SEMGS-AVATAR: SEMANTIC-ALIGNED DISENTANGLED GENERATION OF 4D HUMAN AVATARS FROM A SINGLE IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Disentangled generation of 4D human avatars from a single image has achieved remarkable progress in recent years. However, existing methods suffer from two challenges: (1) Spatially, they struggle to accurately assemble separate body parts due to unclear rules on how parts fit together, leading to overlapping parts and broken boundaries. (2) Temporally, they fail to maintain smooth motion because estimated poses jump from frame to frame, causing the reconstructed avatar to shake and drift. To address these challenges, we propose **SemGS-Avatar**, a framework for high-fidelity 4D human avatar generation through semantic-aligned disentangled Gaussian splatting. Specifically, SemGS-Avatar consists of two core modules: Semantic-Aligned Gaussian Composer and Temporal Consistency Pose Optimization. The former explicitly aligns the semantic attributes of 3D Gaussians with the structured somatic manifold of the SMPL-X model to ensure region-consistent part binding and precise structural arrangement, enabling accurate body-part assembly and boundary refinement. The latter integrates a motion interpolation network with sliding-window kinematic smoothing to stabilize dynamic poses, ensuring smooth and coherent animation. Benefiting from the above designs, our method not only produces high-fidelity 4D human representations with sharper textures and cleaner part boundaries, but also reduces pose jitter and improves temporal stability. Extensive experiments demonstrate that our SemGS-Avatar outperforms existing methods in both visual fidelity and video-driven animation stability.

1 INTRODUCTION

In recent years, generating controllable 4D human avatars from single-image inputs has witnessed rapid progress (Pang et al., 2025). Compared with multi-view reconstruction pipelines that require complex capture setups, as well as traditional static 3D human generation methods that lack motion modeling, single-image 4D human avatar generation provides a more flexible and scalable solution with broader real-world applicability. Among these approaches, disentangled generation, which models body, clothing, hair, and accessories as separate components, has emerged as a promising paradigm, enabling finer texture generation and detail control (Nam et al., 2025a;b). Despite these advances, achieving high-fidelity reconstruction while maintaining stable and temporally coherent animation from a single image remains challenging, particularly when disentangled components must be composed into a structurally consistent and temporally stable whole.

In single-image 3D human generation, conventional methods often rely on holistic representations that lack explicit part-level disentanglement (Saito et al., 2019). As a result, fine-grained structures such as clothing wrinkles, layered garments, and boundary transitions between body parts are difficult to model accurately, leading to blurred textures or structural artifacts. To improve visual fidelity, several recent works introduce part-level disentangled representations, enabling separate modeling of clothing and body components (Nam et al., 2025a;b; Pang et al., 2025). However, **Spatially**, these methods typically operate without clear rules on how these separate body parts should fit together. Without explicit guidance to align the boundaries of independently generated components (such as a jacket and a torso), the assembled avatar often suffers from severe physical conflicts: parts either clip and overlap unnaturally, or exhibit noticeable stitching artifacts and unnatural deformations around

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.

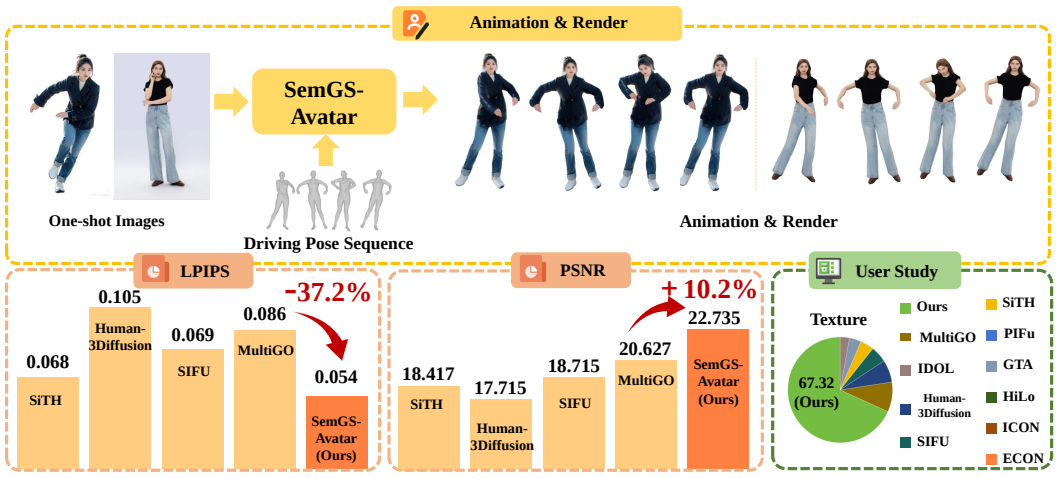


Figure 1: **Representative results of SemGS-Avatar.** Given a single reference image and a driving pose sequence, our method reconstructs high-fidelity 3D humans and generates stable 4D animations with superior PSNR, LPIPS, and user preference.

the boundaries (Xiu et al., 2023). Therefore, achieving semantically consistent and topologically aligned part assembly remains a key challenge. **Temporally**, extending static 3D models to video-driven 4D animation often fails to maintain smooth trajectories (Saito et al., 2020). Due to inaccurate inter-frame pose estimation, joint parameters frequently jump between consecutive frames. Lacking temporal constraints, this instability propagates to the avatar, causing it to constantly shake and drift over long sequences. This severely degrades realism, especially under fast motion, self-occlusion, or tracking interruptions.

To address the above challenges, we propose **SemGS-Avatar**, a framework for generating high-fidelity 4D human avatars from a single image through semantic-aligned disentangled Gaussian splatting. The core idea of SemGS-Avatar is to establish explicit semantic consistency between disentangled part-level 3D Gaussians and the anatomical structure of a parametric human model, ensuring that independently generated components remain spatially coherent during assembly and animation. Specifically, SemGS-Avatar consists of two core components. The **Semantic-Aligned Gaussian Composer** assigns semantic labels to both SMPL-X body parts and generated 3D Gaussians based on segmentation maps. By establishing a part-level semantic alignment mechanism, it enforces explicit semantic consistency between disentangled Gaussian components and the SMPL-X surface, enabling accurate body-part assembly, structural refinement, and clear boundary delineation. This design effectively alleviates cross-part misalignment and overlapping artifacts that commonly arise in part-based generation. To address instability in dynamic driving, the **Temporal Consistency Pose Optimization** integrates a motion interpolation network to recover missing or unreliable poses with sliding-window kinematic smoothing applied at test time. This combined strategy successfully suppresses high-frequency pose tracking noise, reducing inter-frame jitter and drift. Benefiting from these two complementary designs, SemGS-Avatar produces high-fidelity 4D human representations with sharper textures, cleaner part boundaries, and highly stable temporal behavior. Extensive experiments demonstrate that our SemGS-Avatar consistently outperforms existing state-of-the-art disentangled avatar methods in both visual fidelity and video-driven animation stability.

The main contributions of this work are summarized as follows:

- We propose SemGS-Avatar, a framework for single-image 4D human avatar generation that enforces spatial semantic consistency in disentangled Gaussian assembly and improves temporal stability in video-driven animation.
- To achieve precise semantic alignment and structural refinement in part-based assembly, we design a Semantic-Aligned Gaussian Composer that explicitly binds disentangled part-level 3D Gaussians to anatomically corresponding regions on an SMPL-X surface.

- We employ a test-time pose refinement strategy combining motion interpolation with sliding-window kinematic smoothing to address inaccurate inter-frame pose estimation and mitigate pose jitter and drift.

2 RELATED WORK

Single-View 3D Human Generation. Reconstructing 3D humans from a single image has evolved from implicit representations (PIFu (Saito et al., 2019; 2020), ICON/ECON (Xiu et al., 2022; 2023)) to diffusion-based approaches (TeCH (Huang et al., 2024), SiTH (Ho et al., 2024), PSHuman (Li et al., 2025), ConTex-Human (Gao et al., 2024), CHROME (Dutta et al., 2025)) and 3DGS models (HumanSplat (Pan et al., 2024), Human-VDM (Liu et al., 2024b), HumanRef (Zhang et al., 2024), MultiGO (Zhang et al., 2025a)). To improve controllability, disentangled methods (PARTE (Nam et al., 2025a), DeClotH (Nam et al., 2025b), Disco4D (Pang et al., 2025)) model garments and body parts separately. Concurrently, SAGA (Chen et al., 2024) and DAMA (Eskandar et al., 2026) explore mesh-aligned representations to constrain Gaussians via surface-adherence or body-anchoring. However, these methods rely on dense multi-view inputs or hard physical anchoring, which struggle under noisy monocular initializations. In contrast, our SemGS-Avatar tackles single-image 4D generation via a Semantic-Aligned Gaussian Composer. By introducing a soft semantic-topological binding mechanism, it explicitly maps part-level Gaussians to SMPL-X regions. This design avoids bad local minima and prevents cross-part texture bleeding under monocular occlusion, achieving structurally coherent composition and clean boundary delineation.

4D Human Animation. To reconstruct animatable representations rapidly, feed-forward (HumanRAM (Yu et al., 2025), LHM (Qiu et al., 2025a)) and prior-guided pipelines (AdaHuman (Huang et al., 2025), PERSONA (Sim & Moon, 2025), IDOL (Zhuang et al., 2025), TeRA (Wang et al., 2025)) have been developed. For dynamic modeling, foundational monocular Gaussian pipelines (HUGS (Kocabas et al., 2023), 3DGS-Avatar (Qian et al., 2024)) have been advanced by works targeting parameter-driven deformation (GaussianAvatar (Hu et al., 2024b), ParDy-Human (Jung et al., 2023)), structural alignment (HAHA (Svitov et al., 2024), EVA (Hu et al., 2024a)), drift mitigation (LoGAvatar (Zhang et al., 2025b)), and pose-driven rendering (AniGS (Qiu et al., 2025b), Animatable 3D Gaussian (Liu et al., 2024a), Disco4D (Pang et al., 2025)). Concurrently, generative 4DGS methods (DreamGaussian4D (Ren et al., 2023), Splat4D (Yin et al., 2025), GAF (Tang et al., 2024)) leverage diffusion priors for consistency. Nonetheless, frame-by-frame pose estimation in these pipelines still causes motion jitter and drift over long sequences. In contrast, our test-time refinement combines motion interpolation with sliding-window kinematic smoothing to suppress tracking noise and stabilize animations.

3 METHODOLOGY

3.1 PRELIMINARIES

3D Gaussian Splatting (3DGS). 3DGS models a scene using differentiable 3D Gaussian primitives. Each Gaussian is defined as $G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$, where μ denotes the spatial mean and Σ describes the spatial extent and orientation. Each Gaussian is associated with rotation r, scaling s, opacity α, and spherical harmonic coefficients c for view-dependent appearance. During rendering, 3D Gaussians are projected onto the image plane via differentiable splatting, and pixel colors are computed using depth-sorted alpha blending.

SMPL-X Parametrization. SMPL-X is a parametric human body model that defines a differentiable function $\mathcal{M}(\beta, \theta, \psi)$ mapping shape parameters β, pose parameters θ, and expression parameters ψ to a topologically consistent 3D human mesh. The shape parameters control body shape variations, pose parameters represent joint rotations, and expression parameters encode facial deformations.

Linear Blend Skinning (LBS). SMPL-X employs LBS to deform a canonical template mesh according to joint transformations. For a template vertex $\mathbf{v}_i$, its posed position $\mathbf{v}_i^{pose}$ is computed as $\mathbf{v}_i^{pose} = \sum_{k=1}^{K} w_{ik} T_k(\theta)\mathbf{v}_i$, where K is the number of joints, w_{ik} is the skinning weight of vertex i

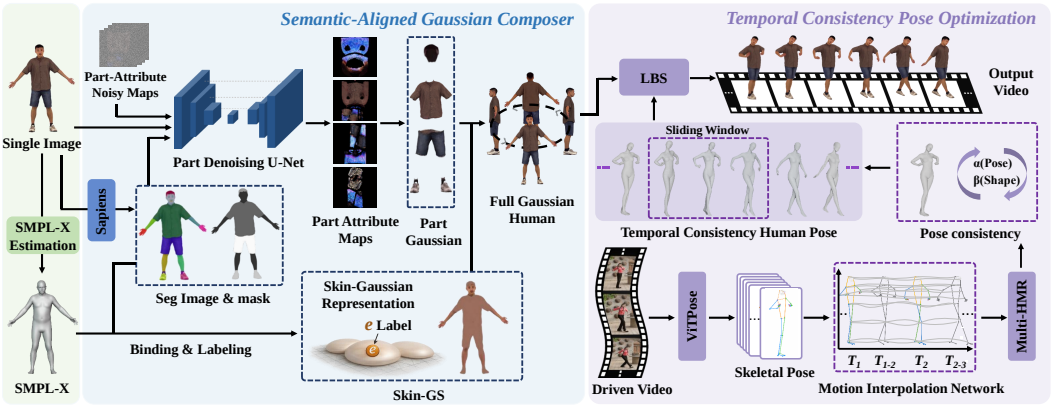


Figure 3: **Overview of SemGS-Avatar.** Given a single image and driving video, SemGS-Avatar generates high-fidelity 4D avatars through two stages. (1) Semantic-Aligned Gaussian Composer reconstructs disentangled part-level 3D Gaussians with explicit SMPL-X topology alignment via semantic binding. (2) Temporal Consistency Pose Optimization refines pose sequences using motion interpolation and sliding-window smoothing, then drives the avatar via LBS for stable animation.

for joint k, and $T_k(\theta)$ is the transformation matrix of joint k. This enables smooth articulation while preserving geometric continuity.

3.2 Overview

Given a single image and a driving motion video, we reconstruct a high-fidelity 4D human avatar that preserves part-level structural consistency and temporal coherence. As illustrated in Fig. 3, SemGS-Avatar consists of two stages: **Semantic-Aligned Gaussian Composer** and **Temporal Consistency Pose Optimization**. In the first stage, we assign semantic labels to Gaussian primitives based on predicted segmentation and SMPL-X parameters, then use a pretrained part-generation U-Net to reconstruct disentangled 3D Gaussian components that are explicitly aligned with a Gaussianized SMPL-X surface (Skin-GS), enforcing topology-aware binding and preventing cross-part misalignment. In the second stage, we employ a motion interpolation network combined with sliding-window kinematic smoothing applied at test time to refine driving poses and reduce temporal jitter.

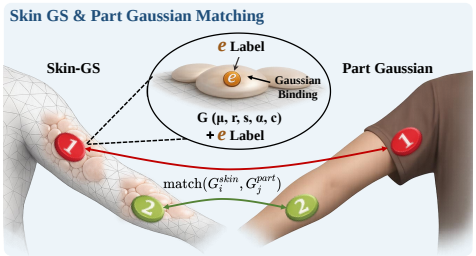


Figure 2: **Skin-GS and Part Gaussian Matching.** Under semantic-guided binding, Skin and Part Gaussians are paired to ensure consistent part-level correspondence.

3.3 Semantic-Aligned Gaussian Composer

This stage reconstructs a part-composable 3D Gaussian representation from a single image while preserving spatial semantic consistency and sharp boundaries. Specifically, we utilize a pretrained deep encoder-decoder network to predict part-level 3D Gaussian attributes, while simultaneously estimating SMPL-X parameters from the reference image. We then attach skin-consistent Gaussians to the SMPL-X surface, forming a semantic-aligned Skin-GS prior.

As shown in Fig. 2, unlike prior disentangled methods that generate part Gaussians in unconstrained individual spaces without structural coordination, we introduce an explicit semantic-topological binding mechanism realized via Deformable Semantic-Geometric Projection Mapping. By establishing explicit semantic consistency between decentralized part generative spaces and the structured somatic manifold of the SMPL-X surface, each Gaussian primitive is projected onto its semantically corresponding anatomical region. Formulating part assembly as a boundary-preserving constrained

manifold projection ensures structurally valid attachment under articulated deformation, effectively preventing cross-part misalignment, stitching artifacts, and unnatural boundary deformations.

During the generation process, we feed the input image, part-level Gaussian noise, and the semantic segmentation map S into the generator. The network outputs a set of pixel-aligned attribute maps $\{N_k\}_{k=1}^{K}$ for K semantic parts, parameterizing the localized Gaussian scales, rotations, opacities, and spherical harmonics coefficients. This generator is pre-trained under joint supervision of L1 pixel reconstruction, perceptual, and semantic cross-entropy losses to ensure high fidelity. Based on the predicted attribute maps, we construct a set of 3D Gaussians for each body part:

$$\mathcal{G}_k = \{g_i^k\}_{i=1}^{n_k}, \quad k \in \{1, \ldots, K\}, \tag{1}$$

where n_k is the number of Gaussians in the k-th part. Each Gaussian primitive is defined as:

$$g_i^k = (\mu_i^k, \Sigma_i^k, \alpha_i^k, c_i^k, e_i^k), \tag{2}$$

where μ_i^k, Σ_i^k, α_i^k, c_i^k, and e_i^k represent the Gaussian center, covariance, opacity, color, and semantic label, respectively. The semantic label e_i^k is explicitly assigned according to the segmentation map S and indicates the part category to which the Gaussian belongs.

In the construction of the Skin-GS representation, we first estimate the SMPL-X parameters from the single input image to obtain the parametric human mesh. We then attach a set of skin-colored Gaussians to the surface of the SMPL-X mesh to form the Skin-GS $\mathcal{G}_{\text{skin}}$. Specifically, for each surface sampling point or mesh face center, a Gaussian primitive is generated as follows:

$$\mathcal{G}_{\text{skin}} = \left\{g_j^s\right\}_{j=1}^{n_s}, \quad g_j^s = (\mu_j^s, \Sigma_j^s, \alpha_j^s, c_j^s, e_j^s), \tag{3}$$

where μ_j^s is determined by the center of the mesh face (or surface sampling point), and c_j^s is obtained by sampling the skin appearance color from the input image or modeled via learnable parameters to capture view-dependent appearance. With the assistance of the segmentation map S, we partition the surface of Skin-GS into different semantic regions and assign corresponding semantic labels e_j^s to the Gaussians in each region. As a result, every Gaussian on the Skin-GS carries an explicit part-level semantic label, forming semantic anchor points on the SMPL-X surface to which part-specific Gaussians can be attached.

After obtaining the semantically labeled part Gaussians $\mathcal{G}_k$ and Skin-GS $\mathcal{G}_{\text{skin}}$, we generate the complete 3DGS human representation through a semantic-consistent Gaussian binding and composition mechanism. Specifically, the Gaussians of part k are only allowed to be combined with or attached to the regions of Skin-GS whose semantic label is k :

$$\mathcal{S}_k = \left\{g_j^s \in \mathcal{G}_{\text{skin}} \mid e_j^s = k\right\}. \tag{4}$$

For each part Gaussian g_i^k, we search for a corresponding positional anchor within the set $\mathcal{S}_k$:

$$j^*(i) = \arg\min_{g_j^s \in \mathcal{S}_k} \|\mu_i^k - \mu_j^s\|_2^2, \tag{5}$$

where the anchor $j^*(i)$ serves as a semantic prior rather than a hard positional constraint, restricting Gaussian attachment to anatomically valid regions. Then the complete 3DGS human representation is constructed by combining the Skin-GS with all 3D part Gaussians, which can be expressed as:

$$\mathcal{G}_{\text{full}} = \mathcal{G}_{\text{skin}} \cup \bigcup_{k=1}^{K} \mathcal{G}_k. \tag{6}$$

To avoid motion-induced gradient discontinuities from the discrete anchor $j^*(i)$, we prune conflicting Gaussians based on geometric consistency and relax the assignment into a deformable, boundary-preserving soft projection with continuous weights w_{ij}:

$$w_{ij} = \frac{\exp\left(-\|\mu_i^k - \mu_j^s\|^2/\tau\right)}{\sum_{m \in \mathcal{S}_k} \exp\left(-\|\mu_i^k - \mu_m^s\|^2/\tau\right)}. \tag{7}$$

This weight defines the expected projection coordinates on the canonical skin manifold:

$$\bar{\mu}_i^k = \sum_{g_j^s \in \mathcal{S}_k} w_{ij} \mu_j^s. \tag{8}$$

The Boundary-Preserving Projection Constraint $\mathcal{L}_{\text{match}}$ is formulated as:

$$\mathcal{L}_{\text{match}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|\mu_i^k - \bar{\mu}_i^k\|_2^2. \tag{9}$$

To ensure photorealistic rendering, we combine $\mathcal{L}_{\text{match}}$ with a pixel-wise reconstruction loss $\mathcal{L}_{\text{rec}}$. The overall optimization objective is:

$$\mathcal{L}_{\text{composer}} = \mathcal{L}_{\text{rec}} + \lambda_m \mathcal{L}_{\text{match}}, \tag{10}$$

where $\lambda_m = 0.15$. We freeze the generator and optimize only the 3DGS representation. This constraint acts as a physical regularizer, forcing separate garment boundaries to dynamically slide along the underlying somatic manifold, eliminating overlapping artifacts and unnatural boundary deformations.

3.4 Temporal Consistency Pose Optimization

After obtaining the high-fidelity 3D Gaussian representation, we employ a test-time pose refinement strategy to stabilize the human deformation field. Because offline-generated 3DGS lacks temporal context, raw pose trajectories cause localized shearing and tearing in the dynamic Gaussian field. This module addresses these artifacts by governing LBS deformation gradients, combining motion interpolation with sliding-window kinematic smoothing to ensure temporal coherence without requiring additional training or model modifications.

Given a driving video sequence $\{V_t\}_{t=1}^{T}$, we first employ ViTPose to perform 2D keypoint detection on each frame, obtaining the human skeletal pose parameters:

$$\mathrm{H}_t = f_{\text{ViTPose}}(V_t), \tag{11}$$

where $\mathbf{H}_t$ denotes the set of 2D keypoints in the t-th frame. Unlike existing methods that simply copy adjacent poses when video tracking is interrupted or frames are missing, we introduce a lightweight, pretrained motion interpolation network f_{interp}. Built as a temporal convolutional network pre-trained on large-scale human motion database, the network reconstructs missing skeletal keypoints from local temporal context:

$$\tilde{\mathrm{H}}_t = f_{\text{interp}}(\mathrm{H}_{t-1}, \mathrm{H}_t, \mathrm{H}_{t+1}), \tag{12}$$

where $\tilde{\mathrm{H}}_t$ represents the completed, physically plausible skeletal pose. This step recovers plausible poses for frames with tracking failures or occlusions, producing a more continuous motion trajectory.

Subsequently, we employ Multi-HMR to map the completed 2D pose parameters into the SMPL-X parameter space, obtaining frame-wise 3D human pose estimations:

$$\mathcal{M}(\beta_t, \theta_t, \psi_t) = f_{\text{HMR}}(\tilde{\mathrm{H}}_t), \tag{13}$$

where β_t denotes the body shape parameters, θ_t represents the pose parameters, and ψ_t corresponds to the facial expression parameters. This step produces initial per-frame SMPL-X mesh estimations.

To further mitigate temporal jitter and body shape drift across frames, we incorporate temporal consistency constraints into the SMPLify framework, forming an improved Temporal SMPLify optimization strategy. While minimizing the reprojection error, we additionally impose temporal smoothness constraints on both pose and shape parameters. For each frame of the SMPL-X mesh pose, we minimize the following objective:

$$\mathcal{L}_{\text{consistency}} = \sum_{t=1}^{T} \left(\mathcal{L}_{\text{reproj}}^t + \lambda_s \mathcal{L}_{\text{shape}}^t + \lambda_p \mathcal{L}_{\text{pose}}^t \right), \tag{14}$$

where $\lambda_s = 0.01$ and $\lambda_p = 0.1$. The reprojection error is defined as:

$$\mathcal{L}_{\text{reproj}}^t = \left\| \Pi(\mathcal{M}(\beta_t, \theta_t, \psi_t)) - \tilde{\mathrm{H}}_t \right\|_2^2, \tag{15}$$

where $\Pi(\cdot)$ denotes the camera projection function. The shape consistency term and pose smoothness term penalize abrupt changes in body shape and pose parameters between consecutive frames:

$$\mathcal{L}_{\text{shape}}^t = \|\beta_t - \beta_{t-1}\|_2^2, \tag{16}$$

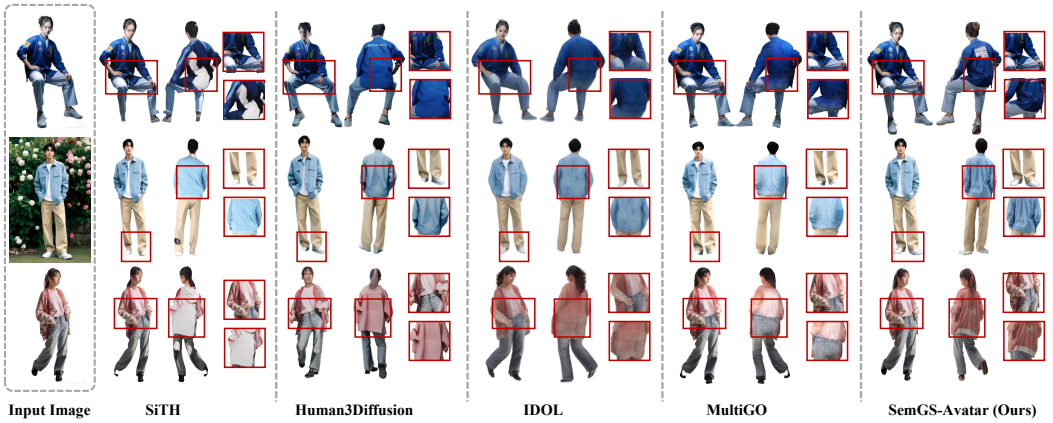


Figure 4: **Qualitative comparison of texture generation.** SemGS-Avatar produces sharper garment details and cleaner part boundaries through semantic-topological alignment, remaining robust under complex poses and in-the-wild images.

$$\mathcal{L}_{\text{pose}}^{t} = \|\theta_t - \theta_{t-1}\|_2^2. \tag{17}$$

To further suppress high-frequency jitter, we apply a sliding-window smoothing filter over the pose parameters within a local temporal window. For each frame t, we penalize abrupt acceleration in both pose and shape parameters:

$$\begin{aligned}\mathcal{L}_{\text{window}} = \textstyle\sum_{t=2}^{T-1} \Big(&\|\theta_{t+1} - 2\theta_t + \theta_{t-1}\|_2^2 \\ &+ \|\beta_{t+1} - 2\beta_t + \beta_{t-1}\|_2^2 \Big).\end{aligned} \tag{18}$$

This acceleration penalty mathematically enforces temporal continuity over the local LBS deformation gradient field by maintaining a smooth-varying Jacobian over time. By combining the above constraints, the overall optimization objective is formulated as:

$$\mathcal{L}_{\text{temporal}} = \mathcal{L}_{\text{consistency}} + \lambda_w \mathcal{L}_{\text{window}}, \tag{19}$$

where $\lambda_w = 0.05$. The resulting temporally coherent transformations are applied via LBS to deform the canonical Gaussians, preventing chaotic fluctuations in local Jacobian fields and eliminating high-frequency shearing.

4 Experiments

Implementation details. All experiments are conducted on NVIDIA RTX 4090 GPUs. During the Gaussian assembly stage, Gaussians are initialized from the SMPL-X surface and jointly optimized with part-level components. Attributes are optimized under standard 3DGS settings for 15,000 iterations using Adam (lr = 2e-5), with adaptive densification every 100 steps. For pose refinement, 2D keypoints from ViTPose are mapped to SMPL-X parameters via Multi-HMR. Test-time interpolation and sliding-window smoothing are then applied to mitigate jitter, driving the avatar via LBS. Remaining parameters follow the standard 3DGS framework.

Datasets. We fine-tune the part-generation U-Net on a composite training set sampled from THuman2.1 (Yu et al., 2021), THuman3 (Su et al., 2022), and CustomHuman (Ho et al., 2023). Multi-view images and segmentation masks are rendered from each scan to provide part-level supervision. For evaluation, we report results on THuman3 and CustomHuman for static reconstruction, and on UBC Fashion (Zablotskaia et al., 2019) for video-driven animation.

Evaluation Metrics. For quantitative evaluation, we adopt standard metrics: LPIPS (Zhang et al., 2018) for perceptual similarity, SSIM (Wang et al., 2004) for structural consistency, PSNR (Wang et al., 2004) for pixel-wise accuracy, FVD (Unterthiner et al., 2018) for video quality, and CLIP Score (Hessel et al., 2021) for semantic similarity between rendered output and reference image.

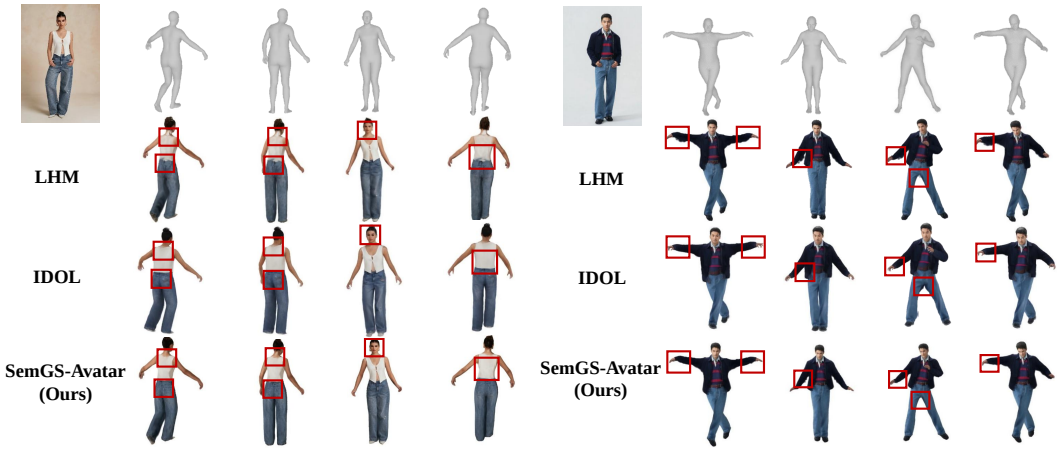


Figure 5: **Comparison of 4D human animation.** SemGS-Avatar avoids Gaussian drifting, produces more stable structures, better preserves body shape and facial identity.

Table 1: **Texture Quantitative Evaluation.** Lower LPIPS indicates better perceptual similarity, while higher SSIM and PSNR indicate better reconstruction quality. Results are reported for front (F) and back (B) views.

	CustomHuman			THuman3.0		
Method	**LPIPS (F/B)↓**	**SSIM (F/B)↑**	**PSNR (F/B)↑**	**LPIPS (F/B)↓**	**SSIM (F/B)↑**	**PSNR (F/B)↑**
GTA	0.073/0.089	0.900/0.892	18.790/18.229	0.063/0.077	0.930/0.928	21.113/20.497
Human-3Diffusion	0.105/0.125	0.858/0.841	17.715/16.628	0.063/0.075	0.913/0.904	20.803/18.500
SiFU	0.069/0.088	0.902/0.892	18.715/18.111	0.060/0.073	0.930/0.924	21.101/20.349
SiTH	0.068/0.084	0.901/0.887	18.417/17.608	0.061/0.077	0.921/0.911	20.326/19.355
IDOL	0.073/0.100	0.901/0.898	19.205/18.688	0.075/0.083	0.910/0.821	20.344/20.317
MultiGO	0.086/0.093	0.931/0.942	20.627/20.177	0.065/0.080	0.923/0.920	20.630/20.534
SemGS-Avatar (Ours)	**0.054/0.076**	**0.963/0.951**	**22.735/21.368**	**0.052/0.059**	**0.964/0.942**	**21.476/20.874**

4.1 Single-image 3D Human Generation

We conduct qualitative comparisons to evaluate texture fidelity against state-of-the-art approaches. As shown in Fig. 4, existing single-view methods often suffer from blurry textures and structural inconsistencies. Due to the inherent ambiguity of monocular inputs and the lack of explicit topological constraints, these methods frequently produce artifacts near part boundaries, including inconsistent clothing-body transitions, color bleeding, and over-smoothed details. Moreover, inaccurate geometric alignment further leads to texture distortions around articulated regions such as shoulders, elbows, and waist, degrading visual realism.

In contrast, SemGS-Avatar generates textures with clearer structural boundaries and more consistent appearance across body parts. Benefiting from the proposed semantic-aligned Gaussian composition module, our method explicitly enforces semantic-topological alignment between disentangled Gaussian components and the SMPL-X surface. This design effectively prevents cross-region texture contamination and ensures that garments, pants, and other components are strictly bound to their corresponding anatomical regions. Furthermore, the Gaussian-based representation preserves fine-grained details while maintaining global color coherence, leading to improved geometry-texture consistency and visually stable results.

In addition to qualitative comparisons, we report quantitative results in Table 1. Our method achieves the best performance across PSNR and other reconstruction metrics compared to existing approaches, validating the effectiveness of our semantic alignment strategy. Our method also remains robust under non-standard poses and in-the-wild images, producing visually plausible results in challenging scenarios.

4.2 Single-image 4D Human Animation

Table 2: **Ablation Study Results.** Results are reported for front (F) and back (B) views on CustomHuman.

Ablation	LPIPS:F/B↓	SSIM:F/B↑	PSNR:F/B↑
w/o Label e	0.073/0.091	0.898/0.850	21.340/20.890
w/o sliding window	0.071/0.090	0.886/0.803	21.365/20.652
w/o f_{interp}	0.069/0.087	0.918/0.901	21.734/21.140
Full Pipeline	**0.054/0.076**	**0.963/0.951**	**22.735/21.368**

Table 3: **Animation Quality and Efficiency.** Comparison of animate results for different methods.

Method	FVD↓	CLIP Score↑	FPS↑
Animate Anyone	148.8	86.6%	2
Champ	155.1	84.3%	1
DisCo4d	167.7	82.4%	1
SemGS-Avatar (Ours)	**143.6**	**88.4%**	**10**

To evaluate animation performance, we present qualitative comparisons using key frames sampled from video-driven sequences. As shown in Fig. 5, existing methods often suffer from texture blurring, pose instability, and inconsistent clothing deformation under fast motion or unreliable driving signals. These issues manifest as local texture flickering, shape drifting around articulated joints, and discontinuous garment silhouettes across adjacent frames. Table 3 shows our method achieves better animation quality and efficiency.

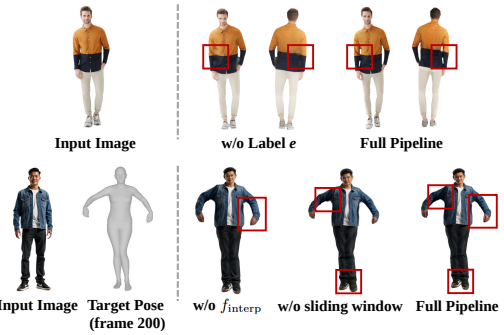


Figure 6: **Visual ablation results.** Comparison of visual ablations. The full model achieves better texture fidelity and animation details.

In contrast, SemGS-Avatar maintains clear and consistent representations across frames. Our temporal consistency optimization strategy enforces smooth pose transitions and suppresses frame-wise estimation noise. In highly articulated regions such as the shoulders, elbows, and waist, the body posture evolves smoothly without abrupt changes or temporal drifting. Meanwhile, clothing geometry preserves consistent shape and sharp boundaries throughout the sequence, without noticeable stretching, collapse, or blurring. Overall, our method achieves superior temporal stability in pose, garment structure, and texture clarity, producing more natural and coherent 4D animation.

4.3 Ablation Studies

To analyze the contribution of each component, we conduct ablation studies on the semantic labeling mechanism, motion interpolation network, and sliding-window kinematic temporal refinement. As shown in Table 2, removing any module degrades performance. Specifically, Fig. 6 shows that removing the semantic label constraint leads to more severe cross-part texture contamination, blurred boundaries, and structural inconsistencies, highlighting the importance of semantic-topological alignment for coherent part composition. Removing either motion interpolation or sliding-window refinement weakens temporal consistency, resulting in frame-to-frame pose fluctuations and reduced animation stability. In contrast, the full pipeline achieves the best overall performance, demonstrating that all components are essential for high visual fidelity and temporal consistency.

5 Conclusion

We propose **SemGS-Avatar**, a framework for high-fidelity, temporally consistent 4D human animation from single views. Unlike prior approaches that treat part-based generation and pose-driven animation as loosely coupled, our method jointly addresses spatial semantic consistency and long-term temporal coherence through semantic-consistent Gaussian binding and a test-time pose optimization. The Semantic-Aligned Gaussian Composer enforces explicit alignment between disentangled Gaussian and the SMPL-X surface, enabling structurally coherent part assembly with clear boundaries and high-fidelity textures. Meanwhile, the Temporal Consistency Pose Optimization reduces pose jitter and drift via motion interpolation and sliding-window kinematic smoothing. Extensive experiments on diverse datasets demonstrate the effectiveness of SemGS-Avatar, achieving state-of-the-art visual fidelity and animation stability.

REFERENCES

Ronghan Chen, Yang Cong, and Jiayue Liu. Saga: Surface-aligned gaussian avatar. *arXiv preprint arXiv:2412.00845*, 2024.

Arindam Dutta, Meng Zheng, Zhongpai Gao, Benjamin Planche, Anwesa Choudhuri, Terrence Chen, Amit K Roy-Chowdhury, and Ziyan Wu. Chrome: Clothed human reconstruction with occlusion-resilience and multiview-consistency from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9124–9135, 2025.

Daniel Eskandar, Berna Kabadayi, Garvita Tiwari, and Gerard Pons-Moll. Dama: Disentangled body-anchored gaussians for controllable multi-layered avatars. *arXiv preprint arXiv:2605.21001*, 2026.

Xiangjun Gao, Xiaoyu Li, Chaopeng Zhang, Qi Zhang, Yanpei Cao, Ying Shan, and Long Quan. Contex-human: Free-view rendering of human from a single image with texture-consistent synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10084–10094, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 21024–21035, 2023.

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.

Hezhen Hu, Zhiwen Fan, Tianhao Wu, Yihan Xi, Seoyoung Lee, Georgios Pavlakos, and Zhangyang Wang. Expressive gaussian human avatars from monocular rgb video. *Advances in Neural Information Processing Systems*, 37:5646–5660, 2024a.

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, 2024b.

Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pp. 1531–1542. IEEE, 2024.

Yangyi Huang, Ye Yuan, Xueting Li, Jan Kautz, and Umar Iqbal. Adahuman: Animatable detailed 3d human generation with compositional multiview diffusion. *arXiv preprint arXiv:2505.24877*, 2025.

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

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. *arXiv preprint arXiv:2311.17910*, 2023.

Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16008–16018, 2025.

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

Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models. *arXiv preprint arXiv:2409.02851*, 2024b.

Hyeongjin Nam, Donghwan Kim, Gyeongsik Moon, and Kyoung Mu Lee. Parte: Part-guided texturing for 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8547–8557, 2025a.

Hyeongjin Nam, Donghwan Kim, Jeongtaek Oh, and Kyoung Mu Lee. Decloth: Decomposable 3d cloth and human body reconstruction from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5636–5645, 2025b.

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. *Advances in Neural Information Processing Systems*, 37:74383–74410, 2024.

Hui En Pang, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Disco4d: Disentangled 4d human generation and animation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26331–26344, 2025.

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.

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 from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025a.

Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21148–21158, 2025b.

Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

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.

Geonhee Sim and Gyeongsik Moon. Persona: personalized whole-body 3d avatar with pose-driven deformations from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12670–12680, 2025.

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, 2022.

David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior. In *Proceedings of the Asian Conference on Computer Vision*, pp. 4051–4068, 2024.

Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, and Matthias Niessner. Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. *arXiv preprint arXiv:2412.10209*, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Yanwen Wang, Yiyu Zhuang, Jiawei Zhang, Li Wang, Yifei Zeng, Xun Cao, Xinxin Zuo, and Hao Zhu. Tera: Rethinking text-guided realistic 3d avatar generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10686–10697, 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.

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

Minghao Yin, Yukang Cao, Songyou Peng, and Kai Han. Splat4d: Diffusion-enhanced 4d gaussian splatting for temporally and spatially consistent content creation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 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.

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.

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019.

Gangjian Zhang, Nanjie Yao, Shunsi Zhang, Hanfeng Zhao, Guoliang Pang, Jian Shu, and Hao Wang. Multigo: Towards multi-level geometry learning for monocular 3d textured human reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 338–347, 2025a.

Jingbo Zhang, Xiaoyu Li, Qi Zhang, Yanpei Cao, Ying Shan, and Jing Liao. Humanref: Single image to 3d human generation via reference-guided diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1844–1854, 2024.

Jinsong Zhang, Xiongzheng Li, Hailong Jia, Jin Li, Zhuo Su, Guidong Wang, and Kun Li. Logavatar: Local gaussian splatting for human avatar modeling from monocular video. *Computer-Aided Design*, pp. 103973, 2025b.

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.

Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26308–26319, 2025.

A AI ASSISTANCE STATEMENT

During manuscript preparation, we used an artificial intelligence (AI) language model to assist with writing and language polishing, including grammar checking, sentence rephrasing, readability improvement, and clarity enhancement. The AI tool was not involved in the research ideas, methodology, experimental design, data collection, or data analysis; all scientific content and decisions were developed and verified by the authors. We ensured that the AI-generated or polished text complies with ethical guidelines and does not contribute to plagiarism or scientific misconduct.

B MORE 3D HUMAN GENERATION RESULTS

We provide more qualitative comparisons with representative methods on single-image 3D human generation. As shown in Fig. 8, SemGS-Avatar consistently produces higher-quality reconstructions compared with existing approaches. Specifically, many existing methods suffer from texture blurring, identity degradation, or structural inconsistencies, especially in fine regions such as the face and clothing textures. In contrast, SemGS-Avatar reconstructs clearer textures and more faithful facial details, preserving identity-related features while maintaining consistent body structure. Benefiting from the proposed Semantic-Aligned Gaussian Composer, our framework enforces semantic-to-topology consistency between Gaussian components and the underlying SMPL-X structure. This design enables more accurate part composition and prevents cross-part artifacts, resulting in sharper garment textures, clearer facial appearance, and more coherent overall geometry. The additional examples further verify that SemGS-Avatar achieves superior visual fidelity and structural consistency in 3D human reconstruction compared with existing SOTA methods.

C ADDITIONAL 4D ANIMATION RESULTS

We provide more video animation rendering results to further demonstrate the animation capability of SemGS-Avatar under diverse poses and motion patterns. As shown in Fig. 9, benefiting from the proposed Temporal Consistency Pose Optimization, SemGS-Avatar suppresses pose jitter and accumulated drift that commonly appear in frame-wise animation pipelines. As a result, the generated animations maintain smooth pose transitions and stable body structures, even under challenging articulated motions. In addition, the Semantic-Aligned Gaussian Composer maintains consistent attachment between body parts and the underlying SMPL-X topology, preventing cross-part deformation and structural inconsistencies during animation. This design helps preserve clear garment boundaries, stable clothing geometry, and detailed textures throughout the sequence. Overall, our method produces more stable and coherent animations across long sequences.

D ANIMATION RESULTS ON IN-THE-WILD IMAGES

To further evaluate the generalization capability of SemGS-Avatar, we additionally test our framework on unseen in-the-wild images collected from the Internet. These images exhibit diverse appearances, complex clothing styles, varying lighting conditions, and unconstrained poses that are not observed during training. Despite these challenges, SemGS-Avatar remains capable of reconstructing structurally coherent and visually realistic 3D human representations from a single input image. As shown in Fig. 10, our method produces clear texture details, consistent body structures, and stable animated results across these challenging inputs. These results demonstrate that SemGS-Avatar possesses strong generalization ability and scalability, making it applicable to a wide range of real-world human images beyond the training datasets.

E ANIMATION RESULTS ON CHALLENGE POSES

To evaluate the robustness of SemGS-Avatar on challenging body configurations, we further test our framework on inputs exhibiting complex and uncommon poses. These poses include extreme limb articulations and dynamic movements that are rarely observed during training. Despite these challenges, SemGS-Avatar consistently generates structurally coherent and visually plausible 3D human models from single inputs while producing temporally stable animations. As illustrated in

Table 4: **User Study of 4D Animation.** Higher values indicate stronger human preference.

Method	User Study (%) ↑
Animate Anyone	8.87
Champ	9.34
DisCo4D	9.15
IDOL	10.30
LHM	16.62
SemGS-Avatar (Ours)	**45.72**

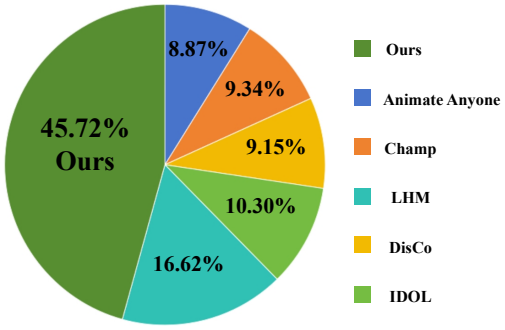


Figure 7: **User Study of 4D Animation.** Higher values indicate stronger human preference.

Fig. 11, our method maintains accurate pose representation, smooth motion transitions, and coherent body structures across frames. These results demonstrate that SemGS-Avatar can effectively handle complex human poses, producing high-quality reconstruction and animation even in difficult scenarios.

F Implementation Details

All experiments are implemented in PyTorch and conducted on NVIDIA RTX 4090 GPUs. The Gaussian parameters, including position, opacity, scale, rotation, and color, are optimized following the standard 3DGS pipeline. The pre-trained part-generation U-Net is fine-tuned on a composite dataset constructed from THuman2.1, THuman3, and CustomHuman, where multi-view renderings and segmentation masks provide part-level supervision. During optimization, the 3D Gaussian attributes are optimized for 15,000 iterations using the Adam optimizer with learning rates aligned with standard 3DGS parameters (e.g., 1.6×10^{-4} for position, 2.5×10^{-3} for scaling, and 2.0×10^{-5} for feature attributes), ensuring consistent and stable reconstruction. Adaptive densification and pruning are performed every 100 iterations to maintain efficient representation while preserving geometric details. For temporal consistency optimization, 2D keypoints are extracted from the driving video using ViTPose and mapped into SMPL-X pose parameters via Multi-HMR. A lightweight motion interpolation network refines the pose sequence to recover missing or occluded frames. We further apply the Temporal SMPLify optimization strategy with shape and pose smoothness terms during test-time refinement, combined with a sliding-window temporal smoothing filter to minimize pose jitter. The optimized SMPL-X poses are then applied to the Gaussian avatar via LBS to produce temporally coherent 4D human animations.

G User Study

To complement objective metrics, which often fail to fully capture perceptual realism and temporal dynamics, we conduct a user study to evaluate the effectiveness of SemGS-Avatar specifically for 4D human animation. We randomly sample 50 examples from the CustomHuman and UBC Fashion datasets to construct the evaluation set. Thirty participants are invited to compare our results with several representative SOTA methods through visual inspection. Participants assess the animations

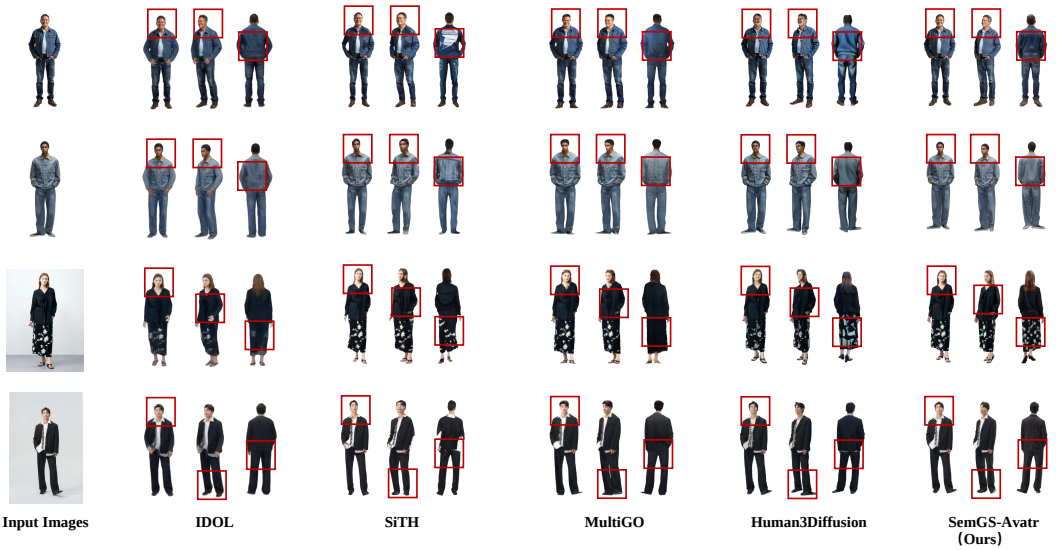


Figure 8: **More Single-image 3D Human Generation Results.** SemGS-Avatar consistently produces higher-fidelity reconstructions than existing methods, with clearer facial details, sharper garment textures, and improved structural consistency, effectively preserving identity-related features.

using a 5-point Likert scale (5 indicating the best quality) based on four animation-focused criteria: (1) motion naturalness and fluidity, (2) identity and appearance consistency across frames, (3) pose accuracy and structural stability, and (4) temporal coherence and smoothness of transitions. The user study results for 4D animation are presented in Table 4 and Fig. 7. SemGS-Avatar consistently receives higher preference scores than existing methods, highlighting its advantages in producing realistic and temporally stable human animations.

H Limitations and Future Works

While SemGS-Avatar demonstrates strong performance in both reconstruction quality and animation stability, there is still room for further improvement. In particular, the current framework relies on the quality of the estimated pose and parametric body model, which may introduce slight inaccuracies in extremely challenging cases. Future work will explore more robust motion modeling and temporal optimization strategies to further enhance animation stability under complex motions. In addition, integrating stronger generative priors could help improve the reconstruction of complex clothing structures and fine geometric details. Extending the framework toward more controllable and interactive human animation is another promising direction.



Figure 9: **Additional 4D Animation Results.** SemGS-Avatar generates smooth and temporally consistent human animations, preserving stable body structures, clear garment boundaries, and detailed textures.

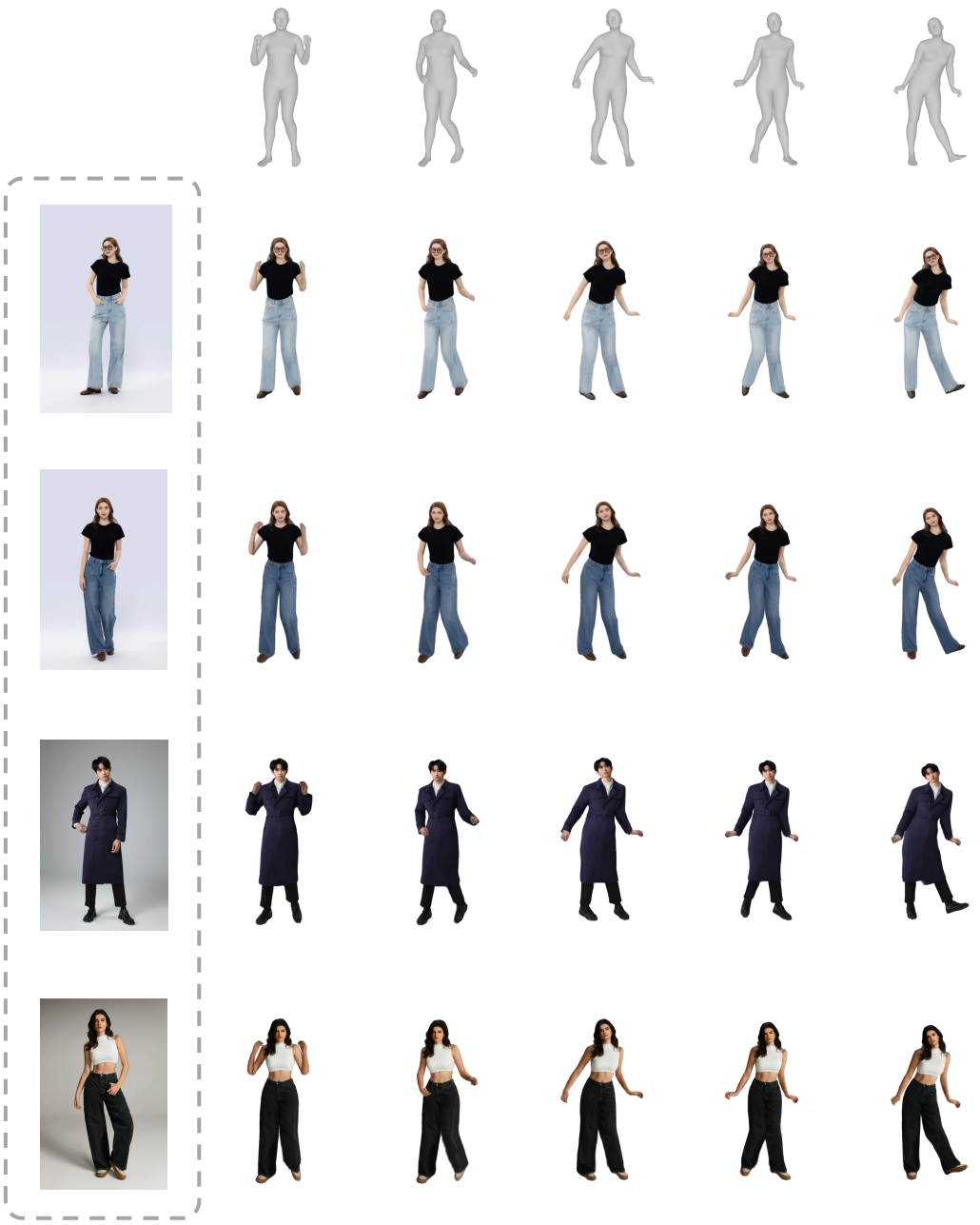


Figure 10: **Animation Results on In-the-wild Images.** SemGS-Avatar generates structurally coherent and visually realistic 3D humans from previously unseen real-world images, preserving clear textures, consistent body structures, and stable animations even under challenging conditions.



Figure 11: **Animation Results on Challenging Poses.** SemGS-Avatar can generate coherent 3D humans and stable animations from single images with complex or extreme poses.