REAL-TIME HUMAN FRONTAL VIEW SYNTHESIS FROM A SINGLE IMAGE VIA UNIFIED GEOMETRY MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Photorealistic human novel view synthesis from a single image is crucial for democratizing immersive 3D telepresence, eliminating the need for complex multi-camera setups. However, current rendering-centric methods prioritize visual fidelity over explicit geometric understanding and struggle with intricate regions like faces and hands, leading to temporal instability. Meanwhile, human-centric frameworks suffer from memory bottlenecks since they typically rely on an auxiliary model to provide informative structural priors for geometric modeling, which limits real-time performance. To address these challenges, we propose **Rosmi**, a geometry-guided framework for instant frontal view synthesis from a single image. By avoiding external geometric modeling and focusing on frontal view synthesis, our model optimizes visual integrity for telepresence. Specifically, Rosmi introduces a novel cascade learning strategy that enables coarse-to-fine geometric feature learning. It first directly learns coarse geometric features, such as SMPL-X meshes and point clouds, and then refines textures through rendering supervision. To achieve real-time efficiency, we distill this unified framework into a lightweight linear attention model. Notably, Rosmi is the first monocular human frontal view synthesis model that achieves real-time inference at 24 FPS, significantly outperforming previous methods in both visual authenticity and structural accuracy.

1 INTRODUCTION

High-fidelity human novel view synthesis (NVS) from a single image is essential for immersive 3D telepresence and virtual video conferencing. While photorealistic rendering can be achieved through advanced hardware setups, such as volumetric scanners and multi-camera arrays, these solutions suffer from high cost and latency that hinder their widespread adoption. This motivates the need for instant reconstruction from monocular inputs such as webcams. However, recovering 3D structure from a single 2D image is inherently ill-posed due to depth ambiguity and structural inconsistency.

Existing single-image NVS methods primarily adopt rendering-centric approaches, including 3D Gaussian splatting (3DGS) (Kerbl et al., 2023) and neural radiance fields (NeRFs) (Mildenhall et al., 2021), which prioritize pixel-level visual fidelity over explicit geometric modeling. While these methods deliver visually compelling results, they often produce unstable geometry and structural artifacts, such as floating densities, temporal jitter, and depth stretching, that are particularly problematic for live telepresence. Diffusion-based methods generate high-fidelity novel views, which, however, require iterative denoising and incur substantial latency that limits real-time applications.

Human-centric NVS is substantially more challenging than general scene reconstruction because it requires accurate modeling of both human structure and high-frequency appearance. Although recent LVSMs and reconstruction models, such as efficient-LVSM (Jia et al., 2026), WorldMirror (Liu et al., 2025), and AnySplat (Jiang et al., 2025b), improve general 3D reconstruction and inference efficiency, they remain insufficient for live telepresence, where expressive human geometry and semantically sensitive regions such as faces and hands demand precise structural modeling. Estimating expressive representations such as SMPL-X (Pavlakos et al., 2019b) is significantly more complex than standard depth prediction, and even small geometric or texture errors may cause noticeable identity drift. To address these challenges, recent methods have incorporated explicit human priors.

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.

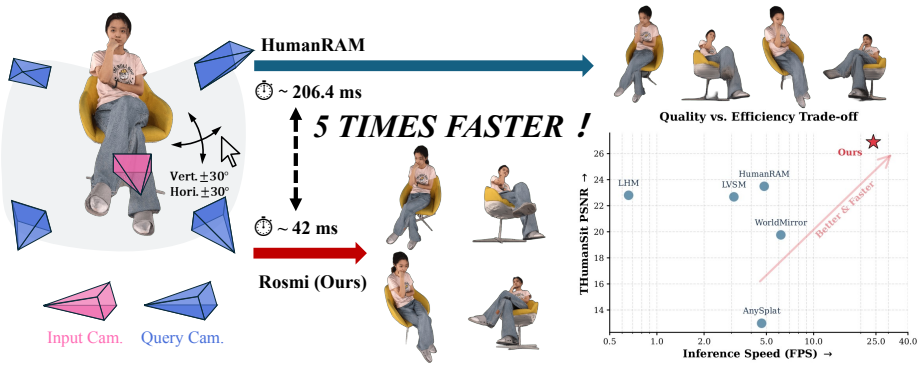


Figure 1: **Real-time high-fidelity human frontal view synthesis. Rosmi** achieves a $5\times$ inference speedup ($\sim$42 ms vs. 206.4 ms) over HumanRAM, while maintaining superior visual quality and reaching real-time frame rates of 24+ FPS.

For example, HumanRAM (Yu et al., 2025) conditions transformer-based NVS on SMPL-X priors. However, it relies on auxiliary models for structural cues, such as rendered position maps, and maintains computationally expensive triplane features, resulting in substantial memory overhead and slower inference. This dependence on external priors limits the computational budget available for high-fidelity rendering and prevents real-time telepresence.

We argue that achieving real-time high-fidelity human NVS requires integrating geometric modeling directly into the framework to enable robust geometric feature extraction without the need for auxiliary models, thereby reducing memory and computational overhead. To this end, we propose **Rosmi**, a geometry-guided framework for instant single-image frontal view synthesis. By explicitly learning structural features within a targeted viewing range, our approach achieves both high visual fidelity and real-time efficiency. Since telepresence typically does not require full 360-degree synthesis, we constrain the model to a frontal frustum of $\pm 30°$, which reduces the complexity of hallucinating heavily occluded regions while preserving visual integrity and stable view transitions.

At the core of Rosmi is our *cascade learning* strategy, which addresses the tradeoff between structural accuracy and real-time inference speed, as shown in Fig. 1. The *cascade* mechanism follows a coarse-to-fine learning paradigm. In the front-end stage, foundational geometric features, including SMPL-X meshes and point clouds, are learned directly through supervision, eliminating the need for memory-intensive auxiliary models during inference. In the back-end stage, the network refines fine-grained textures using view synthesis and 3DGS losses to generate photorealistic outputs. To achieve real-time performance, we introduce a *distillation* phase that compresses the learned geometric and texture features into an efficient backbone. It replaces computationally expensive FlashAttention (Dao et al., 2022; Dao, 2024) with lightweight linear attention (Yang & Zhang, 2024; Xie et al., 2024b), significantly reducing inference latency while preserving structural fidelity. Together, these components enable Rosmi to achieve state-of-the-art visual authenticity and structural accuracy under real-time constraints. In summary, our contributions are as follows:

- We propose Rosmi, the first monocular human frontal view synthesis model capable of identity-preserving reconstruction at 24 FPS, meeting the real-time requirements of telepresence.
- We introduce a *cascade distillation* strategy that decouples foundational geometric learning, including SMPL-X meshes and point clouds, from fine-grained texture synthesis while leveraging lightweight linear attention for efficient inference.
- By focusing on direct geometric feature extraction and limiting synthesis to a practical frontal viewing range, Rosmi overcomes the memory and computational bottlenecks of prior methods while achieving high visual fidelity and structural accuracy for real-time telepresence.

2 Related Works

2.1 Human Novel View Synthesis

Human novel view synthesis has evolved from implicit geometry recovery to real-time volumetric rendering. Early pixel-aligned implicit methods (Saito et al., 2019; 2020; Yu et al., 2021; Varol et al.,

2018; Zheng et al., 2019; Xiu et al., 2022; 2023) recover geometry from spatial features but suffer from limited resolution and slow inference. Subsequent works on NeRFs (Mildenhall et al., 2021; Kwon et al., 2021; Mihajlovic et al., 2022; Hu et al., 2023; Raj et al., 2021) and 3DGS (Kerbl et al., 2023; Kocabas et al., 2024; Kwon et al., 2024; Prospero et al., 2025) have improved photorealistic rendering. Recent feed-forward approaches (Zheng et al., 2024; Hu et al., 2024; Zhou et al., 2025) directly regress 3D attributes, yet often treat geometry as an implicit byproduct, leading to spatial inconsistency, depth ambiguity, and temporal jitter. To improve structural reliability, recent methods increasingly incorporate explicit human priors. Diffusion-based approaches (Poole et al., 2022; Liu et al., 2023; Kolotouros et al., 2023; Ho et al., 2024; Sengupta et al., 2024; Xu et al., 2024) achieve high-quality textures but are typically too slow for real-time use. Meanwhile, HumanRAM (Yu et al., 2025) conditions neural rendering on SMPL-X (Pavlakos et al., 2019b) priors, demonstrating the importance of geometric guidance for sparse inputs (Yu et al., 2025; Jin et al., 2024). However, extracting these priors with auxiliary regression models introduces substantial runtime and memory overhead. This motivates the need for unified frameworks that balance geometric fidelity with computational efficiency.

2.2 Feed-forward NVS and 3D Reconstruction

Generalizable 3D reconstruction and feed-forward NVS have advanced rapidly with large reconstruction models (LRMs), which leverage scalable transformer architectures to learn generic priors from large-scale data (Hong et al., 2023; Xie et al., 2024a). Recent foundation models directly map images to 3D representations or novel views without iterative optimization. For reconstruction, methods such as DUSt3R, CUT3R, and VGGT (Wang et al., 2024; 2025b; Chen et al., 2025b;c; Wang et al., 2025a;c) enable pixel-level 3D reasoning and point cloud regression, while NVS models such as LVSM and RayZer (Jin et al., 2024; Jiang et al., 2025a; Zhao et al., 2025) achieve strong visual quality with minimal handcrafted inductive bias. However, these approaches typically employ monolithic self-attention mechanisms that incur quadratic complexity, limiting real-time high-resolution inference (Jia et al., 2026). Efficient-LVSM (Jia et al., 2026) mitigates this bottleneck through decoupled and linearized attention, while unified frameworks further incorporate geometric priors such as depth, camera intrinsics, and surface normals (Liu et al., 2025; Jiang et al., 2025b; Wang et al., 2025a). These developments motivate lightweight multi-representation architectures that combine flexible geometric guidance with efficient feed-forward inference.

2.3 Human Mesh Reconstruction

Human mesh recovery (HMR) estimates 3D human pose and shape from monocular images and has evolved from optimization-based methods such as SMPLify-X (Pavlakos et al., 2019a) to efficient regression frameworks. Early approaches including HMR, SPIN, and PARE (Kanazawa et al., 2018; Kolotouros et al., 2019; Kocabas et al., 2021) use convolutional networks to directly regress SMPL parameters, but often struggle with fine-grained details and image-to-mesh alignment (Kolotouros et al., 2019; Kocabas et al., 2021). Recent Vision Transformer-based methods advance expressive whole-body recovery with improved facial and hand modeling (Kanazawa et al., 2018; Lin et al., 2023). State-of-the-art methods such as PEAR (Wu et al., 2026) further improve alignment and fidelity through pixel-level supervision and differentiable rendering. By estimating expressive SMPL-X parameters (Pavlakos et al., 2019b), these methods provide strong geometric priors for downstream tasks such as novel view synthesis and animation (Wu et al., 2026; Lin et al., 2023; Pavlakos et al., 2019b). Building on this progress, Rosmi integrates human geometric priors into a linearized feed-forward NVS architecture, avoiding costly external conditions and attention while enabling real-time high-fidelity telepresence.

3 Method

In this section, we present **Rosmi**, a geometry-guided framework for real-time frontal view synthesis that learns and injects explicit 3D human priors to improve structural stability. We first introduce the NVS preliminaries (Sec. 3.1), then describe the cascaded architecture for prior generation and multi-head decoding (Sec. 3.2, Fig. 2). Finally, we present the learning objectives (Sec. 3.3) and cascade distillation strategy for reducing the quadratic cost of standard transformers (Sec. 3.4).

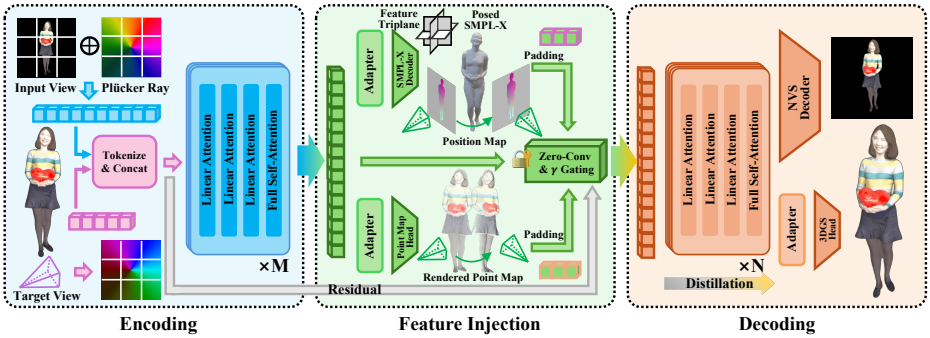


Figure 2: **Overview of the Rosmi architecture.** The framework operates through three cascaded stages: encoding global context, injecting explicit geometric priors (SMPL-X and point clouds), and decoding into NVS or 3DGS (accelerated by a progressive linear attention distillation strategy).

3.1 PRELIMINARY

We formulate single-image NVS as a feed-forward sequence-to-sequence spatial mapping problem. Following recent advancements in transformer-based view synthesis (Jin et al., 2024; Yu et al., 2025), our framework avoids compute-intensive 3D representations, such as explicit cost volumes or rigid epipolar geometric projections, during the initial feature extraction phase. Instead, it leverages the global receptive field of transformers, combined with ray-based spatial conditions, to implicitly establish cross-view correspondences.

Given a single source RGB image $\mathbf{I}_s \in \mathbb{R}^{H\times W\times 3}$ with known camera intrinsics and extrinsics, we compute a pixel-wise Plücker ray (Plucker, 1865) embedding map $\mathbf{P}_s \in \mathbb{R}^{H\times W\times 6}$. The Plücker ray representation is chosen over standard Cartesian coordinates as it inherently avoids origin singularities and preserves 3D line geometry. Specifically, for each pixel, the ray is defined as:

$$\mathbf{r} = (\mathbf{d}, \mathbf{o} \times \mathbf{d}), \tag{1}$$

where $\mathbf{o} \in \mathbb{R}^3$ denotes the camera's optical center and $\mathbf{d} \in \mathbb{R}^3$ represents the normalized viewing direction.

During the tokenization phase, both the source image and its Plücker ray (Plucker, 1865) map are divided into non-overlapping local patches of size $p \times p$. A learnable linear projection layer maps the flattened and concatenated patches into a sequence of high-dimensional source tokens $\mathbf{x}_j \in \mathbb{R}^d$:

$$\mathbf{x}_j = \text{Linear}_{inp}\left([\mathbf{I}_{s,j}, \mathbf{P}_{s,j}]\right), \tag{2}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation, and d represents the latent token embedding dimension. Similarly, the geometric condition for the desired target view is entirely encapsulated by its Plücker ray (Plucker, 1865) map $\mathbf{P}_t$, which is projected into a sequence of target query tokens $\mathbf{q}_j \in \mathbb{R}^d$:

$$\mathbf{q}_j = \text{Linear}_{tar}\left(\mathbf{P}_{t,j}\right). \tag{3}$$

The core of the framework is a multi-layer transformer model $\mathcal{T}$, which receives the concatenated sequence of source tokens $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{l_x}\}$ and target tokens $\mathbf{q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_{l_q}\}$. Using cascaded multi-head self-attention, the transformer implicitly performs cross-view correspondence matching and geometric reasoning. The network then predicts the contextualized target view tokens $\mathbf{y}_j \in \mathbb{R}^d$:

$$\mathbf{y}_1, \ldots, \mathbf{y}_{l_q} = \mathcal{T}(\mathbf{x}_1, \ldots, \mathbf{x}_{l_x}, \mathbf{q}_1, \ldots, \mathbf{q}_{l_q}). \tag{4}$$

In traditional transformer-based NVS models, the output tokens are passed through a final linear decoder, followed by a Sigmoid activation, to regress the RGB color values of the target patches: $\hat{\mathbf{I}}_j^t = \text{Sigmoid}(\text{Linear}_{out}(\mathbf{y}_j)) \in \mathbb{R}^{3p^2}$. The predicted patches are then reassembled into a 2D spatial grid to construct the synthesized novel view. However, relying solely on implicit regression without explicit structural priors often results in severe 3D inconsistencies. To address this issue, we introduce explicit geometric interventions directly within our cascaded architecture, ensuring structural stability and visual fidelity.

3.2 CASCADED MODEL ARCHITECTURE

While transformer-based NVS demonstrates extraordinary zero-shot generalization across diverse scenes, its minimal 3D inductive bias often leads to physically implausible floating artifacts, limb distortions, and severe temporal inconsistencies in dynamic human-centric scenarios. To address these limitations without sacrificing real-time inference speed, our Rosmi introduces a sophisticated cascaded architecture. This architecture is logically divided into three progressive functional stages: encoding, feature injection, and decoding.

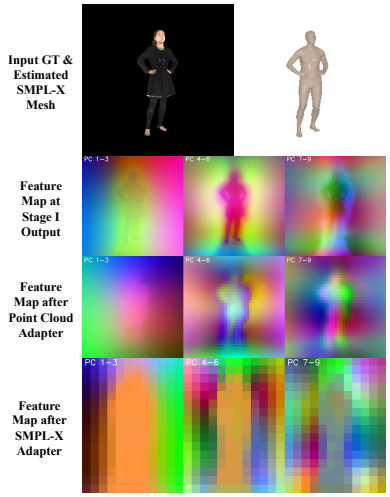


Figure 3: **Visualization of geometry feature injection.** By generating explicit spatial priors (point clouds and SMPL-X) inside the model, our architecture effectively anchors pure data-driven features to capture precise body topology and high-frequency details.

Stage I: Initial Global Context Encoding. The input sequence, comprising concatenated source and target tokens $\mathbf{T}_{in} = [\mathbf{x}, \mathbf{q}]$, is first processed in the encoding stage, which consists of M transformer blocks. This stage acts as a robust generalized feature extractor, establishing rudimentary cross-view correlations based purely on data-driven attention mechanisms. The output is an intermediate latent token representation $\mathbf{T}_{mid} \in \mathbb{R}^{(l_x+l_q)\times d}$, containing rich appearance textures but lacking explicit geometric constraints:

$$\mathbf{T}_{mid} = \mathcal{T}^{(0\to M-1)}\left(\mathbf{T}_{in}\right). \tag{5}$$

Stage II: Structural Geometry Prior Parsing and Feature Injection. Rather than relying on external auxiliary networks to extract structural priors during inference, Rosmi generates them from the intermediate tokens $\mathbf{T}_{mid}$ through two parallel prediction branches. The upper semantic branch employs a spatial adapter and an SMPL-X decoder (Wu et al., 2026) to predict $\mathbf{\Theta}_{smplx} = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\pi}_{cam}\}$, representing body shape, articulation pose, facial expression, and orthographic camera projection. The decoded SMPL-X vertices are projected into the target view using a differentiable rasterizer to obtain a dense pixel-aligned position map, which serves as a coordinate scaffold for querying a feature triplane and producing pose-aware spatial features $\mathbf{F}_{smpl}$. In parallel, to capture geometric details beyond the bare-body topology of SMPL-X, such as clothing and hair, the lower physical branch uses another adapter with a point map head (Ranftl et al., 2021; Liu et al., 2025) to decode $\mathbf{T}_{mid}$ into a 3D spatial representation, yielding a rendered point map and complementary geometric features $\mathbf{F}_{pts}$. Visualizations of the learned features are shown in Fig. 3.

To integrate these dual geometric features without destabilizing the pre-trained weights of the main backbone, both feature maps undergo spatial padding to align their dimensions and are processed by convolutional layers initialized with weights and biases set to zero, denoted by $\text{Conv}_{zero}(\cdot)$. Crucially, to further preserve low-level spatial frequencies and prevent gradient vanishing, we establish a long-skip residual connection directly from the initial tokenized sequence $\mathbf{T}_{in}$. The dual processed features and the residual (He et al., 2016) are combined with the intermediate tokens $\mathbf{T'}_{mid}$, modulated by three distinct learnable gating parameters γ_{res}, γ_{smpl}, and γ_{pts}:

$$\begin{aligned}\mathbf{T'}_{mid} = \mathbf{T}_{mid} &+ \gamma_{smpl} \cdot \text{Flatten}(\text{Conv}_{zero}(\mathbf{F}_{smpl})) \\ &+ \gamma_{pts} \cdot \text{Flatten}(\text{Conv}_{zero}(\mathbf{F}_{pts})) + \gamma_{res} \cdot \mathbf{T}_{in}.\end{aligned} \tag{6}$$

This Zero-Conv & γ Gating design (Zhang et al., 2023) ensures that, at the start of training, the model behaves identically to the vanilla pre-trained backbone, gradually learning to incorporate the structural guidance as the gating parameters γ update.

Stage III: Dual-Head Representation Decoding. The geometry-enhanced tokens $\mathbf{T'}_{mid}$ are passed through the decoding stage, which consists of N transformer blocks (Jin et al., 2024). This stage contextualizes the injected geometric priors with global appearance features, resulting in the final output tokens $\mathbf{T}_{out}$:

$$\mathbf{T}_{out} = \mathcal{T}_{dec}^{(0\to N-1)}(\mathbf{T'}_{mid}). \tag{7}$$

To facilitate comprehensive coarse-to-fine learning, Rosmi employs a dual-head decoding design rather than relying on a single monolithic output. The final tokens $\mathbf{T}_{out}$ are distributed to two parallel specialized decoding heads:

$$(\hat{\mathbf{I}}_t, \mathbf{\Theta}_{gs}) = (\text{NVS_Decoder}(\mathbf{T}_{out}), \text{3DGS_Head}(\mathbf{T}_{out})), \tag{8}$$

where the NVS decoder generates the synthesized novel view $\hat{\mathbf{I}}_t$, leveraging the rich appearance capacity of 2D generative networks, and the 3DGS head outputs volumetric consistency cues $\mathbf{\Theta}_{gs}$, enforcing strict geometric alignment and spatial coherence. This cascaded dual-head design ensures that the final synthesis benefits from both high-quality texture generation of 2D networks (via the NVS decoder) and the structural consistency of explicit 3D representations (via the 3DGS head adapter) (Ranftl et al., 2021; Liu et al., 2025).

3.3 Learning Objective

Rosmi is trained end-to-end using a composite multi-stage loss function to enforce explicit 3D geometry consistency during the intermediate feature injection (Stage II) and photometric fidelity during the final decoding stage (Stage III). The total loss is formulated as:

$$\mathcal{L}_{total} = \lambda_{nvs}\mathcal{L}_{nvs} + \lambda_{gs}\mathcal{L}_{gs} + \lambda_{pts}\mathcal{L}_{pts} + \lambda_{smplx}\mathcal{L}_{smplx}. \tag{9}$$

Intermediate Geometric Supervision: To guarantee that the generated structural priors possess strict spatial awareness, we apply intermediate supervision to both the point map head (Ranftl et al., 2021; Liu et al., 2025) and the SMPL-X decoder (Wu et al., 2026). For spatial geometry, the point cloud loss $\mathcal{L}_{pts}$ aligns the predicted 3D coordinates $\mathbf{X}$ with the pseudo-ground-truth $\mathbf{X}_{gt}$. To capture multi-frequency structural consistency and accommodate aleatoric uncertainty, we compute a multi-scale gradient loss, modulated by a predicted confidence map:

$$\mathcal{L}_{pts} = \sum_{i=1}^{N} \left\| C_i \odot (\mathbf{X}_i - \mathbf{X}_{gt,i}) \right\|_1 + \sum_{i=1}^{N} \left\| C_i \odot (\nabla\mathbf{X}_i - \nabla\mathbf{X}_{gt,i}) \right\|_1 - \alpha \sum_{i=1}^{N} \log C_i. \tag{10}$$

where ∇ denotes the spatial gradient operator, C_i denotes the predicted confidence at the i-th spatial location, $\odot$ denotes element-wise multiplication, and α controls the confidence regularization. The coordinate term enforces point-wise geometric accuracy, while the gradient term preserves local structural consistency and sharp boundaries.

Meanwhile, to enforce consistency in the learned geometric priors, the SMPL-X parameters predicted by the model are constrained using a mean squared error (MSE) against the teacher model's highly accurate SMPL-X parameters:

$$\mathcal{L}_{smplx} = ||\boldsymbol{\beta}_s - \boldsymbol{\beta}_t||_2^2 + ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_t||_2^2 + ||\boldsymbol{\psi}_s - \boldsymbol{\psi}_t||_2^2 + ||\boldsymbol{\pi}_s - \boldsymbol{\pi}_t||_2^2, \tag{11}$$

where $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\pi}$ denote the human body shape, body pose, facial expression, and camera projection parameters, respectively.

Final Photometric Supervision: To ensure high-fidelity visual appearance and authentic identity preservation, the Stage III decoding heads are strictly supervised by the ground-truth target images $\mathbf{I}_t$. The synthesized target image $\hat{\mathbf{I}}_t$ is optimized using a combination of MSE and the learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018) metric. This loss $\mathcal{L}_{nvs}$ captures both pixel-level accuracy and high-frequency perceptual textures:

$$\mathcal{L}_{nvs} = ||\hat{\mathbf{I}}_t - \mathbf{I}_t||_2^2 + \lambda_{lpips} \cdot \text{LPIPS}(\hat{\mathbf{I}}_t, \mathbf{I}_t). \tag{12}$$

Similarly, the explicit volumetric output rendered from the 3DGS head (Ranftl et al., 2021; Liu et al., 2025), denoted by $\hat{\mathbf{I}}_{gs}$, is supervised using an identical metric combination $\mathcal{L}_{gs}$ to enforce robust physical view consistency. We empirically set the balancing coefficients $\lambda_{nvs} = 1.0$, $\lambda_{gs} = 1.0$, $\lambda_{pts} = 100.0$, $\lambda_{smplx} = 1.0$, and $\lambda_{lpips} = 1.0$ during the optimization process.

3.4 Model Distillation

Despite the impressive synthesis quality achieved by the base architecture of Rosmi, the inherent quadratic computational complexity $\mathcal{O}(N^2)$ of standard multi-head self-attention fundamentally restricts its applicability (Vaswani et al., 2017; Dao et al., 2022; Dao, 2024; Yang & Zhang, 2024; Xie et al., 2024b). As the input image resolution increases, the token sequence length N grows

exponentially, making 24 FPS real-time inference mathematically impossible and imposing severe memory bottlenecks. The standard attention mechanism calculates the interaction matrix directly:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \tag{13}$$

To enable real-time performance, we replace standard attention with an advanced linear attention (Yang & Zhang, 2024; Xie et al., 2024b) mechanism, which reduces both time and memory complexity to $\mathcal{O}(N)$. By removing the non-linear softmax operation along the sequence dimension and applying a deterministic feature mapping kernel $\phi(\cdot)$ (e.g., ELU activation plus one) to the Query and Key matrices, we can mathematically exploit the associative property of matrix multiplication:

$$\text{LinearAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \phi(\mathbf{Q})\left(\phi(\mathbf{K})^T\mathbf{V}\right). \tag{14}$$

Directly replacing all attention modules and training from scratch causes catastrophic forgetting of pre-trained spatial priors, degrading texture quality and geometric stability. To address this, we propose a progressive linear attention stitching strategy, as illustrated in Fig. 4. The converged quadratic-attention model serves as the frozen teacher ($\mathcal{T}_{teacher}$), while the hybrid Rosmi network is the student ($\mathcal{T}_{student}$). We divide the $M+N$ transformer layers into groups of six attention blocks and, every 5,000 training steps, progressively replace full or flash attention (Dao et al., 2022; Dao, 2024) (F) with linear attention (Yang & Zhang, 2024) (L). To preserve global modeling capacity, the final attention layer in each group remains full attention. This "majority-linear, minority-full" design allows linear layers to efficiently process local features while periodic full-attention layers retain the global receptive field required for complex 3D reasoning.

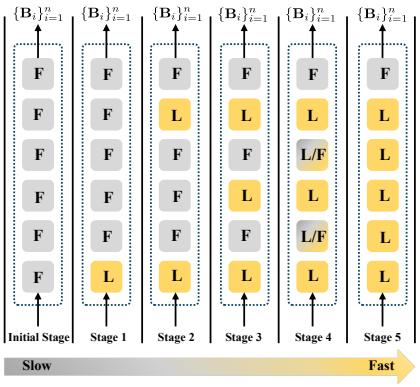


Figure 4: **Progressive distillation.**

To enforce representational consistency during this dynamic transition, we compute a pure distillation loss. We extract the intermediate hidden token states from uniformly sampled layers $\mathcal{S}$ in both the frozen teacher and the evolving student, minimizing their MSE loss:

$$\mathcal{L}_{distill} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} ||\mathcal{T}^{(i)}_{student} - \mathcal{T}^{(i)}_{teacher}||_2^2. \tag{15}$$

This loss is integrated into the total objective $\mathcal{L}_{total}$ in Eq. 9. Through this progressive distillation process, our lightweight hybrid student network effectively inherits the rich geometric and photometric knowledge of the quadratic baseline. This enables Rosmi to achieve robust high-fidelity synthesis at real-time frame rates, decisively breaking the traditional performance barrier.

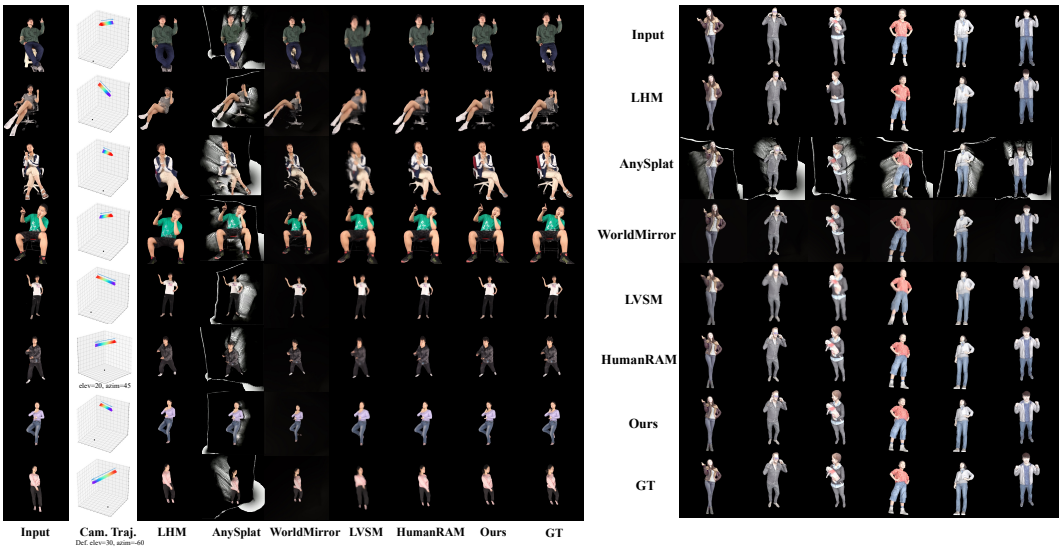
4 Experiment

4.1 Experimental Setup

Datasets and Metrics. We evaluate Rosmi on static, dynamic, and cross-domain datasets. For static scenes, we render training data from THuman2.1 (Yu et al., 2021) (2,445 identities), THuman3.0 (Su et al., 2023) (20 human-garment combinations), THumanSit (Zhang et al., 2025) (4,700 interacting poses), and 2K2K (Han et al., 2023) (2,050 poses), sampling 100 random views within a 30-degree, 2-meter spherical region around a central view. The test set contains 50 uniformly sampled identities from THuman2.1, THuman3.0, and 2K2K. For dynamic scenes, we use 6 ActorsHQ (Işık et al., 2023) identities with a 4:1 frame split for fine-tuning and testing, selecting 13 cameras within 30 degrees for supervision and camera 127 as the frontal input. For cross-domain zero-shot evaluation, we extract ∼2,000 frames from MVHumanNet (Xiong et al., 2024) (Part 1), using a randomly selected middle-level camera as input and a neighboring camera within 30 degrees as target. We evaluate visual fidelity using PSNR, SSIM, and LPIPS (Zhang et al., 2018) against avatar-based LHM (Qiu et al., 2025), feed-forward reconstruction models AnySplat (Jiang et al., 2025b) and WorldMirror (Liu et al., 2025), and NVS models LVSM (Jin et al., 2024) and HumanRAM (Yu et al., 2025). Computational efficiency is measured by inference FPS.

Table 1: **Quantitative comparison on static datasets.** The best and second best results are in **bold** and underlined.

res=512	THuman2.1			THumanSit			2K2K			FPS↑
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
LHM(Qiu et al., 2025)	24.75	0.9357	0.0623	22.80	0.9081	0.0913	21.06	0.8767	0.0943	0.66
AnySplat(Jiang et al., 2025b)	15.44	0.7365	0.2704	12.98	0.6141	0.3605	14.50	0.6847	0.2860	4.66
WorldMirror(Liu et al., 2025)	23.92	0.3289	0.0853	19.76	0.3008	0.1549	20.81	0.3182	0.0986	6.18
LVSM(Jin et al., 2024)	26.46	0.9452	0.0603	22.68	0.8851	0.1208	21.78	0.8897	0.0922	3.10
HumanRAM(Yu et al., 2025)	27.96	0.9551	0.0371	23.49	0.9010	0.0814	26.81	0.9369	**0.0343**	4.84
Ours	**31.23**	**0.9670**	**0.0280**	**26.88**	**0.9300**	**0.0563**	**27.57**	**0.9382**	0.0375	**24.07**



(a) Qualitative comparisons on THuman2.1 and THumanSit. (b) Qualitative comparisons on 2K2K.

Figure 5: **Qualitative comparisons on static human datasets.**

Implementation Details. The full-attention teacher model is trained on eight A6000 GPUs (batch size 32) for 20,000 steps. We then distill it into the linear-attention student model on a single A6000 GPU (batch size 4) for 40,000 steps. Fine-tuning on ActorsHQ is conducted for 6,000 steps using a single GPU. Linear attention is implemented based on SANA (Xie et al., 2024b), with layer parameters $M = 12$ and $N = 24$. All FPS tests are conducted on a single A6000 GPU. Additional implementation details are provided in Appendix A.

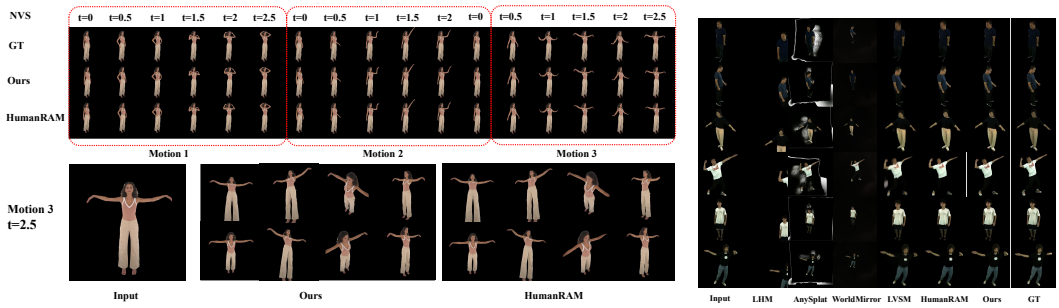
Table 2: **Comparison on video datasets.** Best results are in **bold**.

res=512	ActorsHQ		
	PSNR↑	SSIM↑	LPIPS↓
WorldMirror(Liu et al., 2025)	21.91	0.3944	0.0980
HumanRAM(Yu et al., 2025)	27.52	0.9466	0.0380
Ours	**29.62**	**0.9570**	**0.0307**

4.2 Comparison

As shown in Table 1, Rosmi significantly outperforms baselines on static datasets, achieving PSNRs of 31.23, 26.88, and 27.57 dB on THuman2.1, THumanSit, and 2K2K (vs. HumanRAM's 27.96, 23.49, and 26.81 dB). Our model also maintains SOTA or near-SOTA SSIM and LPIPS. This advantage stems from our dual-geometry prior injection, which anchors the feature space to deterministic topologies, avoiding depth ambiguity and allocating capacity to high-frequency texture synthesis. Moreover, linear attention boosts inference speed to 24.07 FPS (nearly 5× faster than HumanRAM). Qualitatively (Figs. 5a and 5b), baselines suffer from severe texture blurring and floating artifacts, especially on faces and hands.

For dynamic scenes shown in Table 2, Rosmi achieves a PSNR of 29.62 dB on ActorsHQ, which is competitive with HumanRAM and WorldMirror. However, as shown in Fig. 6a, baseline methods suffer from morphological distortions and flickering. In contrast, Rosmi's explicit geometric priors ensure strong spatiotemporal coherence. Finally, in cross-domain zero-shot evaluation on MVHumanNet, illustrated in Table 3 and Fig. 6b, Rosmi achieves a PSNR of 25.30 dB, demonstrating robust identity preservation on unseen data. More results are provided in Appendix B.



(a) Spatiotemporal consistency evaluation on ActorsHQ. (b) Cross-domain evaluation.

Figure 6: **Qualitative evaluation on dynamic and cross-domain datasets.** (a): **Rosmi** maintains stable rendering across continuous motion sequences on ActorsHQ. (b): **Rosmi** demonstrates strong zero-shot generalization on MVHumanNet.

Table 3: **Quantitative comparison on cross-domain datasets.**

res=512	MVHumanNet		
	PSNR↑	SSIM↑	LPIPS↓
LHM(Qiu et al., 2025)	21.95	**0.9359**	**0.0789**
AnySplat(Jiang et al., 2025b)	15.30	0.7598	0.2596
WorldMirror(Liu et al., 2025)	22.03	0.2256	0.1116
LVSM(Jin et al., 2024)	22.41	0.8779	0.1277
HumanRAM(Yu et al., 2025)	22.34	0.8851	0.0971
Ours	**25.30**	0.9029	0.0895

Table 4: **Ablation Study.** distil. and feat. denote distillation and feature, respectively.

res=512	Ablation Study			FPS↑
	PSNR↑	SSIM↑	LPIPS↓	
fully linear attention	25.90	0.9167	0.0815	**25.59**
w/o progressive distil.	22.94	0.8758	0.1321	–
teacher model	27.56	0.9376	0.0384	8.77
w/o pos. map feat.	26.43	0.9320	0.0387	–
w/o point map feat.	27.04	**0.9388**	0.0375	–
Full Model	**27.57**	0.9382	**0.0360**	24.07

4.3 Ablation Study

To validate our architectural decisions, we conduct ablation studies in Table 4. For dual-geometry injection, removing the position map reduces PSNR from 27.57 dB to 26.43 dB, as it provides foundational body topology and pose-aware structural guidance. In contrast, removing the point map reduces PSNR to 27.04 dB, as it captures deviations beyond the parametric body surface, such as loose clothing. These complementary priors jointly constrain both body structure and surface geometry. Without geometric priors, the network relies on implicit estimations, leading to significant degradation in spatial consistency and rendering sharpness.

Moreover, progressive distillation is critical for balancing speed and quality. Removing it causes catastrophic forgetting of spatial priors, drastically reducing the PSNR to 22.94 dB. A fully linear attention configuration (replacing all $M + N$ layers) marginally increases speed to 25.59 FPS but drops the PSNR to 25.90 dB due to limited representational capacity. Conversely, the full-attention teacher model yields a high PSNR of 27.56 dB but a sluggish 8.77 FPS, revealing the substantial computational cost of full attention. This comparison indicates that neither fully linear nor fully quadratic attention alone provides a satisfactory speed–quality tradeoff. Our method bridges this gap by preserving sufficient global modeling capacity while improving efficiency, achieving a 27.57 dB PSNR and inference at 24.07 FPS. More ablation results are provided in Appendix B.

5 Conclusion

In this paper, we introduce **Rosmi**, a novel geometry-guided framework for real-time high-fidelity human frontal view synthesis from a single image. By strategically restricting the synthesis to a practical frontal viewing frustum and decoding explicit structural priors from the backbone, e.g., SMPL-X meshes and 3D point clouds, our approach effectively eliminates the reliance on computationally intensive external conditions during inference. To further enhance efficiency, we propose a progressive cascade distillation strategy that dynamically transitions the network from quadratic full attention to lightweight linear attention. This key innovation addresses the traditional tradeoff between structural accuracy and inference speed, enabling Rosmi to achieve state-of-the-art visual authenticity and spatiotemporal consistency while delivering real-time performance at 24 FPS while preserving identity across viewpoint changes. Ultimately, Rosmi paves the way for accessible immersive 3D telepresence without the need for complex multi-camera setups.

AI USE STATEMENT

Large Language Models (LLMs) were used as supporting tools during manuscript preparation, primarily for language refinement and literature search. Their use included improving sentence clarity and readability, revising phrasing, checking grammar, and assisting in the discovery of potentially relevant prior work. All retrieved references and related information were independently examined and verified by the authors before inclusion in the manuscript.

LLMs did not contribute to the formulation of the research problem, development of the proposed method, experimental design, implementation, data analysis, or interpretation of the results. All technical ideas, methodological choices, experiments, analyses, and conclusions were produced by the authors. The authors reviewed all LLM-assisted material and remain fully responsible for the content, accuracy, and scientific integrity of this work.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of the results presented in this work. The paper provides detailed descriptions of the Rosmi framework, including its cascaded architecture, internal geometric modeling with SMPL-X (Pavlakos et al., 2019b) and point-cloud representations, feature injection mechanism, learning objectives, and progressive attention distillation strategy. We also report the training schedules, model configurations, optimization settings, camera setups, and hardware used throughout our experiments.

Our experiments cover static, dynamic, cross-domain, and in-the-wild settings. The public datasets used in this work include THuman2.1 (Yu et al., 2021), THuman3.0 (Su et al., 2023), THuman-Sit (Zhang et al., 2025), 2K2K (Han et al., 2023), ActorsHQ (Işık et al., 2023), and MVHuman-Net (Xiong et al., 2024). We further provide implementation details and controlled ablations for the major architectural components and efficiency-oriented designs. The core code, model configurations, and evaluation scripts will be publicly released after the review process to facilitate replication and future research.

ETHICS STATEMENT

This work studies real-time human frontal-view synthesis from a single image, with potential applications in telepresence and virtual communication. Rosmi is trained on publicly available human datasets and relies on pretrained components for geometric and visual representation learning. As a result, the model may inherit biases associated with the demographic distribution, appearance, clothing, poses, and capture conditions represented in the underlying training data.

Realistic human view synthesis can also be misused to create deceptive or non-consensual synthetic media, including impersonation of identifiable individuals. We therefore encourage appropriate consent when processing personal imagery, clear disclosure of synthesized content, and safeguards against deceptive or unauthorized use. Any released models and code are intended to support research and responsible development of human view-synthesis systems.

REFERENCES

Wenyue Chen, Peng Li, Wangguandong Zheng, Chengfeng Zhao, Mengfei Li, Yaolong Zhu, Zhiyang Dou, Ronggang Wang, and Yuan Liu. Synchuman: Synchronizing 2d and 3d generative models for single-view human reconstruction. *arXiv preprint arXiv:2510.07723*, 2025a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025b.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025c.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 (CVPR)*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 538–549, 2024.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

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.

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. *arXiv preprint arXiv:2410.01425*, 2024.

Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi: a language for high-performance computation on spatially sparse data structures. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. doi: 10.1145/3592415. URL `https://doi.org/10.1145/3592415`.

Xiaosong Jia, Yihang Sun, Junqi You, Songbur Wong, Zichen Zou, Junchi Yan, Zuxuan Wu, and Yu-Gang Jiang. Efficient-lvsm: Faster, cheaper, and better large view synthesis model via decoupled co-refinement attention. *arXiv preprint arXiv:2602.06478*, 2026.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4929, 2025a.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025b.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11127–11137, October 2021.

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.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *Advances in neural information processing systems*, 36:10516–10529, 2023.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021.

Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision*, pp. 451–468. Springer, 2024.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21159–21168, June 2023.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023.

Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.

Yixing Lu, Junting Dong, Youngjoong Kwon, Qin Zhao, Bo Dai, and Fernando De la Torre. Gas: Generative avatar synthesis from a single image. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12883–12893. IEEE, 2025.

Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*, pp. 179–197. Springer, 2022.

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

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019a.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019b.

Julius Plucker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, (155):725–791, 1865.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Lorenza Prospero, Abdullah Hamdi, Joao F Henriques, and Christian Rupprecht. Gst: Precise 3d human body from a single image with gaussian splatting transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6007–6017, 2025.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14184–14194, 2025.

Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pva: Pixel-aligned volumetric avatars. *arXiv preprint arXiv:2101.02697*, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

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

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

Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, and Cristian Sminchisescu. Diffhuman: Probabilistic photorealistic 3d reconstruction of humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1439–1449, 2024.

Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1581–1593, 2023. doi: 10.1109/TPAMI.2022.3168569.

Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 20–36, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning, 2025c. URL https://arxiv.org/abs/2507.13347.

Jiahao Wu, Yunfei Liu, Lijian Lin, Ye Zhu, Lei Zhu, Jingyi Li, and Yu Li. Pear: Pixel-aligned expressive human mesh recovery, 2026. URL https://arxiv.org/abs/2601.22693.

Desai Xie, Sai Bi, Zhixin Shu, Kai Zhang, Zexiang Xu, Yi Zhou, Sören Pirk, Arie Kaufman, Xin Sun, and Hao Tan. Lrm-zero: Training large reconstruction models with synthesized data. *Advances in Neural Information Processing Systems*, 37:53285–53316, 2024a.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformer, 2024b. URL https://arxiv.org/abs/2410.10629.

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, Shuguang Cui, and Xiaoguang Han. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19801–19811, June 2024.

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

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

Yuanyou Xu, Zongxin Yang, and Yi Yang. Photorealistic text-to-3d avatar generation with constrained geometry and appearance. In *European Conference on Computer Vision*, pp. 397–405. Springer, 2024.

Songlin Yang and Yu Zhang. Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism, January 2024. URL https://github.com/fla-org/flash-linear-attention.

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.

Zhiyuan Yu, Zhe Li, Hujun Bao, Can Yang, and Xiaowei Zhou. Humanram: Feed-forward human reconstruction and animation model using transformers. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–13, 2025.

Jiajun Zhang, Yuxiang Zhang, Hongwen Zhang, Xiao Zhou, Boyao Zhou, Ruizhi Shao, Zonghai Hu, and Yebin Liu. Ins-hoi: Instance aware human-object interactions recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

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 conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Qitao Zhao, Hao Tan, Qianqian Wang, Sai Bi, Kai Zhang, Kalyan Sunkavalli, Shubham Tulsiani, and Hanwen Jiang. E-rayzer: Self-supervised 3d reconstruction as spatial visual pre-training. *arXiv preprint arXiv:2512.10950*, 2025.

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.

Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Boyao Zhou, Shunyuan Zheng, Hanzhang Tu, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian+: Generalizable pixel-wise 3d gaussian splatting for real-time human-scene rendering from sparse views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Appendix

A More Implementation Details

A.1 Static Dataset Construction

To construct the static training dataset, we employ a customized rendering pipeline based on Taichi (Hu et al., 2019). For each 3D mesh, we normalize the human height to approximately 1.80 m (with a minor random perturbation of ± 0.05 m) and apply random horizontal and depth translations (up to ± 0.1 m) to increase spatial diversity. To ensure consistent frontal alignment, we utilize the SMPL-X (Pavlakos et al., 2019b) orientation parameters to correct the global facing direction. For each subject, we render a total of 105 views: 1 strict frontal center view (0° yaw, 0° pitch), 100 randomly sampled views that are uniformly distributed within a viewing frustum of $[-30^\circ, 30^\circ]$ for both yaw and pitch, and 4 extreme corner views ($\pm 30^\circ, \pm 30^\circ$). The rendering distance is fixed at a radius of 2.0 m. For evaluation, we uniformly sample 50 identities from THuman2.1 (Yu et al., 2021), THumanSit (Zhang et al., 2025), and 2K2K (Han et al., 2023), utilizing all 100 sampled views per identity for testing, while the remaining subjects are allocated strictly to the training set.

A.2 Dynamic Dataset Configuration

For dynamic evaluations on ActorsHQ (Işık et al., 2023), we account for the significant differences in camera coordinate systems by fine-tuning the model directly on this dataset. We select Sequence 1 from 6 identities (Actor01, Actor02, Actor04, Actor06, Actor07, and Actor08). From the approximately 2,500 frames per sequence, we allocate 2,000 frames for training and the remainder for evaluation. We designate Camera 127, which is positioned strictly in the frontal viewing direction, as the fixed input source view. During training, target supervision is sampled exclusively from cameras whose viewing angles fall within a 30° relative offset from Camera 127.

A.3 Cross-Domain and In-the-Wild Evaluation

To evaluate cross-domain zero-shot generalization, we randomly extract 2,000 frames from MVHumanNet (Xiong et al., 2024) (Part 1), which encompasses over 10 identities and 30 diverse appearances. Since the MVHumanNet camera rig is stratified into three elevation layers, we select the middle-tier cameras as the input source and randomly sample adjacent cameras within a 30° range as target views. For in-the-wild testing, we use high-resolution internet images (Chen et al., 2025a) or standard smartphone photographs, assuming a default frontal input view. We compute the foundational geometric constraints, such as the Plücker ray (Plucker, 1865) maps, by directly applying the camera intrinsic and extrinsic parameters derived from our static scene rendering pipeline.

A.4 Training Details

The training process is divided into two primary stages to ensure effective feature extraction and real-time efficiency:

- **Teacher Pre-training:** We initially train a teacher model with full attention to acquire strong geometric priors. This stage is conducted on 8 NVIDIA RTX A6000 GPUs for 20,000 steps with a batch size of 32. The framework is optimized using the AdamW optimizer with a weight decay of 1e-5. We apply decoupled learning rates: 1e-5 for the main LVSM (Jin et al., 2024) backbone and 2e-4 for the newly initialized dual-prior injection heads. We employ a Cosine Annealing learning rate scheduler, with a minimum learning rate of 1e-6.
- **Student Distillation:** The pre-trained teacher model is distilled into a lightweight student model over 40,000 steps on a single NVIDIA RTX A6000 GPU, with a batch size of 4. We use a fixed learning rate of 1e-6 for the distillation. To ensure representational consistency, an explicit MSE distillation loss $\mathcal{L}_{distill}$ is computed across the hidden states of evenly spaced layers between the frozen teacher and student models. Specifically, we employ a progressive linear attention stitching strategy: starting at step 5,000, standard attention

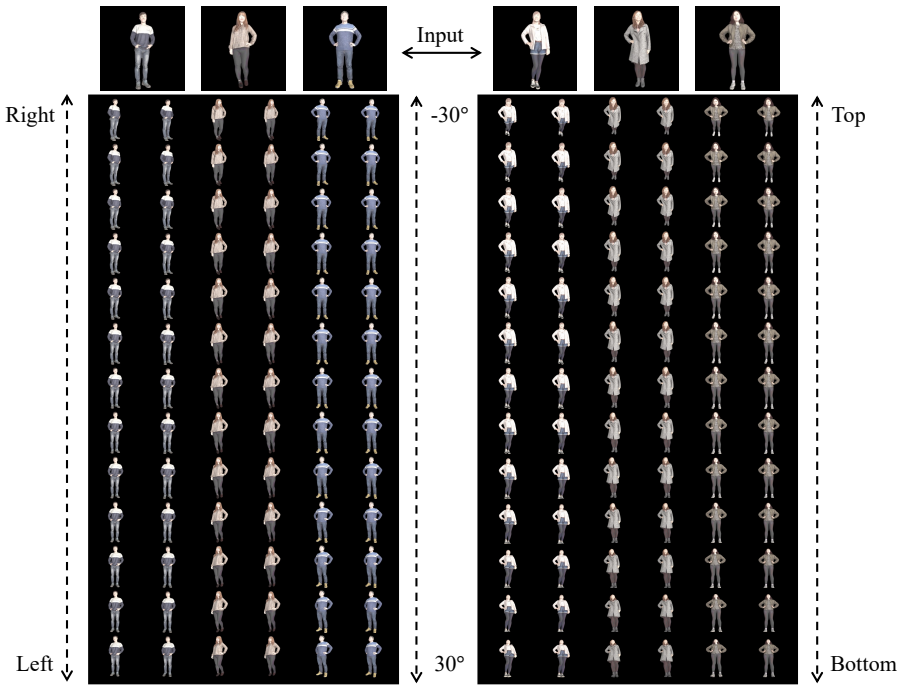


Figure 7: **Evaluation of spatial consistency in different directions. Left:** Results for 3 IDs during horizontal camera rotation, moving from right to left (from $-30°$ to $30°$). **Right:** Results for 3 IDs during vertical camera rotation, moving from top to bottom (from $-30°$ to $30°$). In both sets, images are rendered at $5°$ intervals. Rosmi synthesizes sharper details in high-frequency regions, such as faces and hands, compared to baselines, successfully avoiding severe floating artifacts.

modules are periodically replaced with linear attention equivalents every 5,000 steps. By step 35,000, 30 out of the 36 transformer blocks are converted to linear attention, while full attention is retained at fixed intervals (e.g., every 6 blocks) to maintain global receptive capacity. Finally, the distilled student model is fine-tuned on the ActorsHQ (Işık et al., 2023) dataset for 6,000 steps on a single GPU.

For both the pre-training and distillation stages, we incorporate dual geometric priors, i.e., SMPL-X (Pavlakos et al., 2019b) position maps and depth-guided point clouds, to improve structural and visual accuracy. Position maps are rendered using a differentiable rasterizer, NVDiffRasterizer (Laine et al., 2020), and decoded via tri-plane neural textures with a feature dimension of 32, while point clouds are processed using a Gaussian splatting renderer (Kerbl et al., 2023) to impose explicit physical constraints. These features are seamlessly integrated into the backbone using Zero-Conv (Zhang et al., 2023) gating mechanisms, parameterized by γ. This integration ensures that both structural and textural information is effectively utilized, enhancing the model's ability to synthesize visually consistent and geometrically accurate outputs.

B More Experiments

B.1 Spatiotemporal Consistency

We further evaluate the spatial and temporal consistency of Rosmi under continuous viewpoint and motion variations. As illustrated in Fig. 7, we assess spatial consistency under both horizontal and vertical camera rotations from $-30°$ to $30°$ at $5°$ intervals. Rosmi consistently preserves sharp details in high-frequency regions, such as faces and hands, while avoiding the floating artifacts commonly observed in baseline methods.

We further examine long-sequence temporal consistency on ActorsHQ (Işık et al., 2023). As shown in Fig. 8, we present results for Actor01, Actor02, and Actor04 across different viewpoints and mo-

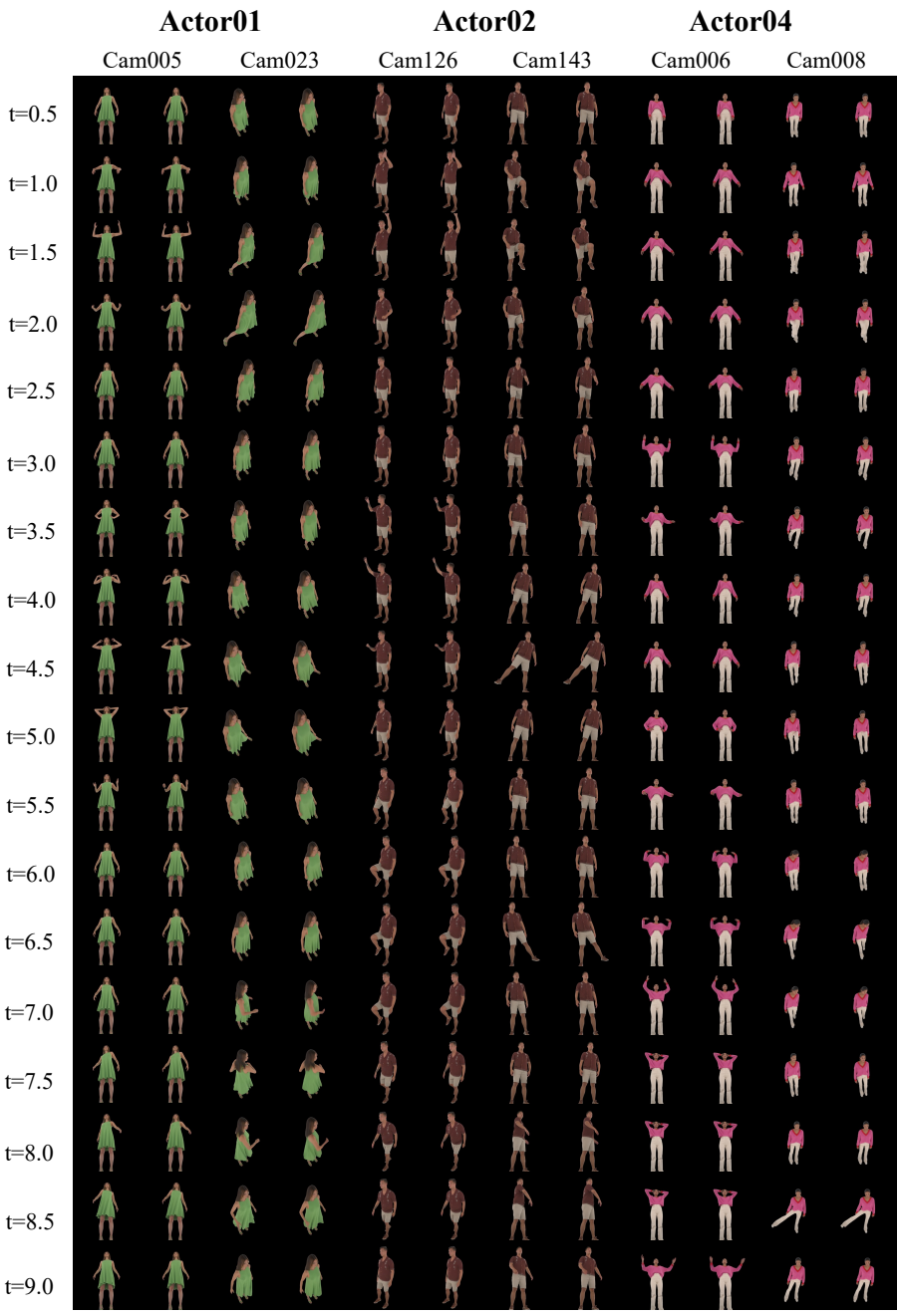


Figure 8: **Evaluation of temporal consistency on ActorsHQ.** Here, we present long-sequence results for Actor01, Actor02, and Actor04 from different viewpoints. Even with significant poses, our method still maintains temporal consistency and produces reasonable NVS results.

tion sequences. Even under large-scale and complex poses, Rosmi maintains stable human structure and temporally coherent novel views, demonstrating robustness to continuous motion.

In addition to the qualitative temporal comparisons above, we quantify inter-frame stability using a flow-based warping error. Given consecutive rendered frames I_t and I_{t+1} and an optical flow field $F_{t\to t+1}$, we compute



Figure 9: **Zero-shot evaluation results on in-the-wild cases.** To validate the practical applicability of our method, we conduct evaluations using in-the-wild data, including Internet images and real-world photographs. The visually appealing results demonstrate the strong generalization capability and robustness of our approach in unseen and complex scenarios.

$$E_{\text{warp}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left\| I_{t+1}(p) - W(I_t, F_{t \to t+1})(p) \right\|, \tag{16}$$

where $W(\cdot)$ denotes image warping and Ω denotes the evaluated pixel region. We report the metric both over non-occluded regions and over the complete image. Rosmi obtains a warping error of **0.0037** on non-occluded regions and **0.0045** over the full image. The consistently low errors complement the qualitative ActorsHQ results and quantitatively confirm that the rendered appearance remains coherent across consecutive frames with limited temporal flicker.

B.2 In-the-Wild Generalization

We additionally evaluate Rosmi on unconstrained in-the-wild inputs to examine its generalization beyond curated benchmarks. Figure 9 presents zero-shot results on diverse Internet images and real-world photographs. Despite substantial variations in appearance, pose, and image conditions, Rosmi produces visually coherent novel views, indicating strong generalization to unseen scenarios.

B.3 Geometry Prediction

Beyond rendering quality, we visualize the geometric representations learned by the model in Fig. 10. The point clouds and SMPL-X (Pavlakos et al., 2019b) meshes predicted by the 12-layer transformer encoder closely align with their ground-truth counterparts. These results verify that the encoding stage captures reliable human structural information, which subsequently serves as geometric guidance for feature refinement and view synthesis.

B.4 Additional Baseline Comparisons

We further compare Rosmi with two recent human synthesis methods, SyncHuman (Chen et al., 2025a) and GAS (Lu et al., 2025), on THuman2.1, THumanSit, and 2K2K. All methods are evaluated using the same input/target views, image resolution, and evaluation protocol. Table 5 reports both reconstruction quality and inference efficiency. Rosmi consistently achieves substantially higher PSNR and SSIM and lower LPIPS across the three datasets, while maintaining real-time inference at 24.07 FPS. In contrast, SyncHuman and GAS operate far below real-time speed. These results indicate that the performance advantage of Rosmi is not limited to the baselines considered in

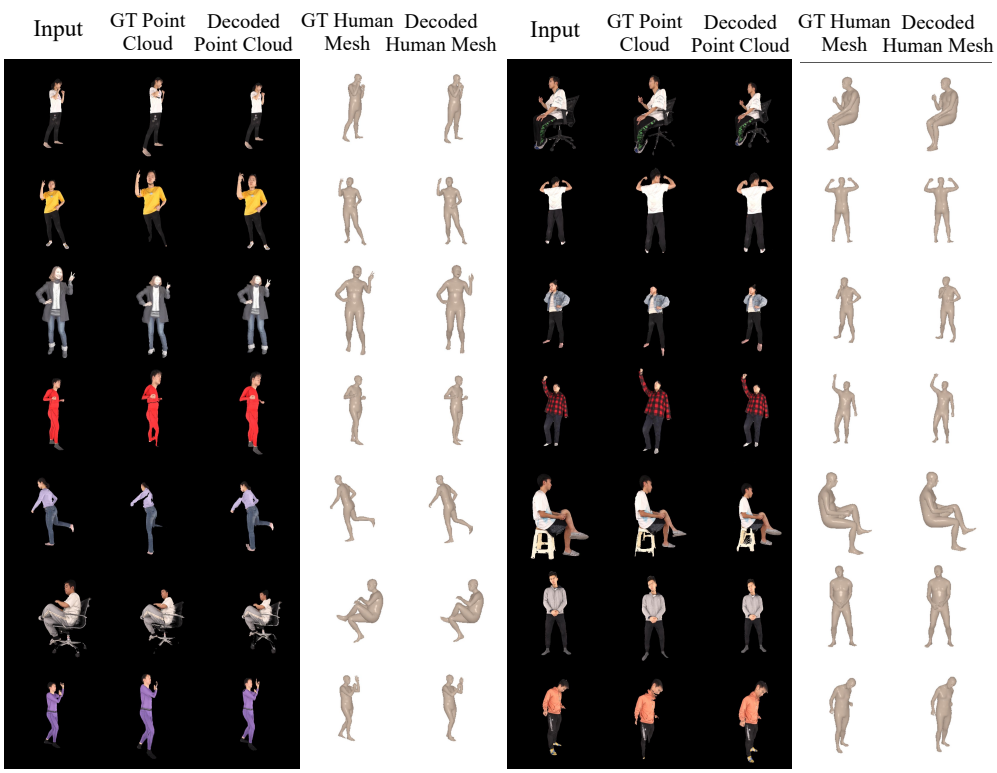


Figure 10: **Point cloud and human mesh from the encoding stage.** This figure illustrates the structural learning capability of our 12-layer transformer encoder. By comparing the predicted point clouds and SMPL-X meshes with their respective ground truth (GT), we observe a high degree of alignment. This indicates that our model effectively extracts precise structural information, ensuring strong geometric priors for the subsequent feature injection stage.

Table 5: **Comparison with additional baselines on static datasets.** We report rendering quality using PSNR, SSIM, and LPIPS together with inference speed.

Method	THuman2.1			THumanSit			2K2K			FPS↑
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
SyncHuman (Chen et al., 2025a)	22.74	0.9185	0.0802	17.13	0.7997	0.7997	21.45	0.8930	0.0760	0.012
GAS (Lu et al., 2025)	24.28	0.9210	0.0792	18.53	0.8166	0.8166	19.54	0.8637	0.8637	0.034
Ours	**31.23**	**0.9670**	**0.0280**	**26.88**	**0.9300**	**0.0563**	**27.57**	**0.9382**	**0.0375**	**24.07**

the main paper, but remains consistent when compared with additional recent human reconstruction and synthesis approaches.

Figure 11 provides representative qualitative comparisons. SyncHuman and GAS can recover the overall human appearance, but exhibit visible degradation around articulated structures and high-frequency regions. Rosmi better preserves body structure and local appearance details across different subjects and poses, consistent with the quantitative improvements in Table 5.

B.5 Identity Preservation

Human NVS is particularly sensitive to identity drift, since small appearance changes in faces, hands, clothing patterns, or other distinctive regions can strongly affect perceptual consistency. We therefore complement conventional pixel-level metrics with an identity consistency metric, denoted as IDSim. Specifically, we compute the cosine similarity between features extracted from corresponding human-body regions in the synthesized view and the source image. A higher score in-



Figure 11: **Qualitative evaluation results of SyncHuman and GAS on THuman2.0 and THumanSit datasets.**

Table 6: **Identity-preservation comparison.** IDSim measures the cosine similarity between human-region features from the synthesized image and the source input. Higher is better.

Dataset	**Ours**	HumanRAM	LVSM	WorldMirror	AnySplat	LHM	SyncHuman	GAS
THuman2.1	**0.8921**	0.8872	0.8893	0.8466	0.7644	0.8903	0.8235	0.8397
THumanSit	0.8841	0.8765	**0.8871**	0.7871	0.7761	0.8592	0.8111	0.8379
2K2K	**0.9325**	0.9281	0.9217	0.7065	0.7926	0.9136	0.9184	0.9057

dicates that the synthesized view better preserves identity-related appearance information from the input.

As shown in Table 6, Rosmi achieves the highest IDSim on THuman2.1 and 2K2K and the second-highest score on THumanSit. Importantly, the improvements are consistent across datasets with substantially different human poses and clothing distributions. This suggests that the explicit geometric guidance does not merely improve pixel-level reconstruction, but also helps stabilize appearance information associated with the input identity.

B.6 Real-Time Speedup

The runtime improvement of Rosmi originates from two complementary design choices rather than from a single acceleration module. Human-centric NVS pipelines that explicitly depend on external human reconstruction models typically contain two major latency components: geometric-prior estimation and subsequent NVS backbone inference. Rosmi addresses both components separately, as summarized in Table 7.

First, our internal geometric learning directly predicts and utilizes structural features inside the NVS framework, removing the need to invoke a separate SMPL-X estimation model during inference. This reduces the geometry-related latency from approximately 0.5 seconds in an external-prior pipeline to below 5 ms. Second, progressive attention distillation substantially accelerates the main transformer backbone, reducing its inference time from approximately 0.11 seconds to below 40 ms. The two improvements are complementary: internal geometry learning removes the auxiliary-model bottleneck, whereas attention distillation reduces the cost of the synthesis backbone itself.

This decomposition also explains why simply replacing the transformer backbone with a faster architecture is insufficient for real-time human NVS: an expensive external human reconstruction model would remain on the critical inference path. Rosmi instead learns the required human geometry internally and accelerates the remaining backbone computation jointly.

B.7 Telepresence Operating Range and Viewpoint Generalization

Rosmi is designed for interactive telepresence rather than unrestricted single-image 360° reconstruction. In this setting, the desired view is continuously adjusted around the frontal direction according to the observer's position. To examine this use case in practice, we construct a viewpoint-controlled

Table 7: **Runtime decomposition.** Rosmi reduces both geometry-prior acquisition and backbone inference latency.

Component	External-prior baselines	Ours
Geometry estimation (t_1)	∼0.50 s	<5 ms
Backbone inference (t_2)	∼0.11 s	<40 ms

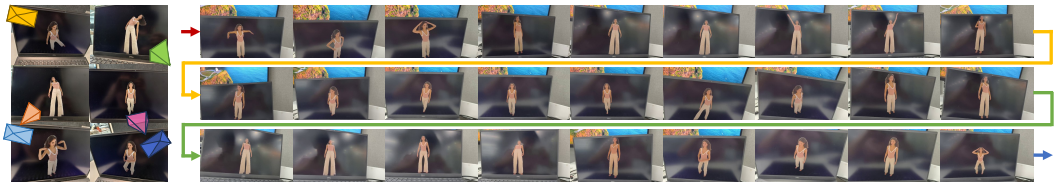


Figure 12: **Real-time viewpoint-controlled telepresence demonstration.** A front-facing camera captures the observer's viewing direction, and Rosmi continuously synthesizes the corresponding human viewpoint for display. This setup illustrates the intended frontal operating range of the model in an interactive telepresence scenario.

telepresence prototype in which a front-facing camera estimates the observer's viewing direction and dynamically updates the synthesized target view.

Figure 12 illustrates the resulting pipeline. The demonstration provides a practical interpretation of the $\pm 30^\circ$ operating range used during training: it covers the continuous viewpoint changes typically required when an observer moves around a frontal display while avoiding the highly ambiguous reconstruction of completely unseen rear body regions.

Although Rosmi is trained primarily within the frontal range, we further evaluate its zero-shot behavior outside this range. Table 8 reports results on 2K2K at progressively larger target-view offsets. Performance remains relatively stable from $\pm 15^\circ$ to $\pm 45^\circ$, while a more noticeable degradation appears at $\pm 60^\circ$. This behavior is expected because larger viewpoint changes expose increasingly large regions that are fully occluded in the input image.

Figure 13 further visualizes this trend. Rosmi continues to produce coherent novel views around $\pm 45^\circ$, despite these viewpoints extending beyond the main training range. At approximately $\pm 60^\circ$, artifacts become more apparent as previously invisible body surfaces occupy a larger portion of the target view. Importantly, the transition is gradual rather than an abrupt failure at the $\pm 30^\circ$ boundary.

For completeness, we also examine a full 360° single-image setting. Back-view synthesis falls below 20 dB PSNR because these regions contain little or no observable evidence in a frontal input image. This experiment further illustrates the fundamental ambiguity of unrestricted single-view human reconstruction and motivates our application-oriented frontal formulation.

B.8 Extended Architectural Ablations

We conduct a more comprehensive ablation study on 2K2K to disentangle the effects of feature injection, residual connections, linear attention, and progressive distillation. The results are summarized in Table 9.

Beyond the individual position-map and point-map ablations reported in the main paper, we first remove the entire geometry feature-injection pathway. The resulting PSNR decreases to 21.83 dB, confirming that explicit structural guidance plays a central role in connecting image appearance with 3D human structure. We also investigate how these features are injected. Directly concatenating geometric features with the backbone representation results in severe degradation, while removing the long residual pathway similarly destabilizes reconstruction. These observations support the use of the proposed gated residual injection rather than naive feature fusion.

We next vary the number of linear-attention layers. Increasing the use of linear attention progressively improves inference speed, while reconstruction metrics remain comparatively stable over a

Table 8: **Zero-shot generalization to extended viewpoints on 2K2K.** The model is evaluated beyond its primary frontal operating range without additional training.

View offset	$\pm 15^\circ$	$\pm 30^\circ$	$\pm 45^\circ$	$\pm 60^\circ$
PSNR↑	27.83	26.74	26.01	24.61
SSIM↑	0.9405	0.9352	0.9314	0.9197



Zero-shot evaluation results on 2K2K validation set (For top 2 rows, the camera moves from top 60° to bottom 60°, while for bottom 2 rows, the camera moves from left 60° to right 60°).

Figure 13: **Zero-shot extrapolation beyond the frontal training range.** Qualitative results on 2K2K for increasingly large viewpoint offsets. The reconstruction remains coherent around $\pm 45^\circ$, while larger appearance and geometry errors emerge near $\pm 60^\circ$ because of newly exposed occluded regions.

broad range. This suggests that a substantial fraction of the transformer can be efficiently linearized. However, replacing attention without an appropriate transition strategy is considerably less stable. The one-stage distillation variant drops to 22.94 dB PSNR, whereas the full-attention teacher preserves high reconstruction quality but runs at only 8.77 FPS. In comparison, the complete progressive strategy achieves 27.57 dB at 24.07 FPS, closely retaining the teacher's reconstruction quality while substantially improving inference efficiency.

B.9 Training-Time Teacher Cost

The full-attention teacher is required only during training to establish accurate geometric and appearance representations before distillation. It is trained for 20,000 steps, corresponding to approximately 8 hours in our setup. Once converged, the teacher is frozen and used to provide intermediate representations for student distillation. Importantly, the teacher is not part of the deployed inference pipeline. Therefore, the final runtime reported for Rosmi reflects only the lightweight student network and does not depend on an external teacher or auxiliary human reconstruction model.

C Failure Cases

While Rosmi achieves real-time high-fidelity synthesis within its targeted domain, it occasionally encounters limitations under highly challenging conditions. As shown in Fig. 14, we observe severe rendering distortions in scenarios involving extreme body poses or when the subject is in close proximity to the camera. We attribute these artifacts primarily to out-of-distribution (OOD) issues. The current training dataset lacks sufficient diversity and coverage for such extreme viewpoint and motion variations. To address these failure cases, future improvements will focus on expanding the richness of training data and incorporating more robust biomechanical constraints.

Table 9: **Extended ablation study on 2K2K.** We analyze geometric feature fusion, residual connections, the number of linear-attention layers, and different distillation strategies.

Metric	Direct concat.	w/o residual	w/o feat. inject.	4 linear	8 linear	12 linear	16 linear	One-stage distill.	Full-attn teacher	Ours
PSNR↑	9.86	13.83	21.83	27.60	27.63	27.58	27.60	22.94	27.56	**27.57**
SSIM↑	0.0174	0.0669	0.8891	0.9514	0.9457	0.9414	0.9413	0.8758	0.9376	0.9382
LPIPS↓	0.6662	0.5078	0.0895	0.0345	0.0353	0.0364	0.0363	0.1321	0.0384	0.0375
IDSim↑	0.5116	0.6611	0.8025	0.9315	0.9319	0.9322	0.9323	0.9321	**0.9333**	0.9325
FPS↑	–	–	–	11.012	14.018	17.307	20.012	24.10	8.77	**24.07**



Figure 14: **Failure cases.** We observe severe distortions in scenarios involving extreme poses or when the subject is in close proximity to the camera. These artifacts are mainly due to OOD issues, which arise from the limited diversity and coverage of such extreme cases in our current training dataset.