS under varying view sparsity and diverse dataset conditions, with particularly strong performance in highly sparse-view settings.

The main contributions of this paper are summarized as follows:

- We propose **CAGE-GS**, a template-free framework for generalizable human Gaussian splatting from sparse calibrated views that addresses geometric ambiguity and unreliable cross-view evidence through progressive camera-aware geometry reasoning, without relying on parametric human templates or ground-truth depth.
- We introduce a **Camera-Aware Warp Refiner** that iteratively refines inverse depth through camera-grounded correspondence and cross-view semantic consistency, and a **Projection-Aware Cross-View Attention** that uses the refined geometry to selectively aggregate spatially consistent multi-view evidence for Gaussian prediction.

- Extensive experiments demonstrate improvements over existing methods across view settings and human datasets, particularly under highly sparse views, while ablations validate camera-grounded geometry refinement and geometry-guided cross-view aggregation.

2 Related Work

2.1 Human Novel View Synthesis

Neural Radiance Fields (NeRFs) Mildenhall et al. (2020) synthesize novel views by querying continuous volumetric representations via ray marching. Building upon this, human-specific NeRFs integrate body priors or skeletal poses to model articulated structures Peng et al. (2021); Kwon et al. (2021), while generalizable methods exploit image-conditioned features and cross-view aggregation to reconstruct unseen performers from sparse inputs Chen et al. (2022); Pan et al. (2023); Hu et al. (2023); Gao et al. (2022); Kwon et al. (2023). Despite their success, implicit volume rendering requires dense point evaluations along each ray, incurring substantial computational costs. Moreover, recovering fine human geometry from sparse observations remains challenging without explicit spatial primitives, motivating more efficient, explicit representations.

3D Gaussian Splatting (3DGS) Kerbl et al. (2023) has emerged as an efficient explicit representation enabling real-time rendering. To capture dynamic humans, early methods optimize subject-specific Gaussians tied to parametric body priors Hu et al. (2024b); Shao et al. (2024); Kocabas et al. (2024); Hu et al. (2024a); Qian et al. (2024). To eliminate time-consuming per-scene optimization, recent feed-forward approaches directly regress 3D Gaussians for unseen subjects from sparse multi-view images in a single forward pass Zheng et al. (2024); Kwon et al. (2025); Xiao et al. (2025); Tu et al. (2025); Pan et al. (2024); Liu et al. (2025). However, sparse inputs provide limited visual cues, making accurate 3D Gaussian positioning and robust cross-view feature fusion particularly challenging in occluded or under-constrained regions.

2.2 Geometry-Guided Cross-View Reasoning

Accurate 3D geometry is pivotal for generalizable human Gaussian reconstruction, as it dictates spatial alignment across views. Existing methods regularize Gaussian positioning via parametric templates (e.g., SMPL) Kwon et al. (2025); Loper et al. (2015); Xiao et al. (2025); Wen et al. (2025) or stereo depth estimation Zheng et al. (2024); Charatan et al. (2024); Chen et al. (2024); Liu et al. (2025). However, template-based priors struggle with non-body topologies (e.g., loose clothing), while stereo estimation suffers from noise and ambiguities under sparse views and wide baselines. Consequently, both paradigms often yield uncertain or inaccurate 3D Gaussian initializations, particularly around complex silhouette boundaries and heavily occluded regions.

To resolve visual ambiguities, recent approaches employ cross-view attention, 3D feature lifting, or visual foundation priors (e.g., VGGT, DINOv2) for feature aggregation Tu et al. (2025); Hu et al. (2025); Kim & Kim (2026); Wang et al. (2025); Oquab et al. (2023). Nonetheless, these modules typically treat the estimated geometry as a static prerequisite. When initial geometry is noisy, cross-view fusion inevitably samples misaligned features across occlusion boundaries, leading to severe feature bleeding. Without dynamic geometric verification, these corrupted features directly degrade Gaussian property prediction, resulting in ghosting artifacts and blurred textures.

To overcome these bottlenecks, **CAGE-GS** introduces a geometry-guided reasoning framework with progressive refinement: it first iteratively rectifies coarse geometric estimates through cross-view correspondence updates, and subsequently constructs projection-consistent sampling candidates to ensure robust, occlusion-aware feature aggregation across sparse views.

3 Method

Given N sparse calibrated human images $\{I_i\}_{i=1}^{N}$ ($I_i \in \mathbb{R}^{H\times W\times 3}$) and camera parameters $\{(\mathbf{K}_i, \mathbf{T}_i)\}_{i=1}^{N}$, our goal is to feed-forwardly reconstruct a generalizable 3D Gaussian representation of an unseen subject for novel-view synthesis. We denote camera projection and unprojection under view i as $\Pi_i(\cdot)$ and $\Pi_i^{-1}(\cdot)$, respectively. As illustrated in Figure 2, CAGE-GS addresses geometric ambiguity and unreliable cross-view evidence through a two-stage design without relying

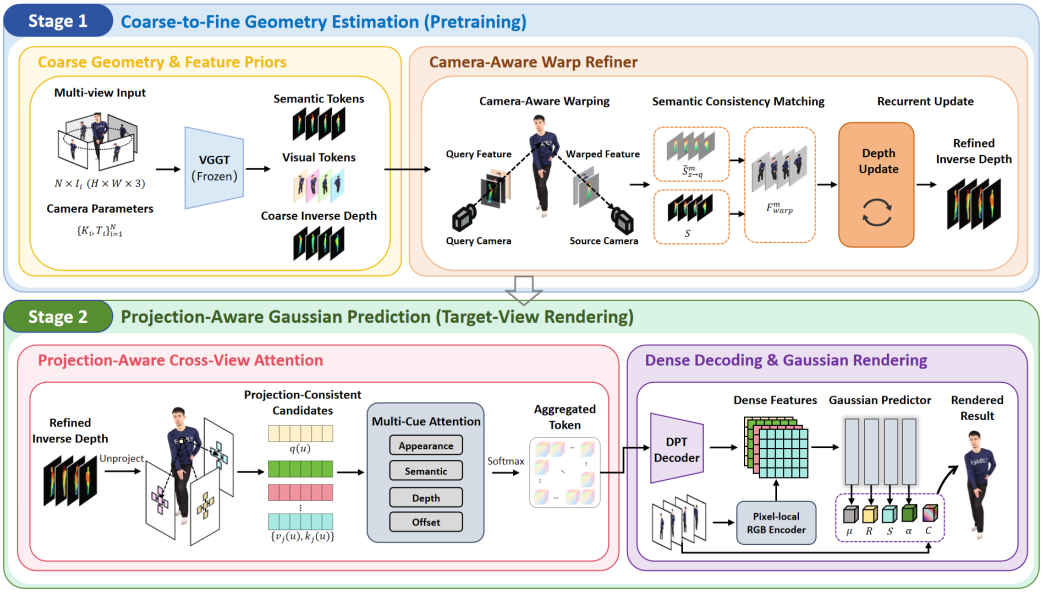

Figure 2: **Overview of CAGE-GS.** Given sparse views, CAGE-GS reconstructs generalizable 3D Gaussians for novel-view synthesis. In Stage 1, coarse geometry from a frozen VGGT backbone is progressively refined by the Camera-Aware Warp Refiner through camera-guided cross-view correspondence. In Stage 2, the Projection-Aware Cross-View Attention exploits refined geometry to selectively aggregate spatially consistent multi-view features for Gaussian prediction.

on parametric templates or ground-truth 3D depth. Given coarse inverse-depth estimates and multi-view features from a frozen VGGT, the **Camera-Aware Warp Refiner** establishes camera-grounded cross-view correspondence and progressively refines the geometry through cross-view consistency. The refined geometry is then exploited by the **Projection-Aware Cross-View Attention** to identify spatially consistent observations and selectively aggregate complementary multi-view features for Gaussian prediction. The resulting features are decoded into 3D Gaussian attributes and rendered for novel-view synthesis under multi-view photometric supervision and geometric regularization.

3.1 Coarse-to-Fine Geometry Estimation

Coarse geometry and feature priors. We first use a frozen VGGT model to extract multi-view visual and semantic features, together with a coarse inverse-depth map ρ_i^0 from its frozen depth head. The inverse depth is normalized within the foreground region to reduce scale variation across subjects and views. To support geometry refinement, a lightweight convolutional encoder further extracts trainable image features $F_i^w \in \mathbb{R}^{H_w \times W_w \times C_w}$, while the VGGT semantic features are adapted to the same resolution and used to assess cross-view consistency.

Camera-Aware Warp Refiner. Starting from the coarse inverse depth, we establish camera-grounded cross-view correspondence and progressively refine the geometry through an iterative process. Rather than treating the coarse geometry as a fixed basis for feature warping, we use it to establish cross-view correspondence and exploit the resulting multi-view evidence to refine the geometry. For a pixel p in query view q, the current inverse depth $\rho_q^m(p)$ determines its 3D location:

$$X_q^m(p) = \Pi_q^{-1}\left(p, \frac{1}{\rho_q^m(p)}\right). \tag{1}$$

This point is reprojected into a neighboring source view s as

$$p_{s\leftarrow q}^m = \Pi_s\left(X_q^m(p)\right). \tag{2}$$

We then sample the source-view image features, semantic features, and foreground mask at the projected location, obtaining $\widetilde{F}_{s\to q}^{w,m}(p)$, $\widetilde{S}_{s\to q}^m(p)$, and $\widetilde{M}_{s\to q}^m(p)$, respectively. A projected correspondence is considered reliable according to its projection validity, foreground membership, and

semantic consistency. We measure the latter using the cosine similarity between the query and warped semantic features:

$$c_s^m(p) = \max\left(0, \cos\left(S_q(p), \widetilde{S}_{s\to q}^m(p)\right)\right). \tag{3}$$

These cues are combined into a correspondence weight $w_s^m(p)$ to aggregate reliable cross-view evidence:

$$F_{\text{warp}}^m(p) = \frac{\sum_{s\in\mathcal{S}_q} w_s^m(p)\widetilde{F}_{s\to q}^{w,m}(p)}{\sum_{s\in\mathcal{S}_q} w_s^m(p) + \epsilon}. \tag{4}$$

The aggregated evidence is then fed back to refine the current geometry. Given the query feature, warped feature, their difference, current inverse depth, and aggregate confidence γ_q^m, the refiner predicts a residual update:

$$\Delta\rho_q^m = \Phi\left(F_q^w, F_{\text{warp}}^m, F_q^w - F_{\text{warp}}^m, \rho_q^m, \gamma_q^m\right). \tag{5}$$

The inverse depth is updated as

$$\rho_q^{m+1} = \mathcal{A}\left(\rho_q^m + \Delta\rho_q^m\right), \tag{6}$$

where $\mathcal{A}(\cdot)$ enforces valid inverse-depth values. A recurrent hidden state is maintained across iterations to stabilize refinement in uncertain regions. The updated geometry is then used to re-establish cross-view correspondence in the next iteration, forming a geometry–correspondence refinement loop: improved geometry yields more reliable correspondence, which in turn provides stronger cross-view constraints for subsequent geometry refinement. After several iterations, the refiner outputs the refined inverse-depth map $\rho_q^\star$, which provides reliable geometric correspondence for subsequent projection-aware feature aggregation.

3.2 Projection-Aware Gaussian Prediction

Projection-Aware Cross-View Attention. Given the refined inverse depth $\rho_q^\star$, we project query tokens into the reference views to establish spatially consistent cross-view candidates. Specifically, for a token location u in view q, its 3D position is obtained by

$$X_q(u) = \Pi_q^{-1}\left(u, \frac{1}{\rho_q^\star(u)}\right). \tag{7}$$

We then project $X_q(u)$ into each reference view and sample local features around the projected location, forming a compact candidate set

$$\mathcal{C}^\ell(u) = \left\{E_s^\ell\left(\Pi_s(X_q(u)) + \delta\right) \mid s \in \{1,\dots,N\}, \delta \in \mathcal{O}\right\}, \tag{8}$$

where $\mathcal{O}$ denotes a small local sampling pattern and the self-view candidate is included when valid. This projection-based selection avoids global attention over spatially unrelated tokens and restricts aggregation to observations that are geometrically plausible for the query point.

For each projected candidate, we jointly measure appearance similarity and semantic–geometric consistency with the query. Specifically, the attention logit is defined as

$$A_j(u) = \frac{\mathbf{q}(u)^\top \mathbf{k}_j(u)}{\sqrt{d}} + \lambda_{\text{sem}} A_{\text{sem}}^j(u) - \lambda_{\text{dep}} A_{\text{dep}}^j(u) - \lambda_{\text{off}} A_{\text{off}}^j(u), \tag{9}$$

and candidates that violate image, foreground, or projected-depth validity are excluded from aggregation. The projection-aware feature is then computed as

$$\widehat{E}_q^\ell(u) = \sum_{j\in\mathcal{V}(u)} \text{Softmax}_j(A_j(u))\,\mathbf{v}_j(u), \tag{10}$$

where $\mathcal{V}(u)$ denotes the valid candidate set.

The aggregated features are introduced through a learned residual gate and processed by an MLP. This projection-aware aggregation is applied to multiple VGGT feature levels to capture complementary information across views and scales.

Dense Decoding and Gaussian Rendering. The aggregated multi-level tokens are decoded by a DPT-style decoder into dense Gaussian features:

$$F_g = \mathcal{D}\left(\{\widehat{E}^{\ell}\}_{\ell\in\mathcal{L}}\right). \tag{11}$$

To preserve fine-grained appearance details, we extract pixel-local features F^i_{local} from each reference image and fuse them with the decoded features:

$$Z_i(p) = F^i_g(p) + \mathcal{F}_{\text{fuse}}\left(\left[F^i_g(p), F^i_{\text{local}}(p)\right]\right). \tag{12}$$

For each foreground pixel p in reference view i, the refined inverse depth provides a geometric prior for the Gaussian center, which is further refined by a predicted offset:

$$\mu_i(p) = \Pi_i^{-1}\left(p, \frac{1}{\rho_i^{\star}(p)}\right) + \Delta\mu_i(p). \tag{13}$$

A Gaussian prediction head regresses the remaining attributes, including rotation $\mathbf{r}_i(p)$, scale $\mathbf{s}_i(p)$, and opacity $\alpha_i(p)$, from $Z_i(p)$. The Gaussian color is directly inherited from the corresponding reference pixel. Collecting Gaussians over all foreground pixels yields the 3D representation $\mathcal{G}$, which is rendered to the target view using a differentiable Gaussian rasterizer:

$$\hat{I}_r, \hat{M}_r = \mathcal{R}\left(\mathcal{G}, \mathbf{K}_r, \mathbf{T}_r\right). \tag{14}$$

3.3 Training Objectives

We define the rendering loss as a combination of the L1 and SSIM losses Wang et al. (2004):

$$\mathcal{L}_r = \alpha\left\|\hat{I}_r - I_r\right\|_1 + (1-\alpha)\left(1 - \text{SSIM}(\hat{I}_r, I_r)\right), \tag{15}$$

where $\hat{I}_r$ and I_r denote the rendered and ground-truth target images, respectively.

Stage 1. We first train the Camera-Aware Warp Refiner with an auxiliary Gaussian rendering branch. The refined inverse depth is back-projected into 3D points, reference RGB values are used as Gaussian colors, and a lightweight head predicts Gaussian rotation, scale, and opacity. The auxiliary Gaussians are then rendered into the target view. The Stage 1 objective is

$$\mathcal{L}_{\text{stage1}} = \lambda_r\mathcal{L}_r + \lambda_c\mathcal{L}_c + \lambda_p\mathcal{L}_p, \tag{16}$$

where $\mathcal{L}_c$ denotes the colored Chamfer loss for cross-view point-cloud consistency, and $\mathcal{L}_p$ regularizes the refined inverse depth toward the coarse VGGT prior.

Stage 2. We initialize the refiner with the Stage 1 weights and jointly fine-tune it with the projection-aware Gaussian prediction network. A frozen copy of the initial refiner is retained as a geometric anchor. The Stage 2 objective is

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_r + \lambda_a\mathcal{L}_a + \lambda_m\mathcal{L}_m, \tag{17}$$

where $\mathcal{L}_a$ is the foreground-masked L1 distance between the current inverse depth and the frozen anchor prediction, and $\mathcal{L}_m$ supervises the rendered target-view silhouette.

The VGGT backbone remains frozen in both stages, and no ground-truth depth, SMPL/SMPL-X parameters, or explicit 3D annotations are required.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation metrics. We evaluate CAGE-GS on three benchmark human datasets: THuman2.0 Yu et al. (2021), HuMMan Cai et al. (2022), and 2K2K Han et al. (2023). Following previous methods Kwon et al. (2025); Xiao et al. (2025), we split the 526 THuman2.0 subjects into 426 for training and 100 for testing. For HuMMan, we follow the official split of 317 training sequences and 22 test sequences. For 2K2K, we use its official split of 2,000 training scans and 50 test scans. We employ PSNR, SSIM Wang et al. (2004), and LPIPS Zhang et al. (2018) to quantitatively evaluate rendering quality over the entire image, where higher PSNR and SSIM and lower LPIPS indicate better performance.

Table 1: Quantitative comparison on THuman2.0 under different numbers of input views. *Template* indicates the use of a parametric human body template (e.g., SMPL/SMPL-X), while *Depth* indicates the use of ground-truth depth. ✓ and × denote usage and non-usage, respectively. Higher PSNR and SSIM are better, while lower LPIPS is better.

Method	Template	Depth	3 views			4 views			5 views		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GPS-Gaussian Zheng et al. (2024)	×	✓	–	–	–	–	–	–	26.54	0.9539	0.0610
EVA-Gaussian Hu et al. (2025)	×	✓	–	–	–	26.31	0.9555	0.0391	27.54	0.9614	0.0297
GHG Kwon et al. (2025)	✓	×	24.84	0.9232	0.0719	25.84	0.9462	0.0633	26.53	0.9571	0.0398
RoGSplat Xiao et al. (2025)	✓	×	26.32	0.9478	0.0530	28.94	0.9615	0.0433	30.98	0.9711	0.0341
GHGS-MVSC Kim & Kim (2026)	✓	×	27.81	0.9521	0.0517	30.93	0.9710	**0.0334**	31.54	0.9729	**0.0269**
CAGE-GS (Ours)	×	×	**29.74**	**0.9661**	**0.0403**	**31.21**	**0.9743**	**0.0334**	**31.61**	**0.9756**	0.0320

Table 2: Quantitative comparison across different human datasets under the four-view setting. A separate model is trained and evaluated on each dataset.

Method	Template	Depth	THuman2.0			2K2K			HuMMan		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
EVA-Gaussian Hu et al. (2025)	×	✓	26.31	0.9555	0.0391	29.72	0.9652	0.0504	26.10	0.9442	0.0679
GHG Kwon et al. (2025)	✓	×	25.84	0.9462	0.0633	25.76	0.9397	0.0599	22.40	0.8915	0.0945
RoGSplat Xiao et al. (2025)	✓	×	28.94	0.9615	0.0433	29.29	0.9609	0.0350	24.94	0.9390	0.0683
GHGS-MVSC Kim & Kim (2026)	✓	×	30.93	0.9710	**0.0334**	26.85	0.9506	0.0571	25.06	0.9392	0.0690
CAGE-GS (Ours)	×	×	**31.21**	**0.9743**	**0.0334**	**32.31**	**0.9749**	**0.0343**	**28.40**	**0.9556**	**0.0527**

Implementation Details. CAGE-GS is implemented in PyTorch Paszke et al. (2019) and trained on a single NVIDIA RTX 4090 GPU with 24 GB of memory. All images are resized to 512×512 for training and evaluation, with a batch size of 1. We use AdamW with betas $(0.9, 0.999)$, a weight decay of 1×10^{-5}, and a learning rate of 1×10^{-4}. The VGGT backbone and depth head remain frozen throughout training. Under the default four-view setting, the Camera-Aware Warp Refiner is pretrained for 100K iterations in Stage 1 and jointly fine-tuned with the Gaussian prediction network for 150K iterations in Stage 2, using a reduced learning rate of 1×10^{-5} for the pretrained refiner. We set $\alpha = 0.8$, $\lambda_r = 1.0$, $\lambda_c = 0.05$, $\lambda_p = 0.03$, $\lambda_a = 0.01$, and $\lambda_m = 0.05$.

4.2 Comparison with the State of the Art

Quantitative evaluation. We first evaluate CAGE-GS under different numbers of input views on THuman2.0, as reported in Table 1. Despite using neither parametric human templates nor ground-truth depth, CAGE-GS consistently achieves strong performance across all view settings. With only three input views, CAGE-GS attains 29.74 dB PSNR and 0.9661 SSIM, surpassing GHGS-MVSC by 1.93 dB and 0.0140, respectively. As the number of input views increases, CAGE-GS further improves reconstruction quality, reaching 31.21 dB PSNR and 0.9743 SSIM with four views and 31.61 dB PSNR and 0.9756 SSIM with five views. Notably, CAGE-GS maintains its advantage even under extreme view sparsity, where reliable geometric correspondence is particularly challenging to establish. These results support the effectiveness of camera-aware geometry enhancement for sparse-view human reconstruction. We further compare CAGE-GS with representative NeRF-based human reconstruction methods on THuman2.0. As reported in Table 3, CAGE-GS achieves better performance across all three metrics, demonstrating the effectiveness of camera-grounded geometric reasoning for sparse-view human reconstruction.

We further evaluate the generalization ability of CAGE-GS across different human datasets under the four-view setting in Table 2. CAGE-GS consistently achieves superior performance on THuman2.0, 2K2K, and HuMMan, demonstrating robustness across diverse human appearances and capture conditions. For example, on 2K2K, CAGE-GS improves PSNR by 3.02 dB over RoGSplat and 5.46 dB over GHGS-MVSC, while achieving better perceptual quality. Similar improvements are observed on HuMMan, confirming its ability to generalize beyond a single dataset. These results further demonstrate that camera-grounded geometric refinement and geometry-guided cross-view aggregation provide an effective solution for sparse-view human reconstruction.

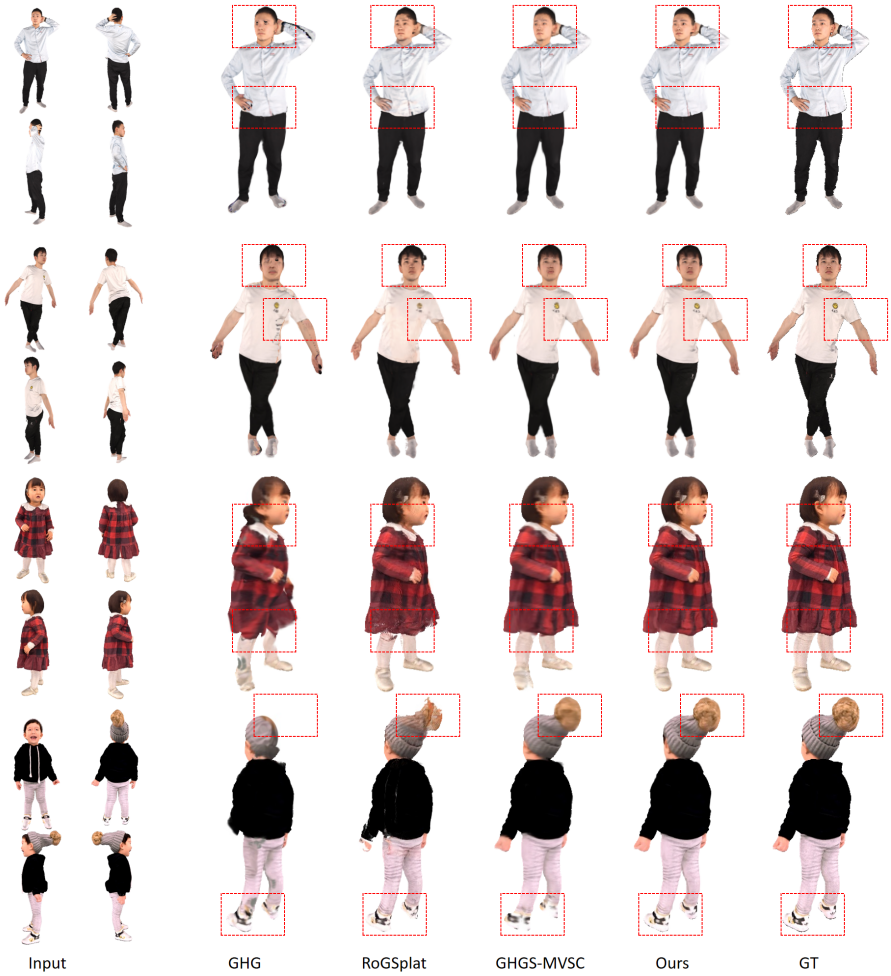

Figure 3: Qualitative comparisons on THuman2.0 (top two rows) and 2K2K (bottom two rows) under the four-view setting. From left to right: input views, GHG Kwon et al. (2025), RoGSplat Xiao et al. (2025), GHGS-MVSC Kim & Kim (2026), our method, and ground truth. CAGE-GS recovers more faithful geometry and sharper appearance details with fewer artifacts.

Qualitative evaluation. Figure 3 presents qualitative comparisons with GHG Kwon et al. (2025), RoGSplat Xiao et al. (2025), and GHGS-MVSC Kim & Kim (2026) under the four-view setting. As highlighted by the red boxes, CAGE-GS produces clearer facial contours and more faithful separation of the hand from the waist on THuman2.0, while preserving details such as the face, chest logo, and arm and garment boundaries. On 2K2K, CAGE-GS better reconstructs the profile face, plaid texture, and layered dress hem, where competing methods exhibit noticeable blurring or boundary distortions. It also preserves fine structures and local boundaries, including the knitted hat, pom-pom, ankle, and shoe boundaries, which are prone to degradation under sparse observations.

Figure 4 further compares the reconstructed geometry. RoGSplat produces noisy depth variations over the coat, while GHGS-MVSC tends to over-smooth the surface geometry. In contrast, CAGE-GS recovers a more coherent torso surface and better preserves the coat opening, hand boundaries, and garment structure, yielding inverse-depth estimates that are more consistent with the ground truth. The qualitative results support the effectiveness of camera-aware geometric refinement for improving both geometric consistency and subsequent appearance reconstruction.

4.3 Ablation Study

We conduct component ablations on the held-out THuman2.0 test set under the four-view setting. As shown in Table 4, removing the Camera-Aware Warp Refiner (CAWR) causes the largest per-

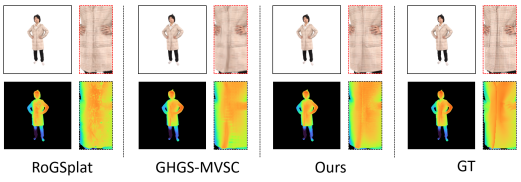

Figure 4: Qualitative comparison of novel-view renderings and inverse-depth maps. CAGE-GS produces sharper details and more coherent geometry than RoGSplat Xiao et al. (2025) and GHGS-MVSC Kim & Kim (2026).

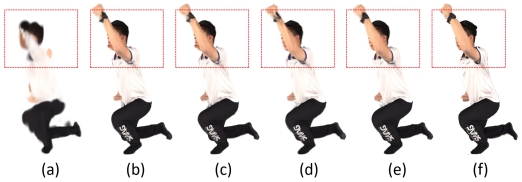

Figure 5: Qualitative ablation on THuman2.0. (a) w/o CAWR, (b) w/o PAA, (c) w/o LE, (d) w/o Joint Training, (e) full model, and (f) ground truth. The full model yields the most complete geometry and sharpest details.

Table 3: NeRF-based comparison on THuman2.0. Best results are shown in bold.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
SHERF Hu et al. (2023)	19.25	0.8942	0.1121
NHP Kwon et al. (2021)	25.74	0.9356	0.0748
GP-NeRF Chen et al. (2022)	23.28	0.9325	0.0798
TransHuman Pan et al. (2023)	27.36	0.9487	0.0505
CAGE-GS (Ours)	**31.21**	**0.9743**	**0.0334**

Table 4: Component ablation on THuman2.0. Best results are shown in bold.

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓
w/o CAWR	22.87	0.9271	0.1026
w/o PAA	30.93	0.9728	0.0340
w/o LE	30.78	0.9723	0.0337
w/o Joint Training	30.34	0.9683	0.0442
Full model	**31.21**	**0.9743**	**0.0334**

formance degradation: PSNR decreases from 31.21 to 22.87 dB, SSIM decreases from 0.9743 to 0.9271, and LPIPS increases from 0.0334 to 0.1026. This substantial degradation confirms that camera-guided cross-view correspondence is essential for recovering reliable geometry. Without joint Stage I training, PSNR decreases by 0.87 dB and LPIPS increases to 0.0442, indicating that joint Stage I training improves the geometry learned by the refiner. Removing PAA or LE also causes consistent performance degradation. PAA improves correspondence-aware cross-view feature selection, whereas LE contributes to the recovery of fine-scale appearance details. Overall, the full model achieves the best performance across all three metrics.

Figure 5 further illustrates the contribution of each component. Without CAWR in (a), the upper body and raised arm become blurred, with incomplete geometry around the torso and legs. Removing PAA in (b) largely preserves the human silhouette but leads to softer facial, sleeve, and rear-leg boundaries, indicating less coherent cross-view aggregation. Without LE in (c), fine appearance details, particularly the lettering on the trousers and clothing edges, become less distinct. Removing joint Stage I training in (d) introduces noticeable smoothing around the raised arm, torso, and rear foot. In contrast, the full model in (e) better separates the arm from the torso, preserves the trouser lettering, and recovers cleaner leg and foot boundaries, yielding results more consistent with the ground truth in (f). These results are consistent with the quantitative ablation study and validate the complementary contributions of the proposed components.

5 Conclusion

We presented CAGE-GS, a template-free framework for generalizable sparse-view human Gaussian splatting. CAGE-GS addresses geometric ambiguity and unreliable cross-view evidence by progressively transforming coarse geometry into reliable cross-view correspondence. The proposed Camera-Aware Warp Refiner uses camera-grounded correspondence to refine inverse depth, while the Projection-Aware Cross-View Attention leverages the refined geometry to aggregate spatially consistent multi-view features for accurate Gaussian prediction. This progressive design connects geometric refinement with cross-view aggregation, allowing reliable geometric evidence to guide accurate appearance and local structure reconstruction. Extensive experiments across different view numbers and human datasets demonstrate high-quality reconstruction without parametric human templates or ground-truth depth, particularly under highly sparse observations and challenging viewing conditions. These results highlight camera-aware geometric reasoning as an effective basis for reliable cross-view aggregation and high-fidelity human Gaussian reconstruction.

6 AI USE STATEMENT

Generative AI tools were used exclusively for language editing and stylistic refinement. All technical content, experimental results, and conclusions were developed and verified by the authors.

REFERENCES

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-Modal 4D Human Dataset for Versatile Sensing and Modeling. In *Proceedings of the European Conference on Computer Vision 2022*, pp. 557–577, 2022.

David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive NeRF for Generalizable and Efficient Neural Human Rendering. In *Proceedings of the European Conference on Computer Vision*, volume 13683, pp. 222–239. 2022.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *Proceedings of the European Conference on Computer Vision*, 2024.

Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3D Human Digitization from Single 2K Resolution Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12869–12879, 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, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. SHERF: Generalizable Human NeRF from a Single Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9352–9364, 2023.

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.

Yingdong Hu, Zhening Liu, Jiawei Shao, Zehong Lin, and Jun Zhang. EVA-Gaussian: 3D Gaussian-Based Real-Time Human Novel View Synthesis under Diverse Multi-View Camera Settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2634–2643, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)*, 42 (4):139:1–139:14, 2023.

Jingi Kim and Wonjun Kim. Generalizable Human Gaussian Splatting via Multi-view Semantic Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–148, 2026.

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.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24741–24752, 2021.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural image-based avatars: Generalizable radiance fields for human avatar modeling. In *Proceedings of International Conference on Learning Representations*, 2023.

Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, Aayush Prakash, and Fernando De La Torre. Generalizable Human Gaussians for Sparse View Synthesis. In *Proceedings of the European Conference on Computer Vision*. 2025.

Ke Li, Mana Masuda, Susanne Schmidt, and Shohei Mori. Radiance Fields in XR: A Survey on How Radiance Fields are Envisioned and Addressed for XR Research. *IEEE Transactions on Visualization and Computer Graphics*, 31(11):9709–9719, 2025.

Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, pp. 37–53, 2025.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6): 248:1–248:16, 2015.

Jia Lu, Taoran Yi, Jiemin Fang, Chen Yang, Chuiyun Wu, Wei Shen, Wenyu Liu, Qi Tian, and Xinggang Wang. Snap-snap: Taking two images to reconstruct 3d human gaussians in milliseconds, 2025. URL https://arxiv.org/abs/2508.14892.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In *Proceedings of the International Conference on 3D Vision*, pp. 800–809, 2024.

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. In *Proceedings of the European Conference on Computer Vision*, pp. 405–421, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2023.

Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In *Advances in Neural Information Processing Systems*, 2024.

Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. TransHuman: A Transformer-Based Human Representation for Generalizable Neural Human Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3544–3555, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 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.

Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1606–1616, 2024.

Hanzhang Tu, Zhanfeng Liao, Boyao Zhou, Shunyuan Zheng, Xilong Zhou, Liuxin Zhang, QianYing Wang, and Yebin Liu. GBC-Splat: Generalizable Gaussian-Based Clothed Human Digitalization under Sparse RGB Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26377–26387, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Jing Wen, Alex Schwing, and Shenlong Wang. Life-gom: Generalizable human rendering with learned iterative feedback over multi-resolution gaussians-on-mesh. In *Proceedings of the International Conference on Learning Representations*, 2025.

Junjin Xiao, Qing Zhang, Yonewei Nie, Lei Zhu, and Wei-Shi Zheng. RoGSplat: Learning Robust Generalizable Human Gaussian Splatting from Sparse Multi-View Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5980–5990, 2025.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-Time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5746–5756, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. GPS-Gaussian: Generalizable Pixel-wise 3D Gaussian Splatting for Real-time Human Novel View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19680–19690, 2024.

A APPENDIX

Additional Two-View Evaluation. We further evaluate CAGE-GS under the more challenging two-view setting on THuman2.0. As shown in Table 5, CAGE-GS achieves 27.5719 dB PSNR, 0.9540 SSIM, and 0.0560 LPIPS, consistently surpassing all competing methods. Compared to the strongest baseline, GHGS-MVSC, our approach improves PSNR by 1.1809 dB. The consistent improvement under severely restricted visual overlap demonstrates the robustness of our progressive geometry rectification and occlusion-aware feature aggregation.

Table 5: Quantitative comparison on THuman2.0 under the two-view setting. Best results among the evaluated methods are shown in bold.

	2-view setting		
Method	PSNR ↑	SSIM ↑	LPIPS ↓
GHG Kwon et al. (2025)	21.7900	0.8766	0.1754
RoGSplat Xiao et al. (2025)	23.5700	0.9255	0.0857
GHGS-MVSC Kim & Kim (2026)	26.3910	0.9480	0.0651
CAGE-GS (Ours)	**27.5719**	**0.9540**	**0.0560**

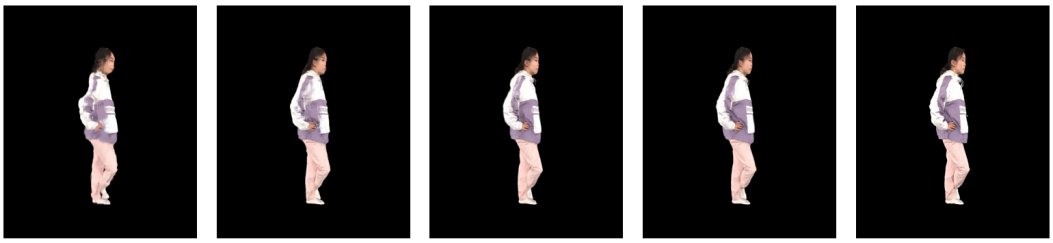

Figure 6: Effect of the number of input views. From left to right, CAGE-GS renderings of the same target view using two, three, four, and five input views, followed by ground truth (GT).

Effect of the Number of Input Views. Figure 6 illustrates novel-view synthesis results of CAGE-GS across two to five input views. Under the extreme two-view setup, minor boundary distortions and geometric ambiguities emerge around the jacket and arms due to restricted visual coverage. As the number of input views increases, additional cross-view constraints progressively resolve these ambiguities, yielding clothing geometries and body silhouettes that align more faithfully with the ground truth. While reconstruction quality continues to improve with additional views, the visual gains become less pronounced beyond four views, with both four- and five-view renderings maintaining high visual fidelity.

Additional Qualitative Comparisons under Varying Input Views. Figure 7 compares CAGE-GS with existing methods across different numbers of input views. With only two views, competing methods show visible artifacts in facial regions and blurred or distorted boundaries around the raised arms and lower body, reflecting the difficulty of recovering fine structures from severely limited observations. CAGE-GS instead preserves more coherent facial details and more accurate human silhouettes. Although additional views progressively alleviate the reconstruction ambiguity for all methods, CAGE-GS consistently yields results that more closely match the ground truth, particularly around faces, hands, and limb boundaries.

Cross-domain qualitative evaluation. To evaluate cross-domain generalization, all models are trained on THuman2.0 and evaluated on 2K2K and HuMMan under the four-view setting without additional fine-tuning. Table 6 reports the quantitative results. On 2K2K, CAGE-GS achieves 28.9404 dB PSNR, improving over RoGSplat by 3.5265 dB while also achieving the best SSIM



Figure 7: Qualitative comparisons under different numbers of input views on THuman2.0. From top to bottom, the rows correspond to the two-, three-, four-, and five-view settings, respectively. From left to right, we show the input views, GHG Kwon et al. (2025), RoGSplat Xiao et al. (2025), GHGS-MVSC Kim & Kim (2026), CAGE-GS (Ours), and ground truth (GT).

and LPIPS. On HuMMan, CAGE-GS achieves the best performance across all three metrics, with a PSNR of 28.6329 dB. These results demonstrate robust reconstruction across different human appearances and clothing styles. Figure 8 further provides qualitative comparisons. CAGE-GS better preserves human shapes, clothing boundaries, facial details, and local appearance structures, with fewer blurring and geometric artifacts under the cross-domain setting.

Table 6: Cross-domain evaluation on 2K2K and HuMMan under the four-view setting. All models are trained on THuman2.0. Best results are shown in bold.

Method	2K2K			HuMMan		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GHGS-MVSC Kim & Kim (2026)	23.9073	0.9296	0.0765	27.5388	0.9530	0.0590
RoGSplat Xiao et al. (2025)	25.4139	0.9400	0.0600	27.7088	0.9537	0.0535
CAGE-GS (Ours)	**28.9404**	**0.9587**	**0.0471**	**28.6329**	**0.9555**	**0.0527**

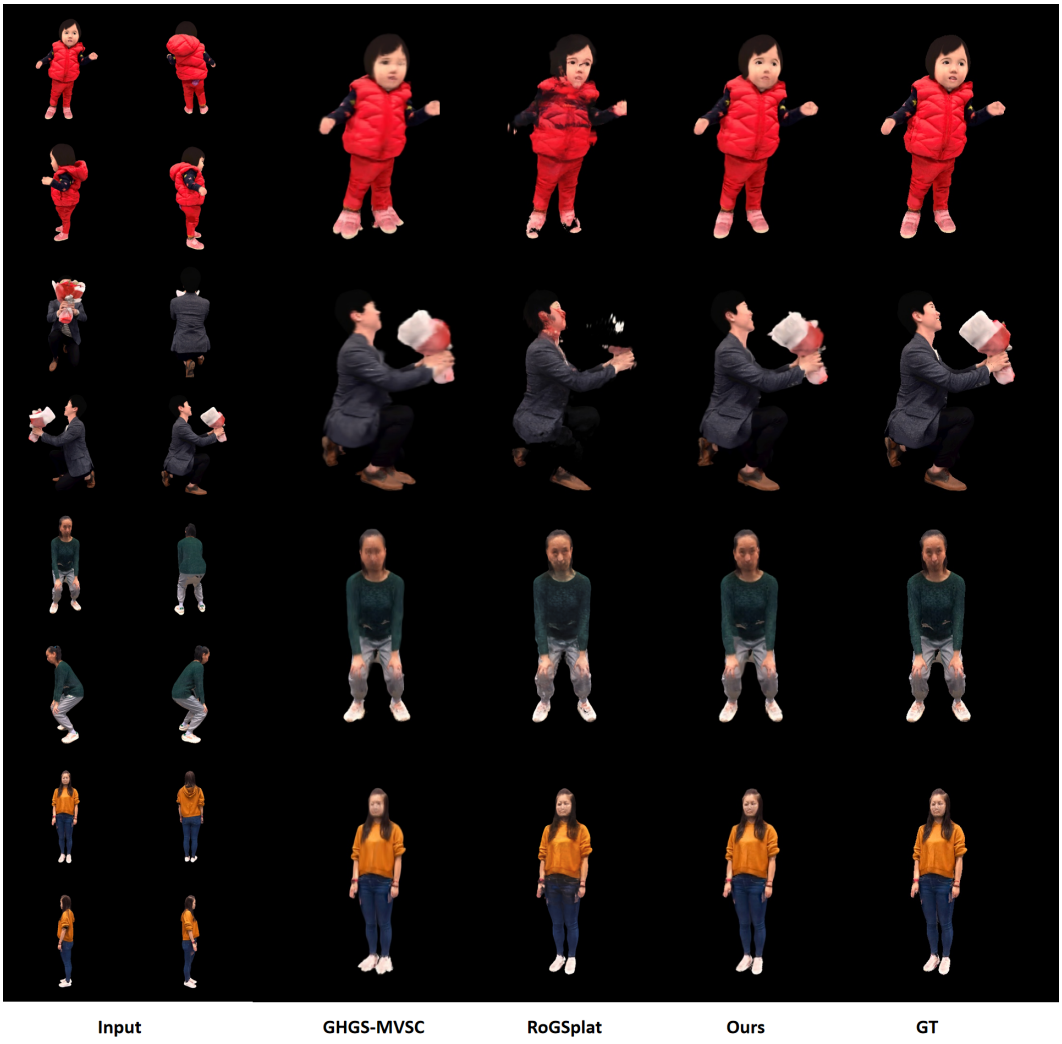

Figure 8: Qualitative cross-domain comparison on 2K2K (top two rows) and HuMMan (bottom two rows). From left to right, we show the input views, GHGS-MVSC Kim & Kim (2026), RoGSplat Xiao et al. (2025), CAGE-GS (Ours), and ground truth (GT).

Efficiency Analysis. Table 7 compares the computational efficiency of different methods under the four-view setting. CAGE-GS substantially improves resource efficiency over GHGS-MVSC, reducing core parameters from 77.94,M to 45.99,M and peak GPU memory from 11.29,GiB to 6.44,GiB, corresponding to reductions of approximately 41% and 43%, respectively. Despite incorporating geometry refinement and projection-aware cross-view aggregation, CAGE-GS maintains an inference throughput of 2.14 FPS, comparable to the 2.92 FPS of GHGS-MVSC, while requiring substantially less GPU memory. Notably, the total parameter count of CAGE-GS includes the frozen pretrained

Table 7: Efficiency comparison under the four-view setting. All methods are evaluated at 512×512 resolution with batch size 1 on a single NVIDIA RTX 4090 GPU.

Method	Core Params (M)	Total Params (M)	Latency ↓ (ms)	P95 ↓ (ms)	FPS ↑	Memory ↓ (GiB)
GHG Kwon et al. (2025)	17.88	17.88	54.96	56.38	18.19	0.96
RoGSplat Xiao et al. (2025)	12.29	12.29	140.73	152.91	7.11	3.78
GHGS-MVSC Kim & Kim (2026)	77.94	987.05	342.59	346.13	2.92	11.29
CAGE-GS (Ours)	45.99	987.75	467.11	470.87	2.14	6.44

geometry backbone, whereas the task-specific components remain relatively compact. These results demonstrate that CAGE-GS achieves a substantially more efficient task-specific architecture without sacrificing comparable inference throughput.