2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3315-4364-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CVPR52734.2025.02453

StdGEN: Semantic-Decomposed 3D Character Generation from Single Images

Yuze He[1,2†] Yanning Zhou[1*] Wang Zhao[2] Zhongkai Wu[1] Kaiwen Xiao[1]
Wei Yang[1] Yong-Jin Liu[2*] Xiao Han[1]
[1] Tencent AI Lab [2] Tsinghua University

Figure 1. Our StdGEN generates high-quality, decomposed 3D characters from a single reference image.

Abstract

We present StdGEN, an innovative pipeline for generating semantically decomposed high-quality 3D characters from single images, enabling broad applications in virtual reality, gaming, and filmmaking, etc. Unlike previous methods which struggle with limited decomposability, unsatisfactory quality, and long optimization times, StdGEN features decomposability, effectiveness and efficiency; i.e., it generates intricately detailed 3D characters with separated semantic components such as the body, clothes, and hair, in three minutes. At the core of StdGEN is our proposed Semantic-aware Large Reconstruction Model (S-LRM), a transformer-based generalizable model that jointly reconstructs geometry, color and semantics from multi-view images in a feed-forward manner. A differentiable multi-layer semantic surface extraction scheme is introduced to acquire meshes from hybrid implicit fields reconstructed by our S-LRM. Additionally, a specialized efficient multi-view diffusion model and an iterative multi-layer surface refinement module are integrated into the pipeline to facilitate high-quality, decomposable 3D character generation. Extensive experiments demonstrate our state-of-the-art performance in 3D anime character generation, surpassing existing baselines by a significant margin in geometry, texture and decomposability. StdGEN offers ready-to-use semantic-decomposed 3D characters and enables flexible customization for a wide range of applications. Project page: https://stdgen.github.io

1. Introduction

Generating high-quality 3D characters from single images has widespread applications in virtual reality, video games, filmmaking, etc. Beyond automatically creating a complete 3D character, there is an increasing demand for the ability to produce decomposable characters, where distinct semantic components like the body, clothes, and hair are disentangled. This decomposition allows for much easier editing, control, and animation of characters, greatly enhancing their usability across various downstream applications.

*Corresponding Authors: amandayzhou@tencent.com, liuyongjin@tsinghua.edu.cn

†Work done during an internship at Tencent AI Lab.

2575-7075/25/$31.00 ©2025 IEEE
DOI 10.1109/CVPR52734.2025.02453

However, creating such decomposable characters from single images is challenging, as each component may face issues such as occlusion, ambiguity, and inconsistencies in their interactions with other components. Existing methods for decomposable avatar generation primarily focus on realistic clothed human models, exploring disentangled 3D parametric [53], explicit [34, 35], or implicit [10, 13, 17, 55] representations alongside various optimization techniques. These optimization approaches often employ score distillation loss [37] to leverage 2D generative priors, which leads to prolonged optimization times and the generation of coarse, high-contrast textures. Additionally, the dependence on parametric human models, such as SMPL-X [29], is inadequate for virtual characters, which often exhibit exaggerated body proportions and complex clothing designs.

To address these limitations, CharacterGen [36] was developed to efficiently generate characters from single images using a multi-view diffusion model and large reconstruction model [14]. Despite showing impressive generation capabilities in various posed images, CharacterGen can only produce holistic avatars in watertight meshes with no decomposability. These meshes require significant manual labor to separate, edit, or animate, limiting their applicability. Moreover, the quality of the generated meshes is often unsatisfactory, particularly in finer details such as the character's face and clothing, as shown in Fig. 5. Therefore, efficiently generating high-quality, decomposable 3D characters remains an open challenge.

In this work, we propose StdGEN, an efficient and effective pipeline to generate decomposable, high-quality, A-pose 3D characters from single images of any pose. At its core is a novel Semantic-aware Large Reconstruction Model (S-LRM), which innovatively introduces semantic attributes to the original Large Reconstruction Model [14], enabling efficient reconstruction of unified geometry, color and semantic fields. Moreover, a differentiable multi-layer semantic surface extraction scheme is proposed to extract decomposed semantic 3D surfaces from reconstructed implicit fields, empowering joint end-to-end training with explicit meshes. Together, S-LRM can efficiently generate semantic-decomposed, complete and consistent 3D surfaces from multi-view input images in a feed-forward manner.

Built upon S-LRM, StdGEN comprises three key stages: multi-view diffusion, feed-forward reconstruction, and mesh refinement, each designed with technical innovations to enable effective decomposition. Specifically, given a single reference image with an arbitrary pose, a specialized efficient multi-view diffusion model first generates multiple A-pose RGB images and normal maps at different camera views. Next, the generated images are processed by our novel S-LRM, to reconstruct geometry, color, and semantics in a feed-forward manner. Once the coarse decomposed surface generation is obtained, we further refine the mesh quality through a proposed iterative multi-layer mesh refinement method, using diffusion-generated 2D images and normal maps as guidance. As a result, StdGEN can generate high-quality decomposable 3D characters from a single reference image within minutes.

To facilitate the semantic decomposed training and testing, we further developed Anime3D++ dataset based on [36] , focusing on anime characters due to the wide internet presence, with finely annotated multi-view multi-pose semantic parts. Our experiments demonstrate that StdGEN surpasses all existing baselines, achieving state-of-the-art performance in both arbitrary and A-pose 3D character generation. Moreover, StdGEN's decomposable generation capability enables flexible customization, which can benefit numerous downstream applications. In summary, this work makes the following contributions:

- We introduce a novel Semantic-aware Large Reconstruction Model (S-LRM), which jointly models geometry, color, and semantic information with efficient tri-plane feed-forward inference. An effective differentiable surface extraction scheme is proposed to facilitate the explicit multi-layer semantic surface reconstruction.
- Building upon (1) our S-LRM, (2) an efficient multi-view diffusion model and (3) iterative multi-layer mesh refinement, we present StdGEN, an efficient pipeline for high-quality decomposed 3D character generation, which for the first time enables semantic-decomposed 3D character creation from arbitrary-posed single images in minutes.
- Our method demonstrates superior performance over existing baselines, with its decomposed modeling unlocking potential for various downstream applications.

2. Related Works

2.1. 3D Generation

To circumvent the need for extensive 3D assets during training, several approaches suggest lifting powerful 2D pre-trained diffusion models [9, 32, 42, 43] for 3D generation. The earliest works [37, 52] incorporate a pre-trained 2D diffusion model for probability density distillation using Score Distillation Sampling (SDS). These approaches gradually optimize a randomly initialized radiance field [1, 4, 49] with volume rendering, making it time-consuming to generate an object. Later research continues to enhance the aesthetics and accuracy of 3D content generation [6, 25, 45, 51, 57] and further investigate different application scenarios [11, 41, 44, 48]. However, relying solely on 2D priors for 3D generation often leads to poor geometry representation, e.g., multi-faced Janus problem, due to the challenges in controlling precise viewpoints through text prompts. The large-scale 3D datasets, e.g. Objaverse [8], unlock the possibility of imposing 3D priors to the model. Several works utilize view-consistent im-

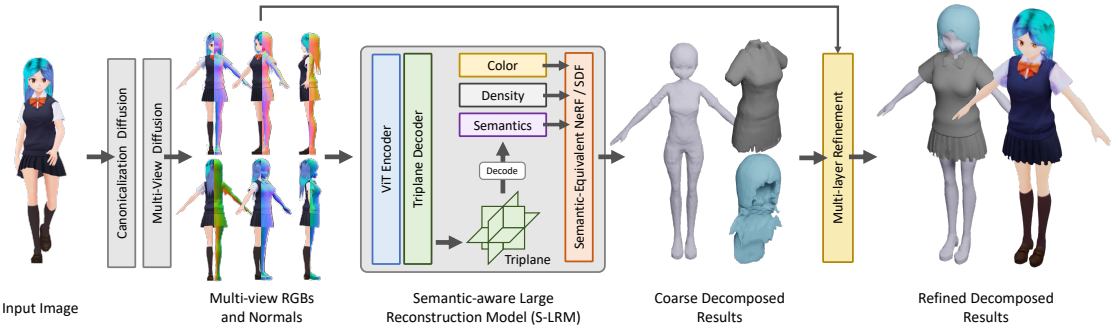

Figure 2. The overview of our StdGEN pipeline. Starting from a single reference image, our method utilizes diffusion models to generate multi-view RGB and normal maps, followed by S-LRM to obtain the color/density and semantic field for 3D reconstruction. Semantic decomposition and part-wise refinement are then applied to produce the final result.

ages to fine-tune the diffusion model. Zero-1-to-3 [26] integrates 3D priors into 2D stable diffusion by fine-tuning the pre-trained model for novel view synthesis (NVS). To further enhance the multi-view consistency, several recent works [19, 27, 28, 47] propose synchronously generating multi-view images in a single generation process and achieving constraints in 3D place through feature interaction in attention mechanism. Besides, the 3D native generation method shows powerful geometric generation ability [24, 30, 66]. However, the ability of these methods to follow instructions is typically moderate; therefore, they face challenges in achieving the desired outcomes in scenarios requiring precise restoration of reference images, e.g., 3D character generation.

2.2. Large Reconstruction Model

Large Reconstruction Model (LRM) [14] leverages the transformer-based model to map the single image feature to implicit tri-plane representation. Instant3D [22] extends LRM by feeding multi-view images instead of a single image. LGM [50], GRM [62] and GS-LRM [65] replace the 3D representation to 3D Gaussians, embracing its efficiency in rendering and low memory consumption. InstantMesh [60] and CRM [58] explicitly model the geometry by equipping the generative pipeline with FlexiCubes [46], achieving high-quality surface extraction and high rendering speed. The following works further explored applying advanced model architecture [63] or 3D presentation [5, 7], aiming to improve the efficiency, realism, and generalization of reconstruction. Integrating with multi-view diffusion models, these LRMs can achieve text-to-3D generation or single image-to-3D generation. Yet all these methods typically produce holistic models. In contrast, our method generates semantically decomposed characters, making downstream processing such as editing and animation much more efficient.

2.3. 3D Character Generation

3D character generation is a challenging problem due to its high precision requirements and the scarcity of data. One line of work leverages 3D-aware GANs to model the distribution of digital humans [2, 13, 20, 33, 64]. Recently the SDS-based methods have shown the possibility of generating a variety of stylized characters [3, 10, 18, 21, 53], yet it suffers from the long optimization times and the difficulty of meticulous style control. Frankenstein [61] concentrates on producing decomposed, textureless 3D meshes based on 2D layouts, restricting the potential for achieving high-fidelity reconstruction from the reference image. CharacterGen [36] calibrates input poses to canonical multi-view images via an image-conditioned multi-view diffusion model, followed by LRM for 3D character reconstruction and multi-view texture back projection, but still exhibits limited geometry and texture quality. Our approach, in contrast, employs a semantic-aware, feed-forward paradigm that generates high-quality, decomposable characters using only one forward pass from an arbitrary reference image, providing significant efficiency and quality improvement.

3. Method

The proposed StdGEN begins with multi-view canonical character generation (Sec. 3.1) from a reference image. To reconstruct a decomposable 3D character from multi-view images, we extend the LRM with a semantic field, enabling semantic-based layered generation (Sec. 3.2). Finally, a multi-level refinement process is designed to enhance the results, improving the geometric structure and providing more detailed textures (Sec. 3.3). An overview of the StdGEN pipeline is shown in Fig. 2.

3.1. Multi-view Generation and Canonicalization

Given a reference character image, our objective is to generate a set of multi-view images and corresponding normal maps that maintain 3D consistency. This process involves two steps: Canonicalizing an arbitrary reference image into an A-pose character and generating multi-view RGBs and normals from the A-pose representation.

Arbitrary Reference to A-pose Conversion. Directly reconstructing a 3D character model in arbitrary poses can be affected by self-occlusion from different viewpoints. There-

fore, we design to first canonicalize the character in 2D space. We select the A-pose as the target representation, as it is more conducive to subsequent multi-view generation, reconstruction, and applications like rigging. Following previous works [16, 36], we employ the Stable Diffusion [42] model, augmented with a ReferenceNet, to translate a 2D arbitrary posed reference image to A-pose image. The ReferenceNet helps to preserve the reference identity in the resulting image. We refer to [36] for more details.

Multi-view RGBs and Normals Generation. This stage provides comprehensive information for the subsequent S-LRM, with multi-view consistent normals crucial for capturing rich surface details in mesh refinement rather than relying on independent normal predictions. We adapt Era3D [23] to generate high-resolution RGBs and normals simultaneously. Using memory-efficient row-wise attention across views and between RGB and normal maps, we ensure fine-grained spatial consistency and enable higher output resolutions. Specifically, it takes a single A-pose image as input, producing orthographic projections of six viewpoints (elevation $0°$, azimuth $-90°, -45°, 0°, 45°, 90°, 180°$). For sharper character details, we increased the resolution through a progressive training approach, ultimately reaching 1024.

Compared with CharacterGen [36], our choice can simultaneously generate high-resolution, multi-view consistent normal maps for mesh refinement. Besides, the two-step design allows for improved editing in the 2D A-pose space, facilitating the generation of decomposed characters for enhanced 3D editing applications.

3.2. Semantic-aware Large Reconstruction Model

Once obtaining multi-view images, [36, 60] use transformer based sparse-view Large Reconstruction Model (LRM) to reconstruct a holistic 3D mesh. In contrast, we aim to generate characters with decomposed components, including the minimal-clothed human model, external clothing, and hair, to produce 3D character models that are not only visually accurate but also functionally versatile for various uses in 3D games and animation pipelines. To achieve this goal, we proposed the Semantic-aware Large Reconstruction Model (S-LRM), which extends NeRF/SDF to simultaneously encode semantics, appearance and geometry information in a feed-forward manner.

As shown in Fig. 2, the S-LRM takes multi-view images into tokens and then feeds into a transformer-based image-to-triplane decoder to output a triplane representation. The triplane features are decoded to obtain semantic, color, density/SDF information. To enhance the training efficiency and reconstruction quality, following InstantMesh [60], we first utilize triplane NeRF representation to render 2D images with volume rendering for training. We then switch to FlexiCubes [46] to extract explicit mesh from triplane-

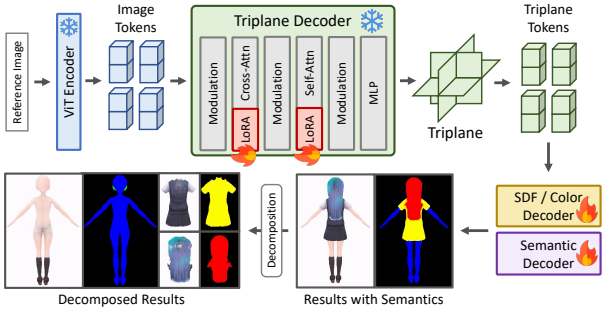

Figure 3. Demonstration of the structure and intermediate outputs of our semantic-aware large reconstruction model (S-LRM).

decoded SDF grid, and use direct mesh rasterization to render images. However, to obtain semantic-decomposed surface reconstruction, both NeRF and SDF implicit representations must be capable of rendering distinct semantic layers into images or extracting separate semantic surfaces using FlexiCubes in a differentiable manner. To achieve that, a novel semantic-equivalent NeRF/ SDF is proposed to extract character parts by specific semantics.

Semantic-equivalent NeRF and SDF. NeRF represents a 3D scene by spatial-variant volume densities with colors*. We extend it with a semantic field, and model them as a learnable function F_Θ that takes sampled point location $\mathbf{x} = (x; y; z)$ as inputs, and outputs color c, density σ and semantic distribution s as: $(\sigma, c, s) = F_\Theta(\mathbf{x})$.

To render per-pixel color $\hat{C}(\mathbf{r})$, a series of 3D points are sampled along the ray $\mathbf{r}$, and the pixel color is computed by integrating the sampled densities σ_i and colors c_i using the volume rendering equation with:

$$\hat{C}(\boldsymbol{r}) = \sum_{i=1}^{N} T_i \alpha_i c_i, \; T_i = \prod_{j=1}^{i-1}(1-\alpha_j) \tag{1}$$

where $\alpha_i = (1 - \exp(-\sigma_i \delta_i))$, $\delta_i = t_{i+1} - t_i$ is the alpha value of samples and distance between adjacent samples.

Given the probability $p_{s,i}$ of semantic s at location i, the pixel color $\hat{C}_s(\mathbf{r})$ under semantic s can be calculated as:

$$\hat{C}_s(\boldsymbol{r}) = \sum_{i=1}^{N} T_{s,i} p_{s,i} \alpha_i c_i, \; T_{s,i} = \prod_{j=1}^{i-1}(1-\alpha_j p_{s,j}) \tag{2}$$

If the probability of a certain semantic at a given location is zero, it should be considered fully transparent under the current semantic category. Furthermore, given that a position is known to be opaque, the probability of the current semantic should be linear to the final equivalent transparency.

Unlike NeRF, SDF does not incorporate the concept of transparency. Instead, positive/negative values represent points outside/inside the surface. Consequently, semantic probabilities cannot be directly applied to SDF for the mesh

*We ignore the view-dependent effects to simplify the discussion.

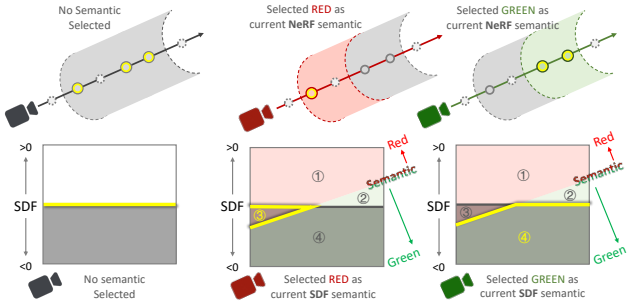

Figure 4. Our semantic-equivalent NeRF and SDF extraction scheme (shown in yellow color).

part extraction. Upon analysis, the extraction of a semantic-equivalent SDF should adhere to the following principles:

1. The zero value of the original SDF serves as a hard constraint. When the original SDF is positive, the equivalent SDF should also be positive;
2. When the original SDF is negative, the equivalent SDF should be zero at the boundaries where the maximum of relevant semantics transits;
3. At locations where the original SDF equals zero, but the probability of the current semantic is not the highest among all semantics, the equivalent SDF should not only maintain its sign but also be greater than zero.

Based on these principles, we propose the following formula for constructing the equivalent SDF:

$$f_{i,s} = \max(f_i, (\max_{r \neq s} p_{i,r}) - p_{i,s}), \tag{3}$$

Where f_i, $f_{i,s}$ are original SDF and equivalent SDF of semantic s at location i, respectively. Fig. 4 illustrates our method's scheme. For red semantics, only region 3 is selected, as regions 1, 2 (SDF>0) and region 4 (non-red) are discarded. Similarly, when the green is chosen, region 4 is correctly extracted. This formulation ensures correct decomposition by specific semantics and is fully compatible with subsequent FlexiCubes mesh extraction. In this way, we can differentiably extract multi-layer semantic surfaces from S-LRM's outputs, greatly facilitating the LRM training and downstream optimization.

Model Structure. Our S-LRM's core structure is derived from InstantMesh [60] with a ViT encoder, an image-to-triplane transformer, and the feature decoder. Several enhancements are made to this foundation to better suit our objectives, with the semantic decoder being a key addition. This component mirrors the structure of the density/color decoder and is designed to generate the semantic field in 3D space. For memory efficient training, we incorporate LoRA [15, 38] structures into all linear layers inside both self-attention and cross-attention blocks.

Multiple semantic level supervision. Current LRMs typically rely solely on 2D supervision, which limits their ability to generate information about objects' internal structures under occlusion; 3D supervision would be effective but often too resource-intensive. To address this, we propose an effective supervision that jointly learns semantics and colors, enabling the acquisition of a 3D semantic field and internal character information using only 2D supervision.

Stage 1: Training on NeRF with Single-layer Semantics. In this initial stage, we train on the triplane NeRF representation. We initialize the model with the pre-trained InstantNeRF, training the newly added LoRA in all attention blocks' linear layers and the newly introduced semantic decoder. We train it under the image, mask and semantic loss:

$$\hat{\mathcal{S}}(\boldsymbol{r}) = \sum_{i=1}^{N} T_i p_i \alpha_i, \ \mathcal{L}_{\text{sem}} = \sum_{k} CE(\hat{\mathcal{S}}_k, \mathcal{S}_k^{gt}) \tag{4}$$

$$\mathcal{L}_1 = \mathcal{L}_{\text{mse}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} \tag{5}$$

$\hat{\mathcal{S}}$ is the semantic map calculated by the probabilities p_i from semantic decoder's output through a softmax layer. $\hat{\mathcal{S}}_k, \mathcal{S}_k^{gt}$ denotes the k-th view of rendered and ground-truth semantic maps, and CE denotes the cross-entropy function.

Stage 2: Training on NeRF with Multi-layer Semantics. Having learned robust surface semantic information in the first stage, we aim to learn the 3D character's internal semantic and color information. We hierarchically supervise from outside to inside according to the spatial relationship of different semantic parts, by masking specific semantics during rendering and supervising with corresponding 2D ground truth. Assuming we aim to preserve a set of semantics $\{P_s\}$, we can render the image and semantic map under current conditions as follows:

$$\hat{C}_P(\boldsymbol{r}) = \sum_{i=1}^{N} T_{P,i}\alpha_i c_i \sum_{s \in P} p_{s,i}, \tag{6}$$

$$\hat{\mathcal{S}}_P(\boldsymbol{r}) = \sum_{i=1}^{N} T_{P,i}\alpha_i p_i \sum_{s \in P} p_{s,i}, \tag{7}$$

$$\text{where } T_{P,i} = \prod_{j=1}^{i-1}(1 - \alpha_j \sum_{s \in P} p_{s,j}), \tag{8}$$

The loss function is defined as:

$$\mathcal{L}_2 = \mathcal{L}_{\text{mse},P} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips},P} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask},P} + \lambda_{\text{sem}} \sum_{k} CE(\hat{\mathcal{S}}_{P,k}, \mathcal{S}_{P,k}^{gt}) \tag{9}$$

This decomposed training approach enables our S-LRM to simultaneously learn color and semantic information for the surface and the object's interior, thus achieving feed-forward 3D content decomposition and reconstruction.

Stage 3: Training on Mesh with Multi-layer Semantics. We switch to mesh representation [46] for efficient high-resolution training. We then extract the equivalent SDF via:

$$f_{i,P} = \max(f_i, (\max_{s \notin P} p_{i,s} - \max_{s \in P} p_{i,s})), \tag{10}$$

		A-pose Conditioned Input				Arbitrary-pose Conditioned Input			
		SSIM↑	LPIPS↓	FID↓	CLIP Similarity↑	SSIM↑	LPIPS↓	FID↓	CLIP Similarity↑
Multi-view Comparisons in 2D	SyncDreamer [27]	0.870	0.183	0.223	0.864	0.845	0.217	0.328	0.839
	Zero-1-to-3 [26]	0.865	0.172	0.500	0.885	0.842	0.209	0.481	0.878
	Era3D [23]	0.876	0.144	0.095	0.908	0.842	0.195	0.094	0.900
	CharacterGen [36]	0.886	0.119	0.063	0.928	0.871	0.139	0.056	0.919
	Ours	**0.958**	**0.038**	**0.004**	**0.941**	**0.920**	**0.071**	**0.014**	**0.935**
Character Comparisons in 3D	Magic123 [39]	0.886	0.142	0.192	0.887	0.849	0.197	0.256	0.862
	ImageDream [54]	0.856	0.171	0.846	0.836	0.823	0.218	0.875	0.818
	OpenLRM [12]	0.889	0.151	0.406	0.878	0.863	0.191	0.707	0.844
	LGM [50]	0.876	0.151	0.282	0.902	0.838	0.203	0.480	0.884
	InstantMesh [60]	0.888	0.126	0.107	0.906	0.846	0.202	0.285	0.886
	Unique3D [59]	0.889	0.136	0.030	0.919	0.856	0.190	0.042	0.903
	CharacterGen [36]	0.880	0.124	0.081	0.905	0.869	0.134	0.119	0.901
	Ours	**0.937**	**0.066**	**0.010**	**0.941**	**0.916**	**0.084**	**0.011**	**0.936**

Table 1. Quantitative comparison of A-pose and arbitrary pose inputs for 2D multi-view generation and 3D character generation.

Subsequently, we input the equivalent SDF into FlexiCubes to obtain the mesh, render the image and semantic map, and supervise using the following loss function:

$$\mathcal{L}_3 = \mathcal{L}_2 + \lambda_{\text{normal}} \sum_k M_P^{gt} \otimes \left(1 - \hat{N}_{P,k} \cdot N_{P,k}^{gt}\right) + \lambda_{\text{depth}} \sum_k M_P^{gt} \otimes \left\| \hat{D}_{P,k} - D_{P,k}^{gt} \right\|_1 + \lambda_{\text{dev}} \mathcal{L}_{\text{dev}} \quad (11)$$

where $\hat{D}_{P,k}$, $\hat{N}_{P,k}$, denotes the rendered depth and normal; $D_{P,k}^{gt}$, $N_{P,k}^{gt}$ and M_P^{gt} denote the ground truth depth, normal, and mask of the k-th view under semantic set P, respectively; $\mathcal{L}_{\text{dev}}$ denotes the deviation loss of FlexiCubes.

3.3. Multi-layer Refinement

Given the limitations in geometric and texture detail achievable by large reconstruction models, further optimization of the mesh post-reconstruction is necessary. Recent methods [24, 59] utilizing high-resolution normal maps for mesh optimization have shown promising results, albeit primarily for holistic meshes. We propose an iterative optimization mechanism for multi-layer mesh refinement.

To ensure thorough optimization at each level, we employ a staged approach following our S-LRM: Initially, we extract different parts of the mesh by specifying various semantics and optimize only the base minimal-clothed human model; upon completion, we overlay the clothing mesh, fixing the base and optimizing solely the clothing component; finally, we add the hair mesh, fixing the previous two layers and optimizing only the hair. The optimization process is guided by the multi-view normal maps generated earlier through diffusion models, each optimization step involves a differentiable rendering of the mesh to compute losses and gradients, followed by vertex adjustments based on these gradients, and re-mesh operations including edge collapse, split and flips. The loss function is defined as follows:

$$\mathcal{L}_{r1} = \lambda'_{\text{mask}} \sum_k ||\hat{M}_k - M_k^{\text{pred}}||_2^2 + \lambda_{\text{col}} \mathcal{L}_{\text{col}} + \lambda'_{\text{normal}} M_k^{\text{pred}} \otimes \sum_k ||\hat{N}_k - N_k^{\text{pred}}||_2^2 \quad (12)$$

where $\hat{M}_k$, $\hat{N}_k$ are rendered masks and normal maps, M_k^{pred}, N_k^{pred} are diffusion-generated masks and normal maps under k-th view, respectively. $\mathcal{L}_{\text{col}}$ is the collision loss modified from [35] to ensure outer-layer mesh in the outer normal direction of inner-layer mesh:

$$\mathcal{L}_{\text{col}} = \frac{1}{n} \sum_{i=1}^{n} \max\left((v_j - v_i) \cdot n_j, 0\right)^3 \quad (13)$$

where v_i represents the i-th vertex of the outer-layer mesh, v_j is its nearest neighbor of v_i on the inner-layer mesh, and n_j denotes the normal vector associated with v_j. Upon completing the optimization process, the mesh undergoes an additional ExplicitTarget Optimization phase, similar to that employed in Unique3D [59]. This stage aims to eliminate multi-view inconsistencies and further refine the geometry. Finally, the optimized meshes are colorized by the back projection of the multi-view images.

4. Experiments

4.1. Anime3D++ Dataset

We develop the Anime3D++ Dataset to meet the quality and decomposability requirements based on the original Anime3D [36] Dataset. We initially gathered around 14,000 3D anime character models from VRoid-Hub, then filtered these down to 10,811 high-quality models, standardized in A-pose with arms angled 45° downward. Renderings include RGB images, depth, normal, and semantic maps for layered supervision. Each character is rendered in full, only a base minimal-clothed human model and a base human model plus clothing. Please refer to the supplementary materials (Sec. E) for more details.

4.2. Results and Comparisons

Given that current methods cannot generate layered 3D models from a single arbitrary character image, we compare our non-layered generation results with other methods on the test split of our Anime3D++ dataset. We further decoupled the pose canonicalization component to ensure fairness, conducting both 2D and 3D comparisons for arbitrary

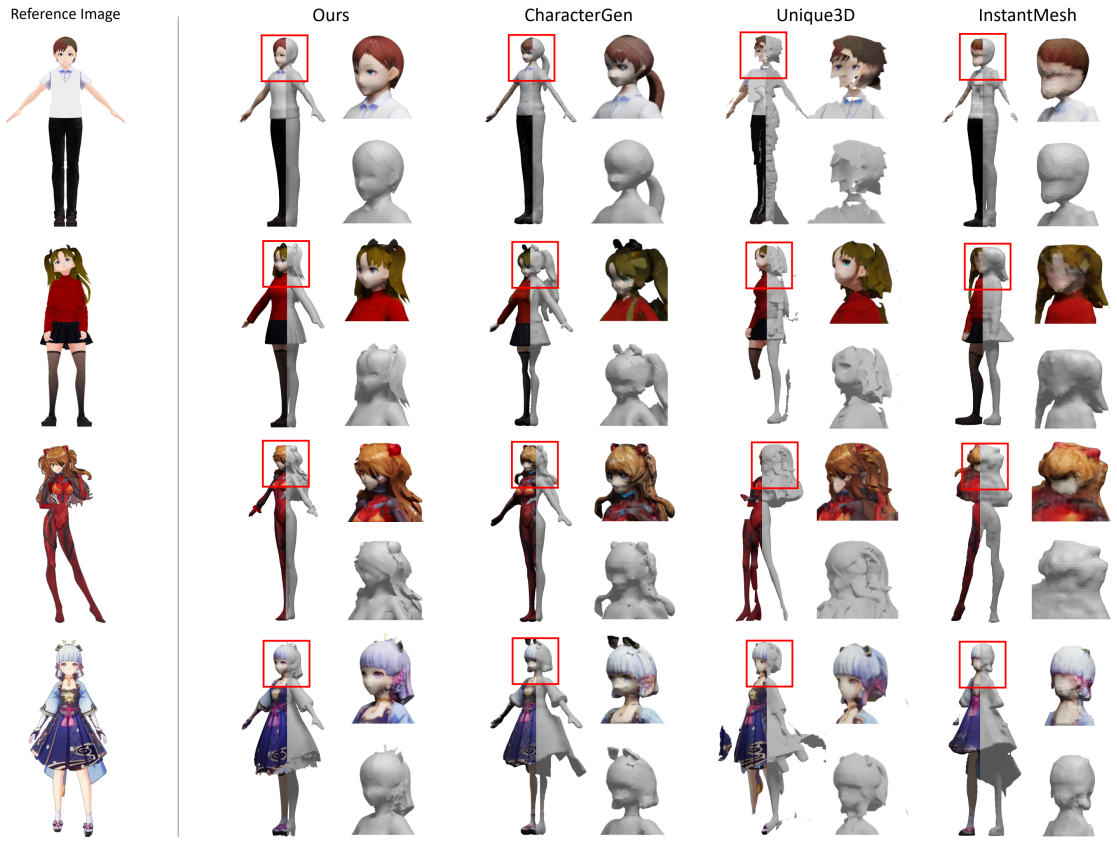

Figure 5. Qualitative comparisons on geometry and appearance of generated 3D characters.

pose and A-pose character reference inputs. When using A-pose inputs, all methods are compared against A-pose ground truth. For arbitrary pose inputs, we follow CharacterGen's settings, comparing our method and CharacterGen (both capable of canonicalization) against the A-pose ground truth, while other methods are compared under the original pose ground truth. Generation quality and fidelity are evaluated using SSIM [56], LPIPS [67], and FID. We also compute CLIP [40] cosine similarity between the front ground-truth image and the multi-views or 3D renderings.

Quantitative Results. For 2D multi-view generation results, we compare our method with Zero-1-to-3 [26], SyncDreamer [27], Era3D [23], and CharacterGen [36]. For 3D Character Generation results, we compare with Magic123 [39], ImageDream [54] (SDS-based optimization); OpenLRM [12, 14], LGM [50], InstantMesh [60] (feed-forward methods); Unique3D [59] (direct mesh reconstruction); and CharacterGen. 3D Results are uniformly rendered as eight equidistant images at elevation=0 and aligned using horizontal mask registration.

As shown in Tab. 1, Our method performs better in both standard and arbitrary pose settings. Existing 2D multi-view generation methods often struggle to preserve adequate geometric and appearance information in generated images. Among 3D methods, SDS-based approaches typically exhibit blurred geometry and suffer from the Janus Problem, and feed-forward methods generally lack geometric and texture precision. Unique3D achieves high metrics due to high-resolution supervision but suffers from unstable mesh initialization, impacting visual quality. CharacterGen demonstrates a notable advantage to other methods in arbitrary pose settings but the advantage diminishes in A-pose, showing effective pose canonicalization but limited reconstruction ability. In contrast, our method consistently achieves superior results across all cases.

Qualitative Results. Fig. 5 reveals several limitations across different methods. InstantMesh's results were constrained by grid resolution, leading to insufficient texture precision. While Unique3D achieved higher resolution, its heavy reliance on depth-based mesh initialization made it susceptible to geometric collapse under inaccurate depth estimations. CharacterGen exhibited low geometric and texture fidelity despite employing multi-view back-projection, and frequently produced visually disruptive black artifacts. In contrast, our method demonstrated superior geometric accuracy and texture fidelity performance.

Decomposed Results. Fig. 6 illustrates our method's capability to reconstruct decomposed characters from single reference images, organized from left to right as follows: the input reference image, the reconstructed 3D character (ap-

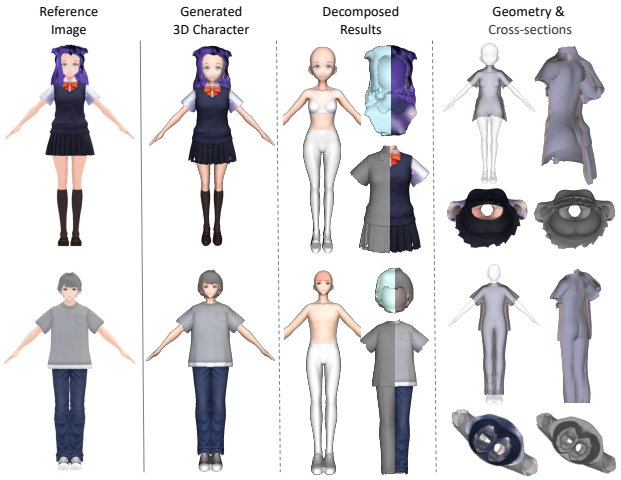

Figure 6. Decomposed outputs of our method, presented in texture, mesh, and cross-section.

Method	Overall Quality	Fidelity	Geometric Quality	Texture Quality
CharacterGen [36]	4.0%	4.5%	6.2%	4.2%
InstantMesh [60]	1.8%	2.0%	1.8%	2.5%
Unique3D [59]	11.8%	18.3%	10.5%	17.0%
Ours	**82.4%**	**75.2%**	**81.5%**	**76.3%**

Table 2. User study results comparing different methods across four dimensions. The values represent the percentage of times each method was selected as the best in the respective dimension.

plied shading for better visualization), the semantically decomposed components and geometries. The reconstructed meshes demonstrate high geometric precision, while the semantic decomposition is also accurate, successfully decoupling the base human model, clothing, and hair. This level of decomposition represents a significant advancement in character reconstruction from single images. The rightmost figure presents a cross-sectional view of the clothing, revealing that our reconstructed clothing is entirely internally hollow. This substantially enhances the potential for integration with downstream applications like realistic physics simulations and easier rigging.

User Study. To comprehensively evaluate our method's performance, we conducted a user study involving 28 volunteers. The study utilized 16 randomly selected 3D meshes and their corresponding reference images, drawn from both web-collected images and the Anime3D++ test split. Participants were asked to compare the randomly shuffled results of four approaches based on overall quality, fidelity, geometric quality, and texture quality. For each comparison group, participants identified the best result in each dimension. Tab. 2 shows that our method achieved superior performance across all dimensions, demonstrating its effectiveness in generating high-quality 3D characters.

Ablation Study. We further conduct ablation studies on character decomposition and multi-layer refinement. Please see supplementary material (Sec. A) for details.

4.3. Applications

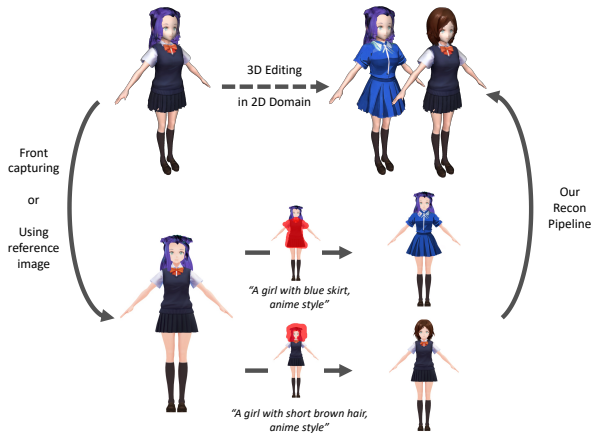

Figure 7. Our pipeline enables diverse 3D editing using only text prompts, masks, and in-painting diffusion in the 2D domain.

Our method's capability to generate decomposed characters in A-pose not only facilitates simpler rigging but also enables 3D editing through user-specified prompts and in-painting masks in the 2D domain. As illustrated in Fig. 7, our framework allows for effortless character customization, including outfit changes and hairstyle modifications. The process requires minimal user input: the original reference image or a front-view capture, a roughly sketched in-painting mask, and a text prompt. Utilizing existing in-painting models (HD-Painter [31] in this case), we achieve the desired 2D edits. Subsequently, these 2D modifications are translated into the 3D domain through our reconstruction and optimization pipeline. During the multi-layer refinement stage, the relevant mesh components are selectively replaced with the edited versions while preserving the unmodified parts. This bridges the gap between 2D image editing and 3D character customization and streamlines the 3D character modification process. We further provide animation comparisons in supplementary material (Sec. B).

5. Conclusion

In this paper, we introduce StdGEN, an innovative pipeline for generating semantic decomposed high-quality 3D characters from single images. A novel semantic-aware large reconstruction model is proposed to jointly reconstruct geometry, color and semantics, together with an efficient 2D diffusion model and iterative multi-layer refinement module to enable the high-quality generation from arbitrary-posed single image. StdGEN achieves superior performance over existing baselines in terms of geometry, texture and decomposability, and offers ready-to-use, semantic-decomposed, customizable 3D characters, demonstrating significant potential for various applications.

Acknowledgement. This work was partially supported by Beijing Science and Technology plan project (Z231100005923029), Beijing Natural Science Foundation (L247007), and the Natural Science Foundation of China (62461160309, 62332019).

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[2] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields, 2023. 3

[3] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models, 2023. 3

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 2

[5] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields, 2024. 3

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22246–22256, 2023. 2

[7] Ruikai Cui, Xibin Song, Weixuan Sun, Senbo Wang, Weizhe Liu, Shenzhou Chen, Taizhang Shang, Yang Li, Nick Barnes, Hongdong Li, and Pan Ji. Lam3d: Large image-point-cloud alignment model for 3d reconstruction from single image, 2024. 3

[8] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects, 2023. 2

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[10] Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai. Tela: Text to layer-wise 3d clothed human generation. *arXiv preprint arXiv:2404.16748*, 2024. 2, 3

[11] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023. 2

[12] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 6, 7

[13] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 2, 3

[14] 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. 2, 3, 7

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[16] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 4

[17] Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5734–5745, 2023. 2

[18] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars, 2023. 3

[19] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 3

[20] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors, 2022. 3

[21] Taeksoo Kim, Byungjun Kim, Shunsuke Saito, and Hanbyul Joo. Gala: Generating animatable layered assets from a single scan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1535–1545, 2024. 3

[22] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *The Twelfth International Conference on Learning Representations*. 3

[23] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 4, 6, 7

[24] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 3, 6

[25] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution

text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[26] 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*, pages 9298–9309, 2023. 3, 6, 7

[27] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *The Twelfth International Conference on Learning Representations*, 2024. 3, 6, 7

[28] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 3

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

[30] Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8744–8753, 2024. 3

[31] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: high-resolution and prompt-faithful text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2312.14091*, 2023. 8

[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[33] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations, 2022. 3

[34] 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. *arXiv preprint arXiv:2406.12459*, 2024. 2

[35] Bo Peng, Yunfan Tao, Haoyu Zhan, Yudong Guo, and Juyong Zhang. Pica: Physics-integrated clothed avatar. *arXiv preprint arXiv:2407.05324*, 2024. 2, 6

[36] Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACM Transactions on Graphics (TOG)*, 43(4): 1–13, 2024. 2, 3, 4, 6, 7, 8

[37] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 2

[38] Zhangyang Qi, Yunhan Yang, Mengchen Zhang, Long Xing, Xiaoyang Wu, Tong Wu, Dahua Lin, Xihui Liu, Jiaqi Wang, and Hengshuang Zhao. Tailor3d: Customized 3d assets editing and generation with dual-side images. *arXiv preprint arXiv:2407.06191*, 2024. 5

[39] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 6, 7

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[41] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 2

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[44] Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Control4d: Dynamic portrait editing by learning 4d gan from 2d diffusion-based editor. *arXiv preprint arXiv:2305.20082*, 2023. 2

[45] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[46] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Trans. Graph.*, 42(4):37–1, 2023. 3, 4, 5

[47] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*. 3

[48] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 2

[49] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 5459–5469, 2022. 2

[50] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3, 6, 7

[51] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439*, 2023. 2

[52] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation, 2022. 2

[53] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Xin Li, Wenping Wang, Rong Xie, and Li Song. Disentangled clothed avatar generation from text descriptions. *arXiv preprint arXiv:2312.05295*, 2023. 2, 3

[54] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 6, 7

[55] Yi Wang, Jian Ma, Ruizhi Shao, Qiao Feng, Yu-Kun Lai, and Kun Li. Humancoser: Layered 3d human generation via semantic-aware diffusion model. *arXiv preprint arXiv:2408.11357*, 2024. 2

[56] 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. 7

[57] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[58] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 3

[59] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 6, 7, 8

[60] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3, 4, 5, 6, 7, 8

[61] Han Yan, Yang Li, Zhennan Wu, Shenzhou Chen, Weixuan Sun, Taizhang Shang, Weizhe Liu, Tian Chen, Xiaqiang Dai, Chao Ma, et al. Frankenstein: Generating semantic-compositional 3d scenes in one tri-plane. *arXiv preprint arXiv:2403.16210*, 2024. 3

[62] Xu Yinghao, Shi Zifan, Yifan Wang, Chen Hansheng, Yang Ceyuan, Peng Sida, Shen Yujun, and Wetzstein Gordon. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation, 2024. 3

[63] Chubin Zhang, Hongliang Song, Yi Wei, Yu Chen, Jiwen Lu, and Yansong Tang. Geolrm: Geometry-aware large reconstruction model for high-quality 3d gaussian generation, 2024. 3

[64] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: a 3d generative model for animatable human avatars, 2022. 3

[65] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *European Conference on Computer Vision*, 2024. 3

[66] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 3

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7