Gen-3Diffusion: Realistic Image-to-3D Generation via 2D & 3D Diffusion Synergy

Yuxuan Xue[1,2] Xianghui Xie[1,2,3] Riccardo Marin[1,2] Gerard Pons-Moll[1,2,3]

[1]University of Tübingen [2]Tübingen AI Center [3]Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract—Creating realistic 3D objects and clothed avatars from a single RGB image is an attractive yet challenging problem. Due to its ill-posed nature, recent works leverage powerful prior from 2D diffusion models pretrained on large datasets. Although 2D diffusion models demonstrate strong generalization capability, they cannot guarantee the generated multi-view images are 3D consistent. In this paper, we propose **Gen-3Diffusion**: Realistic Image-to-3D Generation via 2D & 3D Diffusion Synergy. We leverage a pre-trained 2D diffusion model and a 3D diffusion model via our elegantly designed process that synchronizes two diffusion models at both training and sampling time. The synergy between the 2D and 3D diffusion models brings two major advantages: 1) **2D helps 3D in generalization**: the pretrained 2D model has strong generalization ability to unseen images, providing strong shape priors for the 3D diffusion model; 2) **3D helps 2D in multi-view consistency**: the 3D diffusion model enhances the 3D consistency of 2D multi-view sampling process, resulting in more accurate multi-view generation. We validate our idea through extensive experiments in image-based objects and clothed avatar generation tasks. Results show that our method generates realistic 3D avatars and objects with high-fidelity geometry and texture. Extensive ablations also validate our design choices and demonstrate the strong generalization ability to diverse clothing and compositional shapes. Our code and pretrained models will be publicly released on our project page.

Index Terms—3D Generation, Object Reconstruction, Human Reconstruction, Synchronized Diffusion Models

1 Introduction

Creating realistic 3D content is crucial for numerous applications, including AR/VR, as well as in the movie and gaming industries. Methods for creating a 3D model from a single RGB image are especially important to scale up 3D modelling and make it more consumer-friendly compared to traditional studio-based capture methods. However, this task presents substantial challenges due to the extensive variability in object shapes and appearances. These challenges are further intensified by the inherent ambiguities associated with monocular 2D views.

Moreover, beyond general object modeling, the generation of realistic clothed avatars presents a particularly demanding set of challenges. This complexity arises from the diversity of human body shapes and poses, which is compounded by a wide array of clothing, accessories, and occlusion by interacting objects. Such challenges are accentuated by the relative scarcity of large-scale 3D human datasets as compared to those available for objects, highlighting the critical need for advanced modeling techniques that can effectively navigate these complexities.

Recent image-to-3D approaches can be categorized into Direct-Reconstruction-based and Multi-View Diffusion-based methods. Direct-Reconstruction-based approaches directly predict a 3D representation that can be rendered from any viewpoint. Due to the explicit 3D representation, these methods produce an arbitrary number of consistent viewpoint renderings. For objects reconstruction, recent approaches such as LRM [1] and TriplaneGaussian [2] directly predict the NeRF or 3D Gaussian Splats from the input context view. However, these non-generative models directly regress the 3D representation in a deterministic manner, which easily leads to blurry unseen regions in Fig. 2. For clothed avatar reconstruction, recent popular approaches obtain the 3D model based on common template [3], [4], [5], [6] which utilize the SMPL [7] body model as the shape prior and perform the clothed avatar reconstruction. However, underlying SMPL body template highly limits the 3D representation of challenging human appearance, such as large dress, occlusion by interacting objects, etc. Examples can be found in Fig. 3. Furthermore, human reconstructors are trained on relative small-scale datasets due to the limited amount of high-quality 3D human data, which further restricts their ability to generalize to diverse shapes and textures. Last but not least, all above mentioned image-to-3D reconstruction works, regardless object-oriented or human-oriented, are typically deterministic which produce blurry textures and geometry in the occluded regions.

Multi-view diffusion-based methods [8], [9], [10] are proposed to synthesize desired novel views from single RGB image. These methods distill the inherent 3D structure presented in pretrained 2D diffusion models [11]. Typically, they fine-tune a large-scale 2D foundation model [11] on a large 3D dataset of objects [12], [13], [14], to generate novel views at given camera poses. Thanks to the pertaining on large-scale image datasets, Multi-view diffusion methods show strong generalization capability to unseen objects. However, since these models diffuse images purely in 2D without explicit 3D constraints or representation, the resulting multi-views often lack 3D consistency [15], [16]. The 3D inconsistent multi-view images further restrict downstream applications such as sparse-view 3D reconstruction [17].

E-mail:yuxuan.xue@uni-tuebingen.de

Fig. 1. Given a single image of a person or an object, our method **Gen-3Diffusion** creates realistic 3D objects or clothed avatars with high-fidelity geometry and texture. We use Gaussian Splatting to flexibly represent various shapes which can be extracted to high-quality textured meshes.

To address these challenges, we propose **Gen-3Diffusion**: Realistic Image-to-3D Generation via 2D & 3D Diffusion Synergy. We design our method based on two key insights: 1). 2D multi-view diffusion models (MVDs) provide strong shape priors that help 3D reconstruction; 2). Explicit 3D representation produces guaranteed 3D consistent multi-views that improve the accuracy of sampled 2D multi-view images. To leverage the benefits of both 2D MVD and explicit 3D representation, we propose a novel framework that synchronizes a 3D diffusion model with a 2D MVD model at each diffusion sampling step.

Specifically, we first introduce a novel 3D diffusion model that directly regresses 3D Gaussian Splatting (3D-GS [18]) from intermediately denoised multi-views images of 2D MVD. The predicted 3D-GS can be rendered into multi-view images with guaranteed 3D consistency. At every iteration, 2D MVD denoises multi-view images conditioned on input view, which are then reconstructed to 3D-GS by our 3D diffusion model. The 3D-GS are then re-rendered to multi-views to continue the diffusion sampling process. This 3D lifting during iterative sampling improves the 3D consistency of the generated 2D multi-view images while leveraging a large-scale foundation model trained on billions of images.

In summary, our contributions are:

- We propose a novel 3D-GS diffusion model for 3D reconstruction, which bridges large-scale priors from 2D multi-view diffusion models and the efficient explicit 3D-GS representation.
- A sophisticated joint diffusion process that incorporates reconstructed 3D-GS to improve the 3D consistency of 2D diffusion models by refining the reverse sampling trajectory.
- Our proposed formulation enables us to achieves superior performance and generalization capability than prior works, both in fileds of objects reconstruction ($Gen3D_{\text{object}}$) and clothed human reconstruction ($Gen3D_{\text{avatar}}$). Our code and pretrained models will be publicly released on our project page.

2 Related Works

2.1 Novel View Synthesis

Significant progress has been made in recent years in synthesizing images at target camera poses given multi-view observations. NeRF [20] and 3D Gaussian splitting (3D-GS) [18] are two popular representations for novel view synthesis. NeRF [20] uses neural networks to represent the continuous radiance fields and obtains new images via volumetric rendering. Despite impressive results, the training and rendering speed is slow and lots of efforts [21], [22] have been made to speed up NeRF. Alternatively, 3D-GS [18] represents the radiance with a discrete set of 3D Gaussians and renders them with rasterization which is highly efficient.

Optimizing NeRF or 3D-GS is time-consuming and requires dense multi-view images. Recently, Zero-1-to-3 [8] proposes a novel idea in fine-tuning pretrained image diffusion models [11] to generate the desired target views in a zero-shot manner. Considering the power of the pretrained StableDiffusion [11], Zero-1-to-3 has seen 5B 2D images and 10M 3D objects, demonstrating superior generalization ability in real-world images.

Despite the excellent generalization ability, zero-1-to-3 suffer from severe 3D inconsitency across different views while generating multiple different views. The reason is that the different views are sampled independently from each other. To address the multi-view inconsistency in diffusion sampling, multi-view diffusion models [9], [10], [23], [24] are proposed to generate multiple views simultaneously with information exchange across all sampled views using dense pixel-level attention in the latent space. In our observation, the dense multi-view attention improves the multi-view consistency, but also limits the ability of free novel view synthesis in Zero-1-to-3. Moreover, the generated multi-view images still have no guarantee of 3D consistency due to the lack of a common 3D representation during the diffusion sampling process.

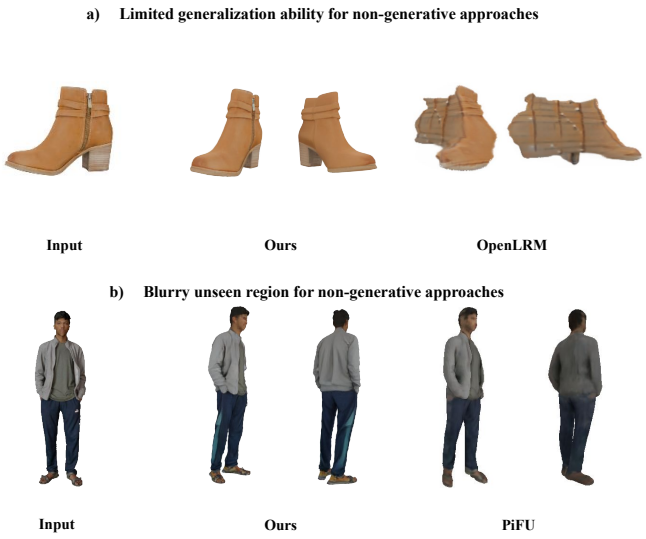


Fig. 2. **Motivation for generative 3D reconstruction design**. Unlike methods [1], [19] that deterministically regress 3D from single images, our Gen-3Diffusion learns conditional distribution and samples a plausible 3D-GS, resulting in high-fidelity and realistic unseen regions.

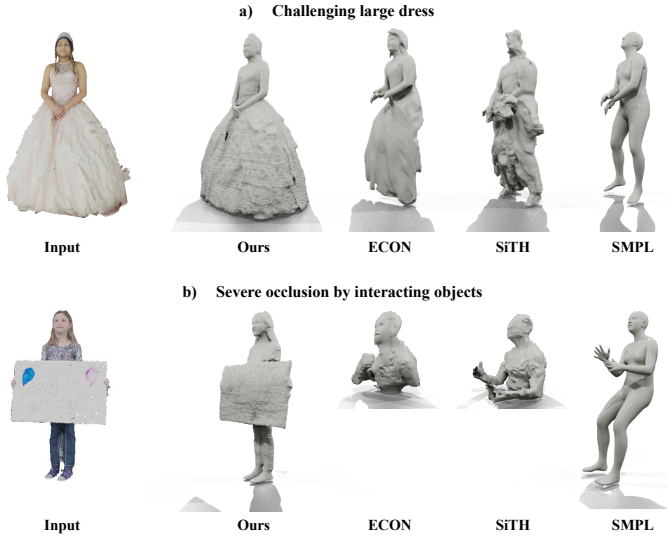


Fig. 3. **Motivation for template-free avatar reconstruction design**. Methods [3], [5] relying on SMPL [7] template suffer from inaccurate SMPL estimation and cannot represent challenging dresses or object interaction. Our Gen-3Diffusion is template-free and leverages shape prior from 2D diffusion models, can faithfully handle above challenges.

2.2 3D Objects from Image

Obtaining high-quality 3D objects from a single Image is an attractive but challenging task. Early object reconstruction works [25], [26], [27], [28], [29] focus mainly on geometry. Recently, with the emergence of differentiable rendering technologies, many works try to directly regress a 3D representation such as NeRF [1], [30] or 3D-GS [2], [31] from single RGB images. However, these methods are deterministic and do not learn the distribution of the underlying 3D scene, which can result in blurry rendering results at the inference time as in Fig. 2.

With the advance of 2D diffusion models [11] and efficient 3D representation [32], recent works can reconstruct 3D objects with detailed textures [1], [2], [9], [16], [17], [23], [33], [34], [35]. One popular paradigm is first using strong 2D models [8], [10], [36] to produce multi-view images and then train another model to reconstruct 3D from multi-view images [16], [17], [23], [37], [38]. In practice, their performance is limited by the accuracy of the multi-view images generated by 2D diffusion modes. Some works have tried to train another network that learns to correct the noisy multi-view images [16], [34], [37]. However, the network can be overfitted to error patterns from specific models, leading to limited generalization ability. Instead of correcting the multi-views in the last step output which is too late, we inject 3D consistency information early in the sampling stage, resulting in more accurate multi-views and 3D reconstruction.

2.3 Clothed Avatar from Image

Creating realistic human avatar from consumer grade sensors [39], [40], [41], [42], [43], [44] is essential for downstream tasks such as human behaviour understanding [45], [46], [47], [48], [49] and gaming application [50], [51], [52], [53], [54]. Researchers have explored avatar creation from monocular RGB [55], [56], Depth [43], [57] video or single image [4], [5], [19], [58], [59].

Avatar from single image is particularly interesting and existing methods can be roughly categorized as template-based [3], [4], [5], [6] and template-free [19], [58], [59], [60]. Template-free approaches [19], [58], [59], [60] directly predict a human occupancy field conditioned on a single image. This is flexible to represent diverse human clothing yet not robust to challenging poses due to the lack of shape prior. To leverage the shape prior information from human body models [7], [61], template-based approaches first estimate parametric body mesh from the image and then reconstruct the clothed avatar. Despite the impressive performance, these methods rely on the naked body model [7], [61] and they are affected by the inaccurate body mesh estimation which is common in extremely loose clothing or occlusion introduced by interacting objects as shown in Fig. 3. In this work, instead of using naked body models, we leverage the shape prior from pre-trained image diffusion models. This allows us to represent diverse clothed avatar shapes, including large dress and interacting objects. Furthermore, our method is not limited by the errors from monocular body mesh estimation methods.

3 Background

3.1 Denoising Diffusion Probabilistic Models

DDPM [62] is a generative model which learns a data distribution by iteratively adding (forward process) and removing (reverse process) the noise. Formally, the forward process iteratively adds noise to a sample $\mathbf{x}_0$ drawn from a distribution $p_{\text{data}}(\mathbf{x})$:

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\alpha_t, \bar{\alpha}_t$ schedules the amount of noise added at each step t [62]. To sample data from the learned distribution, the reverse process starts from $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it until $t = 0$:

$$\begin{aligned} \mathbf{x}_{t-1} &\sim \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \tilde{\beta}_{t-1}\mathbf{I}), \\ \text{where } \tilde{\beta}_{t-1} &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t) \end{aligned} \tag{2}$$

A network parametrized by θ is trained to estimate the posterior mean μ_θ at each step t. One can also model conditional distribution with DDPM by adding the condition to the network input [63], [64].

3.2 2D Multi-View Diffusion Models

Many recent works [8], [9], [10], [23], [24], [65] propose to leverage strong 2D image diffusion prior [11] pretrained on billions images [66] to generate multi-view images from a single image. Among them, ImageDream [10] demonstrated a superior generalization capability to unseen objects [17]. Given a single context view image $\mathbf{x}^{\text{c}}$, ImageDream generate 4 orthogonal target views $\mathbf{x}^{\text{tgt}}$ with a model ϵ_θ, which is trained to estimate the noise added at each step t. With the estimated noise ϵ_θ, one can compute the "clear" target views $\tilde{\mathbf{x}}_0^{\text{tgt}}$ with close-form solution in Eq. (1):

$$\tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t)). \tag{3}$$

This *one-step* estimation of $\tilde{\mathbf{x}}_0^{\text{tgt}}$ can be noisy and inaccurate, especially when t is large and $\mathbf{x}_t^{\text{tgt}}$ is extremely noisy and does not contain much information. Thus, the iterative sampling of $\mathbf{x}_t^{\text{tgt}}$ is required until $t = 0$. To sample next step $\mathbf{x}_{t-1}^{\text{tgt}}$, standard DDPM [62] computes the posterior mean μ_θ from current $\mathbf{x}_t^{\text{tgt}}$ and estimated $\tilde{\mathbf{x}}_0^{\text{tgt}}$ at step t with:

$$\begin{aligned} \mu_\theta(\mathbf{x}_t^{\text{tgt}}, t) &:= \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \tilde{\mathbf{x}}_0^{\text{tgt}}) \\ &= \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t^{\text{tgt}} + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\tilde{\mathbf{x}}_0^{\text{tgt}}, \\ &\text{where } \beta_t = 1-\alpha_t. \end{aligned} \tag{4}$$

Afterwards, $\mathbf{x}_{t-1}^{\text{tgt}}$ can be sampled from Gaussian distribution with mean μ_{t-1} and variance $\tilde{\beta}_{t-1}\mathbf{I}$ (Eq. (2)) and used as the input for the next iteration. The reverse sampling is repeated until $t = 0$ where 4 clear target views are generated.

Although multi-view diffusion models [9], [10], [24] generate multiple views together, the 3D consistency across these views is not guaranteed due to the lack of an explicit 3D representation. Thus, we propose a novel 3D consistent diffusion model, which ensures the multi-view consistency at each step of the reverse process by diffusing 2D images using reconstructed 3D Gaussian Splats [18].

3.3 3D Diffusion with Differentiable Rendering

Although DDPM [62] has emerged as a powerful class of generative models capable of capturing the distributions of complex signals, it can only model distributions for which training samples are directly accessible. Thus, directly training DDPM to learn the distribution of NeRF or 3D-GS requires pre-computing feature planes or Gaussians from 3D object scans, which is exorbitant. Recent works [67], [68], [69] propose to learn the distribution of 3D representation by diffusing the rendered images through differentiable rendering. In contrast to novel view diffusion models in Sec. 3.2, these works directly learn image-conditional 3D radiance field generation, instead of sampling from the distribution of novel views conditioned on a context view.

Given a single context view image $\mathbf{x}^{\text{c}}$, Diffusion-with-Forward (DwF) [68] generates Pixel-NeRF [30] from the noisy view $\hat{\mathbf{x}}_t^{\text{tgt}}$ and render to clear target view $\hat{\mathbf{x}}_0^{\text{tgt}}$:

$$\hat{\mathbf{x}}_0^{\text{tgt}} = \texttt{renderer}(\text{NeRF}_\phi(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, t)). \tag{5}$$

Similar to Eq. (3), the *one-step* estimation of $\hat{\mathbf{x}}_0^{\text{tgt}}$ can be noisy and inaccurate, especially when t is large. Thus, one can use standard DDPM to sample $\mathbf{x}_{t-1}^{\text{tgt}}$ using Eq. (2) and perform the iterative denoising.

Inspired by Diffusion-with-Forward [68], we learn the image-conditional 3D-GS generation in a diffusion-based framework. In this scenario, 3D-GS is efficient and thus more appropriate than NeRF for iterative sampling and rendering. Our $\texttt{renderer}(\cdot)$ is the differentiable rasteraizer implemented and accelerated by CUDA, which achieves around 2700 times faster rendering than volume-rendering-based $\texttt{renderer}(\cdot)$ in [68]. Moreover, our 3D-GS diffusion model can be enhanced by the 2D multi-view priors from 2D diffusion models in Sec. 3.2. We describe this model in more details in Sec. 4.1.

4 GEN-3DIFFUSION

Overview. Given a single RGB image, we aim to create a realistic 3D model consistent with the input. We adopt an image-conditioned 3D generation paradigm due to inherent ambiguities in the monocular view. We introduce a novel 3D Gaussian Splatting (3D-GS [18]) diffusion model that combines shape priors from 2D multi-view diffusion models with the explicit 3D-GS representation. This allows us to jointly train our 3D generative model and a 2D multi-view diffusion model end-to-end and improves the 3D consistency of 2D multi-view generation at inference time.

In this section, we first introduce our novel generative 3D-GS reconstruction model in Sec. 4.1. We then describe how we leverage the 3D reconstruction to generate 3D consistent multi-view results by refining the reverse sampling trajectory (Sec. 4.3) of 2D diffusion model. An overview of our 2D & 3D diffusion synergy can be found in Fig. 4.

4.1 3D-GS Diffusion Model

Given a context image $\mathbf{x}^c$, we use a conditional diffusion model to learn and sample from a plausible 3D distribution. Previous works demonstrated that 3D generation can be done implicitly via diffusing rendered images of a differentiable 3D represetation [67], [68], [69] such as NeRF [20], [30].

In this work, we propose a 3D-GS diffusion model g_ϕ, which is conditioned on input context image $\mathbf{x}^c$ to perform reconstruction of 3D Gaussian Splats $\mathcal{G}$. Diffusing directly in the space of $\mathcal{G}$ parameters requires pre-computing Gaussian Splats from scans, which is exorbitant. Instead, we diffuse

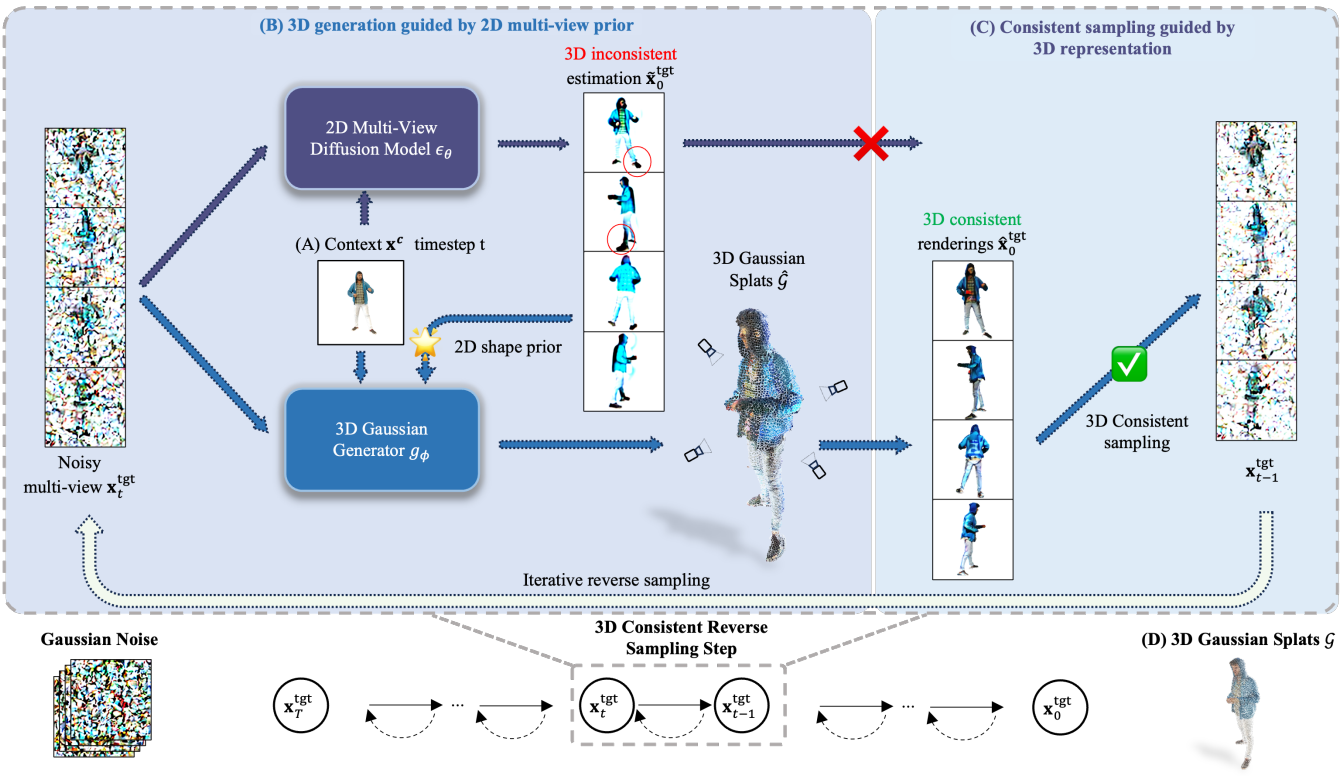


Fig. 4. **Method Overview.** Given a single RGB image (A), we sample a realistic 3D object represented as 3D Gaussian Splatting (D) from our learned distribution. At each reverse step, our 3D generation model g_ϕ leverages 2D multi-view diffusion prior from ϵ_θ which provides a strong shape prior but is not 3D consistent (B, Sec. 4.2). We then refine the 2D reverse sampling trajectory with generated 3D renderings that are guaranteed to be 3D consistent (C, Sec. 4.3). Our tight coupling ensures 3D consistency at each sampling step and obtains high-quality 3D Gaussian Splats.

the multi-view renderings of $\mathcal{G}$ using a differentiable rendering function $\texttt{renderer}(\cdot)$ to learn the conditional 3D distribution.

We denote $\mathbf{x}_0^{\text{tgt}}$ as the ground truth images at target views to be diffused and $\mathbf{x}_0^{\text{novel}}$ as the additional novel views for supervision. At training time, we uniformly sample a timestep $t \sim \mathcal{U}(0, T)$ and add noise to $\mathbf{x}_0^{\text{tgt}}$ using Eq. (1) to obtain noisy target views $\mathbf{x}_t^{\text{tgt}}$. Our generative model g_ϕ takes $\mathbf{x}_t^{\text{tgt}}$, diffusion timestep t, and the conditional image $\mathbf{x}^c$ as input, and estimates 3D Gaussians $\hat{\mathcal{G}}$:

$$\begin{aligned} \hat{\mathcal{G}} &= g_\phi(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^c), \\ \text{where } \mathbf{x}_t^{\text{tgt}} &= \sqrt{\bar{\alpha}_t}\mathbf{x}_0^{\text{tgt}} + \sqrt{1-\bar{\alpha}_t}\epsilon, \text{ and } \epsilon \sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \tag{6}$$

We adopt an asymmetric U-Net Transformer proposed by [17] for g_ϕ to directly predict 3D-GS parameters from per-pixel features of the last U-Net layer. The context image $\mathbf{x}^c$ is attended onto the noisy image $\mathbf{x}_t^{\text{tgt}}$ using dense pixel-wise attention. More specifically, the H $\times$ W $\times$ 14 feature map is reshaped in H $*$ W $\times$ 14, where a total number of H $*$ W 3D-GS are available, each has a center $\mathbf{o} \in \mathbb{R}^3$, a scaling factor $\mathbf{s} \in \mathbb{R}^3$, a rotation quaternion $\mathbf{q} \in \mathbb{R}^4$, an opacity value $\boldsymbol{\alpha} \in \mathbb{R}^1$, and a color feature $\mathbf{c} \in \mathbb{R}^3$. For more implementation details regarding the asymmetric U-Net Transformer, please refer to [17].

To supervise the generative model g_ϕ, we use a differentiable rendering function $\texttt{renderer}(\cdot) : \{\mathcal{G}, \pi^{\text{p}}\} \mapsto \mathbf{x}^{\text{p}}$ to render images at target views π^{tgt} and additional novel views π^{novel}. Denoting $\mathbf{x}_0 := \{\mathbf{x}_0^{\text{tgt}}, \mathbf{x}_0^{\text{novel}}\}$ as ground truth and $\hat{\mathbf{x}}_0 := \{\hat{\mathbf{x}}_0^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{novel}}\}$ as rendered images, we compute the loss on images and generated 3D-GS:

$$\begin{aligned} \mathcal{L}_{gs} &= \lambda_1 \cdot \mathcal{L}_{\text{MSE}}(\mathbf{x}_0, \hat{\mathbf{x}}_0) + \lambda_2 \cdot \mathcal{L}_{\text{Percep}}(\mathbf{x}_0, \hat{\mathbf{x}}_0) \\ &\quad + \lambda_3 \cdot \mathcal{L}_{\text{reg}}(g_\phi(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^c)), \\ \text{where } \hat{\mathbf{x}}_0 &:= \{\hat{\mathbf{x}}_0^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{novel}}\} \\ &= \texttt{renderer}(g_\phi(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^c), \{\pi^{\text{tgt}}, \pi^{\text{novel}}\}), \end{aligned} \tag{7}$$

here $\mathcal{L}_{\text{MSE}}$ denotes the Mean Square Error (MSE) and $\mathcal{L}_{\text{Percep}}$ is the perceptual loss based on VGG-19 [70]. We also apply $\mathcal{L}_{\text{reg}}$, a geometry regularizer [71], [72] to stabilize the generation of $\hat{\mathcal{G}}$.

With this, we can train a generative model that diffuses 3D-GS *implicitly* by diffusing 2D images $\mathbf{x}_t^{\text{tgt}}$. At inference time, we can generate 3D-GS given the input image by denoising 2D multi-views sampled from Gaussian distribution. We initialize $\mathbf{x}_T^{\text{tgt}}$ from $\mathcal{N}(0, \mathbf{I})$, and iteratively denoise the rendered images of predicted $\hat{\mathcal{G}}$ from our model g_ϕ. At each reverse step, our model g_ϕ estimates a clean state $\hat{\mathcal{G}}$ and render target images $\hat{\mathbf{x}}_0^{\text{tgt}}$. We then calculate target images $\mathbf{x}_{t-1}^{\text{tgt}}$ for the next step via Eq. (4) and repeat the process until $t = 0$, obtaining clear images $\hat{\mathbf{x}}_0^{\text{tgt}}$ and a clean 3D-GS $\hat{\mathcal{G}}$.

Our generative 3D-GS reconstruction model archives superior performance on in-distribution reconstruction yet generalizes poorly to unseen categories (Sec. 5.4 Fig. 10). Our key insight for better generalization is leveraging strong priors from pretrained 2D multi-view diffusion models for 3D-GS generation.

Algorithm 1 Joint 2D & 3D Diffusion Training

Input: Dataset of posed multi-view images $\mathbf{x}_0^{\text{tgt}}$, π^{tgt}, $\mathbf{x}_0^{\text{novel}}$, π^{novel}, a context image $\mathbf{x}^{\text{c}}$, text description y
Output: Optimized 2D multi-view diffusion model ϵ_θ and 3D-GS generative model g_ϕ

repeat
  $\{\mathbf{x}_0^{\text{tgt}}, \mathbf{x}_0^{\text{novel}}, \mathbf{x}^{\text{c}}, y\} \sim q(\{\mathbf{x}_0^{\text{tgt}}, \mathbf{x}_0^{\text{novel}}, \mathbf{x}^{\text{c}}, y\})$
  $t \sim \text{Uniform}(\{1, \dots, T\})$; $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}_t^{\text{tgt}} = \sqrt{\bar{\alpha}_t}\mathbf{x}_0^{\text{tgt}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$
  $\tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t))$
  $\hat{\mathcal{G}} = g_\phi\left(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^{\text{c}}, \tilde{\mathbf{x}}_0^{\text{tgt}}\right)$ // Enhance conditional 3D generation with 2D diffusion prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ from ϵ_θ
  $\{\hat{\mathbf{x}}_0^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{novel}}\} = \texttt{renderer}\left(\hat{\mathcal{G}}, \{\pi^{\text{tgt}}, \pi^{\text{novel}}\}\right)$
  Compute loss $\mathcal{L}_{total}$ (Eq. (9))
  Gradient step to update ϵ_θ, g_ϕ
until converged

Algorithm 2 3D Consistent Guided Sampling

Input: A context image $\mathbf{x}^c$ and text y; Converged 2D diffusion model ϵ_θ and 3D generative model g_ϕ
Output: 3D Gaussian Splats $\mathcal{G}$ of the 2D image $\mathbf{x}^c$

$\mathbf{x}_T^{\text{tgt}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  $\tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t))$
  $\hat{\mathcal{G}} = g_\phi\left(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^{\text{c}}, \tilde{\mathbf{x}}_0^{\text{tgt}}\right)$
  $\hat{\mathbf{x}}_0^{\text{tgt}} = \texttt{renderer}\left(\hat{\mathcal{G}}, \pi^{\text{tgt}}\right)$
  $\mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t^{\text{tgt}} + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0^{\text{tgt}}$ // Guide 2D sampling with 3D consistent multi-view renderings
  $\mathbf{x}_{t-1}^{\text{tgt}} \sim \mathcal{N}\left(\mathbf{x}_{t-1}^{\text{tgt}}; \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}\right), \tilde{\beta}_{t-1}\mathbf{I}\right)\right)$
end for
return $\mathcal{G} = g_\phi\left(\mathbf{x}_0^{\text{tgt}}, \tilde{\mathbf{x}}_0^{\text{tgt}}, \mathbf{x}^{\text{c}}, t = 0\right)$

4.2 3D Diffusion with 2D Multi-View Priors

Pretrained 2D multi-view diffusion models (MVD) [10], [24], [36] have seen billions of real images [66] and millions of 3D data [12], which provide strong prior information and can generalize to unseen objects [17], [34]. Here, we propose a simple yet elegant idea for incorporating this multi-view prior into our generative 3D-GS model g_ϕ. We can also leverage generated 3D-GS to guide 2D MVD sampling process which we discuss in Sec. 4.3.

Our key observation is that both 2D MVD and our proposed 3D-GS generative model are diffusion-based and share the same sampling state $\mathbf{x}_t^{\text{tgt}}$ at timestep t. Thus, they can be tightly *synchronized*. This enables us to couple and facilitate information exchange between 2D MVD $\boldsymbol{\epsilon}_\theta$ and 3D-GS generative model g_ϕ at the same diffusion timestep t. To inject the 2D diffusion priors into 3D generation, we first compute *one-step* estimation of $\tilde{\mathbf{x}}_0^{\text{tgt}}$ (Eq. (3)) using 2D MVD ϵ_θ, and condition our 3D-GS generative mode g_ϕ additionally on it. Formally, our 3D-GS generative model enhanced with 2D multi-view diffusion priors is written as:

$$\begin{aligned} \hat{\mathcal{G}} &= g_\phi(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^c, \tilde{\mathbf{x}}_0^{\text{tgt}}), \\ \text{where } \tilde{\mathbf{x}}_0^{\text{tgt}} &= \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, t)). \end{aligned} \tag{8}$$

The visualization of $\tilde{\mathbf{x}}_0^{\text{tgt}}$ along the whole sampling trajectory in Fig. 5 shows that the pretrained 2D diffusion model $\boldsymbol{\epsilon}_\theta$ can already provide useful multi-view shape prior even in large timestep $t = 1000$. This is further validated in our experiments where the additional 2D diffusion prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ leads to better 3D reconstruction (Tab. 7) as well as more robust generalization to general objects (Fig. 10). By utilizing the timewise iterative manner of 2D and 3D diffusion models, we can not only leverage 2D priors for 3D-GS generation but also train both models jointly end to end, which we discuss in Sec. 4.3.

4.3 Synergy between 2D & 3D Diffusion

Joint Diffusion Training We adopt pretrained ImageDream [10] as our 2D multi-view diffusion model ϵ_θ and jointly train it with our 3D-GS generative model g_ϕ. We observe that our joint training is important for coherent 3D generation, as opposed to prior works that frozen pretrained 2D multi-view models [17], [33]. We summarize our training algorithm in Algorithm 1. We combine the loss of 2D diffusion and our 3D-GS generation loss $\mathcal{L}_{gs}$(Eq. (7)):

$$\mathcal{L}_{total} = \mathcal{L}_{\text{MSE}}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\theta) + \mathcal{L}_{gs} \tag{9}$$

Once trained, one can sample a plausible 3D-GS $\mathcal{G}$ conditioned on the input image from the learned 3D distributions. However, we observe that the multi-view diffusion model $\boldsymbol{\epsilon}_\theta$ can still output inconsistent multi-views along the sampling trajectory (see Fig. 4). On the other hand, our 3D generator produces explicit 3D-GS which can be rendered as 3D consistent multi-views. Our second key idea is to use the 3D consistent renderings to guide 2D sampling process for more 3D consistent multi-view generation. We discuss this next.

3D Consistent Guided Sampling With the shared and *synchronized* sampling state $\mathbf{x}_t^{\text{tgt}}$ of 2D multi-view diffusion model $\boldsymbol{\epsilon}_\theta$ and 3D-GS reconstruction model g_ϕ, we couple both models at arbitrary t during training. Similarly, they are also connected by both using estimated clean multi-views $\mathbf{x}_0^{\text{tgt}}$ at sampling time. To leverage the full potential of both models, we carefully design a joint sampling process that utilizes the reconstructed 3D-GS $\hat{\mathcal{G}}$ at each timestep t to guide 2D multi-view sampling, which is summarized in Algorithm 2.

We observe that the key difference between the clean multi-views estimated $\mathbf{x}_0^{\text{tgt}}$ from 2D diffusion model and our 3D-GS generation lies in 3D consistency: 2D MVD computes multi-view $\tilde{\mathbf{x}}_0^{\text{tgt}}$ from 2D network prediction which can be 3D inconsistent while our $\hat{\mathbf{x}}_0^{\text{tgt}}$ are rendered from explicit 3D-GS representation which are guaranteed to be 3D consistent. Our idea is to guide the 2D multi-view reverse sampling process with our 3D consistent renderings $\hat{\mathbf{x}}_0^{\text{tgt}}$ such that the 2D sampling trajectory is more 3D consistent. Specifically, we leverage 3D consistent multi-view renderings $\hat{\mathbf{x}}_0^{\text{tgt}}$ to

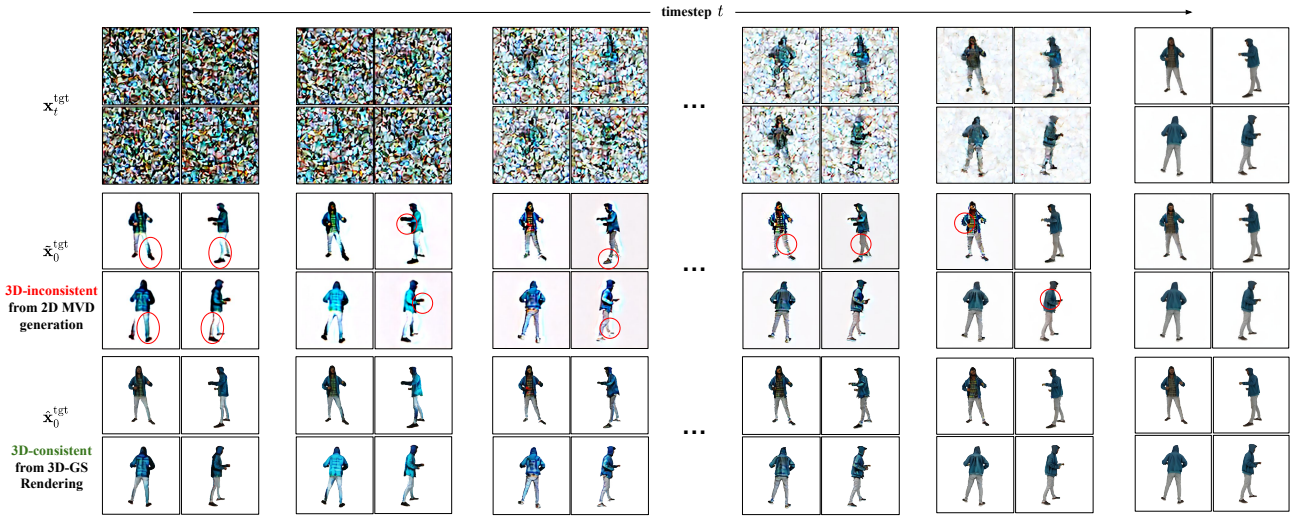


Fig. 5. Visualization of DDPM reverse sampling trajectory. At each individual step, estimated $\tilde{\mathbf{x}}_0^{\text{tgt}}$ can be 3D inconsistency across different views, while the rendering $\hat{\mathbf{x}}_0^{\text{tgt}}$ are 3D consistent and can refine the inconsistency along trajectory (Eq. (10)).

refine the posterior mean $\mu_\theta(\mathbf{x}_t^{\text{tgt}}, t)$ at each reverse step:

$$\begin{aligned} &\text{Original: } \mu_\theta(\mathbf{x}_t^{\text{tgt}}, t) := \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \tilde{\mathbf{x}}_0^{\text{tgt}}) \\ \rightarrow \quad &\text{Ours: } \mu_\theta(\mathbf{x}_t^{\text{tgt}}, t) := \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}), \\ &\text{where } \hat{\mathbf{x}}_0^{\text{tgt}} = \texttt{renderer}(\hat{\mathcal{G}}, \pi^{\text{tgt}}), \\ \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}) &= \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t^{\text{tgt}} + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0^{\text{tgt}} \end{aligned} \tag{10}$$

With this refinement, we guarantee the 3D consistency at each reverse step t and avoid 3D inconsistency accumulation in original multi-view sampling [10]. In Fig. 5, we visualize the evolution of originally generated multi-views $\tilde{\mathbf{x}}_0^{\text{tgt}}$ and multi-views rendering $\hat{\mathbf{x}}_0^{\text{tgt}}$ from generated 3D-GS $\hat{\mathcal{G}}$ along the whole reverse sampling process. It intuitively shows how effective the sampling trajectory refinement is. We perform extensive ablation in Sec. 5.4 showing the importance of the consistent refinement for sampling trajectory.

5 Experiments

In this section, we demonstrate the effectiveness of our method for two image-based 3D reconstruction tasks: general object reconstruction (denoted as *Gen3D*$_{\text{object}}$, Sec. 5.2) and clothed human avatar reconstruction (denoted as *Gen3D*$_{\text{avatar}}$, Sec. 5.3). We also ablate the influence of 2D and 3D diffusion models to our full pipeline in Sec. 5.4.

5.1 Experimental Setup

5.1.1 Datasets

General Object. We use a filtered high-quality Objaverse [12] subset introduced in LGM [17] which consists of around 80K objects to train our *Gen3D*$_{\text{object}}$ model. We evaluate our model on the Google Scanned Object dataset (GSO) [73] according to the same evaluation protocol specified in EscherNet [74].

Clothed Avatar. We train our *Gen3D*$_{\text{avatar}}$ model on a combined 3D human dataset [75], [76], [77], [78], [79], [80], [81], [82], compromising $\sim$ 6000 high quality scans. We evaluate our *Gen3D*$_{\text{avatar}}$ on around 450 subjects from three different datasets: sizer [83], iiit [84], and cape [85].

5.1.2 Implementation Details

Network Architecture. Following [17], our 3D-GS generative model g_ϕ consists of 6 down blocks, 1 middle block, and 5 up blocks, with the input image at 256×256 and output Gaussian feature map at 128×128. For each iteration, we start from 4 Gaussian noisy images $\mathbf{x}_t^{\text{tgt}}$ and concatenating their corresponding 2D prior images $\tilde{\mathbf{x}}_0^{\text{tgt}}$ channel-wise, and our g_ϕ generates in total $128 \times 128 \times 4 = 65536$ number of 3D-GS. For implementation details regarding the U-Net model, please refer to [11], [17].

Training. We trained both our *Gen3D*$_{\text{object}}$ and *Gen3D*$_{\text{avatar}}$ models on 8 NVIDIA A100 GPUs for approximately 5 days. Each GPU was configured with a batch size 2 and gradient accumulations of 16 steps to achieve an effective batch size of 256. Each batch involved sampling 4 orthogonal images with zero elevation angle as target views $\mathbf{x}_0^{\text{tgt}}$, and 12 additional images as novel views $\mathbf{x}_0^{\text{novel}}$ to supervise the 3D generative model Eq. (9). The hyperparameters for training Eq. (9) were set as follows: $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, and $\lambda_3 = 100.0$. During training, we employed the standard DDPM scheduler [62] to construct noisy target images $\mathbf{x}_t^{\text{tgt}}$. The maximum diffusion step T is set to 1000. The AdamW [86] optimizer is adopted with the learning rate of 5×10^{-4} , weight decay of 0.05, and betas of (0.9, 0.95). The learning rate is cosine annealed to 0 during the training. We clip the gradient with a maximum norm of 1.

Inference. Our whole pipeline, including joint sampling of both 2D & 3D diffusion models, takes only about 11.7 GB of GPU memory and 22.6 seconds for inference on NVIDIA A100, which is friendly for deployment. For the adopted pretrained multi-view diffusion model, we use a guidance scale of 5 following [10]. At inference time, we use DDIM scheduler [87] to perform faster reverse sampling. The total reverse sampling steps are set to 50 in all experiments. We use GOF [72] and TSDF [88] to extract the textured mesh from the generated 3D Gaussian Splats.

5.1.3 Evaluation Metrics

We evaluate the 3D reconstruction quality in terms of appearance and geometry. For appearance quality, we compute metrics on directly generated images (novel view synthesis methods) or renderings (direct 3D reconstruction methods) at 32 novel camera views with uniform azimuth and zero elevation angle. The metrics for appearance reported include multi-scale Structure Similarity (SSIM) [89], Learned Perceptual Image Patch Similarity (LPIPS) [90], and Peak Signal to Noise Ratio (PSNR) between predicted and ground-truth views. Moreover, we report the Fréchet inception distance (FID) [91] between synthesized views and ground truth renderings, which reflects the quality and realism of the unseen regions.

For geometry quality, we compute Chamfer Distance (CD), Point-to-Surface distance (P2S), F-score [92] (w/ threshold of $0.01m$), and Normal Consistency (NC) between the extracted geometry and the groundtruth scan. We normalize the extracted geometry into $[-1, 1]$ and perform iterative closest point (ICP) to match the global pose between extracted and groundtruth geometry to ensure alignment, same as [3], [74]. In all experiments, we re-evaluate the baseline models by using their officially open-sourced checkpoints on the same set of reference views for a fair comparison.

5.2 3D Object from Image

We evaluate *Gen3D*$_{\text{object}}$ for novel view synthesis and geometry reconstruction on the GSO dataset [73]. We compare our method against 2D novel view diffusion models Zero-1-to-3 [8], Zero-1-to-3-XL [8], EscherNet [74], and SV3D [93], as well as methods that directly reconstruct 3D models such as LRM [1], TriplaneGaussian [2] and LGM [17]. Since the code for LRM is not publicly available, we adopt the implementation and pretrained model of OpenLRM [94] and TripoSR [33] and compare with them. Notably, many other 2D multi-view diffusion models [9], [10], [23] prioritize 3D generation rather than view synthesis. This limits their methods to generate fixed target views rather than arbitrary free-view synthesis, making them not directly comparable.

We report the quantitative evaluation results in Tab. 1 and show some comparisons in Fig. 6. Novel View Diffusion models [8], [74], [93] achieve good appearance metrics yet they cannot directly produce a 3D representation from the images. Direct reconstruction approaches [2], [33], [94] predicts 3D directly from images. However, the geometry could be over-smooth (TriplaneGaussian [2]) or the texture is not realistic (LGM [17]). Our *Gen3D*$_{\text{object}}$ diffuses 2D images and 3D-GS jointly, which results in better and more 3D-consistent view synthesis and better 3D reconstruction. Please refer to supplementary video for a more comprehensive comparison.

5.3 Realistic Avatar from Image

We evaluate *Gen3D*$_{\text{avatar}}$ in novel view synthesis and the geometry on the Sizer dataset [83], IIIT [84], and CAPE dataset [85], [96], [97]. We compare our approach against prior methods for image-to-avatar reconstruction, including pure geometry-based [4], [5], [95] and textured geometry-based [3], [6], [19] human reconstruction methods. To further assess performance, we also fine-tuned the state-of-the-art object reconstruction method LGM [17] and its underlying multi-view diffusion model [10] on our training data, denoted as LGM$_{\text{ft}}$.

We report quantitative evaluation in Tab. 2 and Tab. 3 and show qualitative comparison for both appearance and geometry via extracted mesh in Fig. 7. It can be seen that template-based approaches heavily rely on accurate SMPL estimations hence they easily fail when the estimations are off. This is common when the person is wearing large dress, has a different shape as the adult SMPL body shape or is interacting with object/accessories. In contrast, our method is template-free hence can flexibly represent all possible body and clothing shapes, leading to more coherent appearance and geometry reconstruction.

To further evaluate the reconstruction quality, we conduct a user study to compare the reconstruction of different methods. We render 20 subjects with texture to compare ours against SiTH and SIFU and another 20 subjects with only geometry to compare ours against ICON and ECON. The subjects are randomly sampled from evaluation dataset of Sizer [83], IIIT [84], CAPE [85]. We release the user study to 70 people from different technical backgrounds. Overall, our results are preffered by 80.3% of users. It clearly shows that our *Gen3D*$_{\text{avatar}}$ significantly outperforms baselines in both geometry and appearance. Please see Fig. 8 for visualization of the user study results.

Our method is a diffusion-based feed-forward approach without any SMPL estimation or test-time optimization process as is typical in template-based methods [3], [4], [5], [6]. This allows us to obtain 3D reconstruction at higher inference speed. We report the runtime comparison (on Nvidia A100) in Tab. 4. Our method is much faster than baseline human reconstruction methods.

5.4 Ablation Studies

In this section, we elaborate our ablation studies which validate our design choices. Note that here we focus on our human model *Gen3D*$_{\text{avatar}}$ and report results on the human datasets as well as generalization to o.o.d unseen objects.

3D Diffusion helps 2D Diffusion. One of our key ideas is leveraging our explicit 3D model to refine the 2D multi-view reverse sampling trajectory, ensuring 3D consistency in 2D Multi-View Diffusion (MVD) generation (see Sec. 4.3 and Eq. (10)). To evaluate this, we compare the multi-view images (4 orthogonal views) generated by pretrained MVD [10], fine-tuned MVD on our human data (MVD$_{\text{ft}}$) and MVD with our 3D consistent sampling (ours), as shown in Tab. 8. The results demonstrate that our proposed method effectively enhances the quality of generated multi-view images by leveraging the explicit 3D model to refine sampling trajectory.

Additionally. we analyze the 3D reconstruction results with the multi-view images generated by these models in Fig. 9. MVD and MVD$_{\text{ft}}$ produce inconsistent multi-view images, which typically lead to floating Gaussian and hence blurry boundaries. In contrast, our method can generate more consistent multi-views, result in better 3D Gaussians

Fig. 6. **Novel view synthesis visualisation of 3D objects from images.** Our *Gen3D*$_{\text{object}}$ is able to directly generate 3D-GS and render to arbitrary desired novel multi-views, which are more detailed and faithful w.r.t. the context image, and more 3D-consistent compared to prior works.

TABLE 1
Comparing object reconstruction methods on GSO dataset [73]. Our method achieves a better appearance and higher-quality 3D geometry.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CD$_{(cm)}$ ↓	P2S$_{(cm)}$ ↓	NC ↑	F-score ↑
Zero-1-to-3 [8]	20.158	0.876	0.109	67.87	–	–	–	
Zero-1-to-3-XL [8]	20.324	0.884	0.107	65.14	–	–	–	–
EscherNet [74]	20.503	0.895	0.107	65.75	–	–	–	–
SV3D [93]	20.975	0.900	0.105	64.72	–	–	–	–
OpenLRM [94]	18.972	0.880	0.133	143.29	9.17	9.37	0.663	0.112
TripoSR [33]	19.820	0.898	0.110	73.26	6.23	6.49	0.734	0.178
TriplaneGaussian [2]	18.067	0.893	0.132	149.92	10.83	14.18	0.601	0.081
LGM	19.089	0.885	0.122	64.16	9.88	12.32	0.579	0.146
Gen3D$_{\text{object}}$	**22.881**	**0.917**	**0.078**	**54.12**	**4.12**	**4.00**	**0.734**	**0.293**

Splats and sharper renderings. We further quantitatively evaluate the impact of our proposed sampling trajectory refinement on final 3D reconstruction in Tab. 6. We compare the reconstruction results of methods with and without our trajectory refinement while using the same 2D MVD and 3D reconstruction models with same setting in Tab. 3 and Tab. 2. It can be clearly seen that our trajectory refinement improves the quality of 3D reconstruction.

2D Diffusion helps 3D Diffusion. Another key idea of our work is the use of multi-view priors $\tilde{\mathbf{x}}_0^{\text{tgt}}$ from 2D diffusion model pretrained on massive data [11], [12], [66] to enhance our 3D generative model. This additional prior

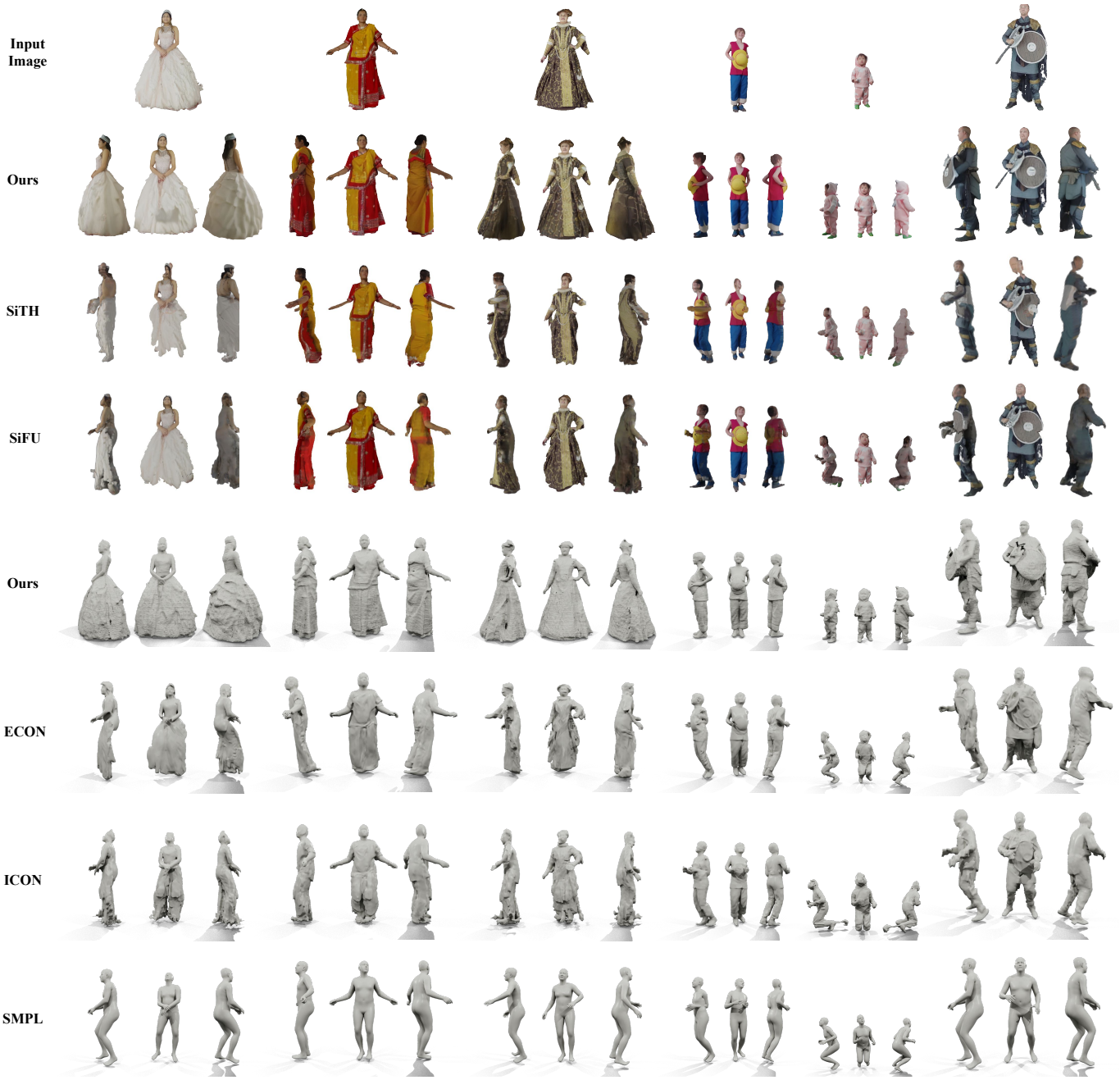


Fig. 7. **Appearance and geometry comparison of human avatar creation methods.** We show novel view renderings of the textured avatar and extracted mesh in row 2-4 and row 5-7 respectively. Prior methods produce blurry textures (SiTH [3], SiFU [6]) or oversmoothed surface (ECON [5], ICON [4]) on unseen backside regions. The reliance on SMPL estimation is susceptible to errors (shown in row 8) and makes them difficult to model loose clothing or diverse human shapes. In contrast, our method adopts template-free 3D-GS representation, and leverages strong prior from 2D MVD models, allowing us to faithfully reconstruct high-fidelity geometry and appearance from single RGB image.

information is pivotal for ensuring accurate reconstruction of both in-distribution human dataset and generalizing to out-of-distribution objects.

We evaluate the performance of our 3D model g_ϕ by comparing generation results with and without the 2D diffusion prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ (refer to Eq. (8) and Eq. (6)). For avatars reconstruction, our powerful 3D reconstruction model can already achieve state-of-the-art performance. Moreover, our $\textit{Gen3D}_{\text{avatar}}$ full model with multi-view prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ generates avatars with higher quality as demonstrated in Tab. 7. We further evaluate it on the GSO [73] dataset which consists of unseen general objects to our $\textit{Gen3D}_{\text{avatar}}$ model. The improvements are even more pronounced in this setting, highlighting the challenges of generating coherent 3D structures from a single 2D image, particularly with unseen objects. These ablation studies effectively proves that the 2D multi-view diffusion prior enhances generalization capability.

6 Overview

6.1 Limitations

Limited by low resolution (256×256) of our adopted pre-trained 2D diffusion model [10], our model cannot recover

TABLE 2
Geometry evaluation for clothed avatars reconstruction on Sizer, IIIT, and CAPE dataset. Our method produces better 3D geometry in all datasets.

	Sizer Dataset [83]				IIIT Dataset [84]				CAPE Dataset [84]			
Method	$CD_{(cm)}\downarrow$	$P2S_{(cm)}\downarrow$	F-score ↑	NC ↑	$CD_{(cm)}\downarrow$	$P2S_{(cm)}\downarrow$	F-score ↑	NC ↑	$CD_{(cm)}\downarrow$	$P2S_{(cm)}\downarrow$	F-score ↑	NC ↑
SMPL [7]	3.94	4.02	0.237	0.743	4.67	4.33	0.204	0.728	5.04	4.91	0.213	0.743
PiFU [19]	2.35	2.31	0.410	0.782	2.70	2.64	0.337	0.764	3.40	3.27	0.314	0.791
FoF [95]	5.37	5.26	0.204	0.676	5.34	5.29	0.188	0.691	5.65	5.52	0.146	0.689
ICON [4]	3.01	3.20	0.285	0.771	4.55	4.53	0.202	0.716	4.28	4.28	0.238	0.762
ECON [5]	2.83	3.04	0.329	0.781	3.86	3.84	0.253	0.744	3.96	4.14	0.286	0.775
SiTH [3]	3.38	3.45	0.285	0.753	4.90	4.83	0.208	0.716	3.76	3.95	0.279	0.785
SiFU [6]	2.69	2.81	0.324	0.778	4.25	4.18	0.216	0.725	3.73	3.71	0.270	0.779
LGM_{ft} [17]	2.80	3.27	0.306	0.556	3.76	4.31	0.245	0.567	3.96	4.23	0.258	0.557
$Gen3D_{avatar}$	**1.06**	**1.05**	**0.627**	**0.794**	**1.44**	**1.39**	**0.531**	**0.781**	**1.89**	**1.84**	**0.491**	**0.801**

TABLE 3
Appearance evaluation for clothed avatars reconstruction on Sizer, IIIT, and CAPE dataset. Our method produces overall better appearance.

	Sizer Dataset [83]				IIIT Dataset [84]				CAPE Dataset [84]			
Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
PiFU [19]	19.22	0.913	0.068	33.50	22.40	**0.905**	0.083	22.41	22.03	0.910	0.082	38.79
SiTH [3]	18.90	0.912	0.063	21.87	19.53	0.901	0.078	19.90	22.20	0.908	0.082	28.46
SiFU [6]	18.01	0.899	0.072	36.64	**22.65**	0.899	0.087	46.76	**22.23**	0.906	0.085	43.63
LGM_{ft} [17]	20.57	0.902	0.077	16.75	20.65	0.879	0.100	15.54	20.46	0.898	0.089	20.33
$Gen3D_{avatar}$	**21.28**	**0.928**	**0.047**	**10.01**	22.13	**0.905**	**0.066**	**9.69**	21.46	**0.912**	**0.064**	**16.40**

TABLE 4
Runtime performance comparison. Our method is faster than template-based human reconstruction methods.

	Time (s)	VRAM (GB)
ICON [4]	60.5	6.3
ECON [5]	45.3	5.9
SiFU [6]	48.9	12.0
SiTH [3]	106.2	22.0
$Gen3D_{avatar}$	22.6	11.7

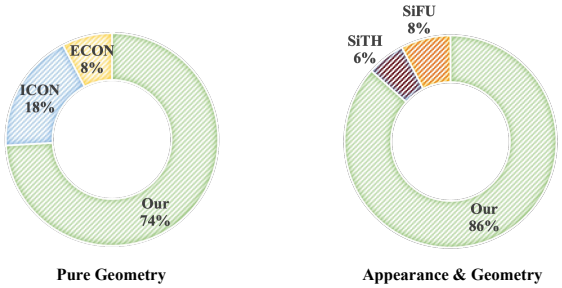


Fig. 8. **Runtime performance & user preference comparison.** Left: Inference time and GPU consumption of SoTA avatar reconstruction approaches. Right: User study statistics of avatar reconstruction on pure geometry or appearance & geometry comparison. Our method is preferred by most people in both geometry and appearance.

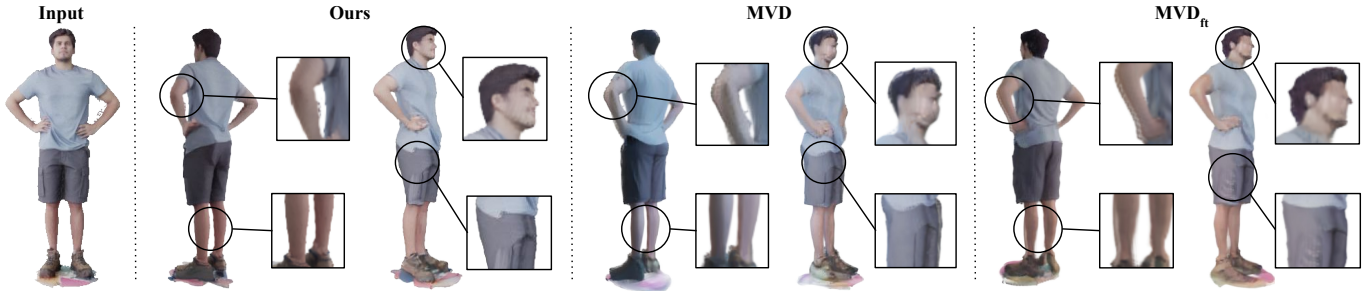


Fig. 9. **3D reconstruction conditioned on different multi-view priors.** Without our 3D-consistent sampling, the 2D diffusion model cannot generate 3D consistent multi-views (MVD, MVD_{ft}), leading to artifacts like floating 3D Gaussians splats. Our method obtains more consistent multi-views hence better 3D-GS and rendering.

TABLE 5
Ablation of 2D multi-view priors in o.o.d. generalization.

Method	LPIPS↓	SSIM↑	PSNR↑	FID↓
Ours w/o $\tilde{\mathbf{x}}_0^{tgt}$	0.189	0.721	14.45	107.14
Ours	**0.194**	**0.778**	**16.12**	**83.89**

Fig. 10. **Ablation of 2D multi-view priors** for object reconstruction. The model is trained for human reconstruction. It can be seen in the left table that our 2D MVD model improves the generalization ability to unseen objects, leading to more plausible object shape as shown on the right image.

TABLE 6
Ablating the influence of 3D consistent guided sampling for 3D-GS generation. Our proposed sampling strategy improves the 3D reconstruction quality by enhancing multi-view consistency of 2D diffusion models.

Method	$CD_{(cm)}\downarrow$	F-score↑	NC ↑	LPIPS↓	SSIM↑	PSNR↑
Our w/o Traj. Ref.	1.57	0.498	0.794	0.064	0.908	21.09
Ours	**1.35**	**0.550**	**0.798**	**0.060**	**0.918**	**21.49**

TABLE 7
Ablating the influence of 2D multi-view priors $\tilde{\mathbf{x}}_0^{\text{tgt}}$. The strong prior from 2D diffusion models enhance the 3D reconstruction quality.

Method	$CD_{(cm)}\downarrow$	F-score↑	NC ↑	LPIPS↓	SSIM↑	PSNR↑
Ours w/o $\tilde{\mathbf{x}}_0^{\text{tgt}}$	1.75	0.498	0.795	0.068	0.912	20.98
Ours	**1.35**	**0.550**	**0.798**	**0.060**	**0.918**	**21.49**

TABLE 8
Ablating the influence of 3D consistent guided sampling for 2D multi-view images generation. Our proposed sampling strategy improves the multi-view image quality from 2D diffusion models.

Method	PSNR ↑	LPIPS ↓	SSIM ↑
MVD	22.32	0.078	0.911
MVD_{ft}	24.14	0.061	0.926
Ours	**24.69**	**0.048**	**0.934**

fine details such as text on the objects. A potential solution is to use a recent powerful high-resolution multi-view diffusion models [65], [98], which provide strong shape priors in higher resolution.

6.2 Future Works

Our **Gen-3Diffusion** is a general framework for image-to-3D reconstruction, which shows the superior reconstruction ability on isolated objects and human subjects. Extending the current framework to scene-level reconstruction yields more difficulties such as different camera poses and z-buffering. Moreover, reconstructing 4D Gaussian Splatting from single RGB videos is attractive but challenging due to more monocular ambiguities. We leave these to future works.

6.3 Conclusion

In this paper, we introduce **Gen-3Diffusion**, a 3D consistent diffusion model for creating realistic 3D objects or clothed avatars from single RGB images. Our key ideas are two folds: 1) Leveraging strong multi-view priors from pre-trained 2D diffusion models to generate 3D Gaussian Splats, and 2) Using the reconstructed explicit 3D Gaussian Splats to refine the sampling trajectory of the 2D diffusion model which enhances 3D consistency. We carefully designed a diffusion process that synergistically combines the strengths of both 2D and 3D models. We compare our image-to-3D model $Gen3D_{\text{object}}$ with 8 state-of-the-art methods and our image-to-avatar model $Gen3D_{\text{avatar}}$ with 6 popular works, show that our approach outperforms them in both appearance and geometry. We also extensively ablate our method which proves the effectiveness of our proposed ideas. Our code and pretrained models will be publicly available on our project page.

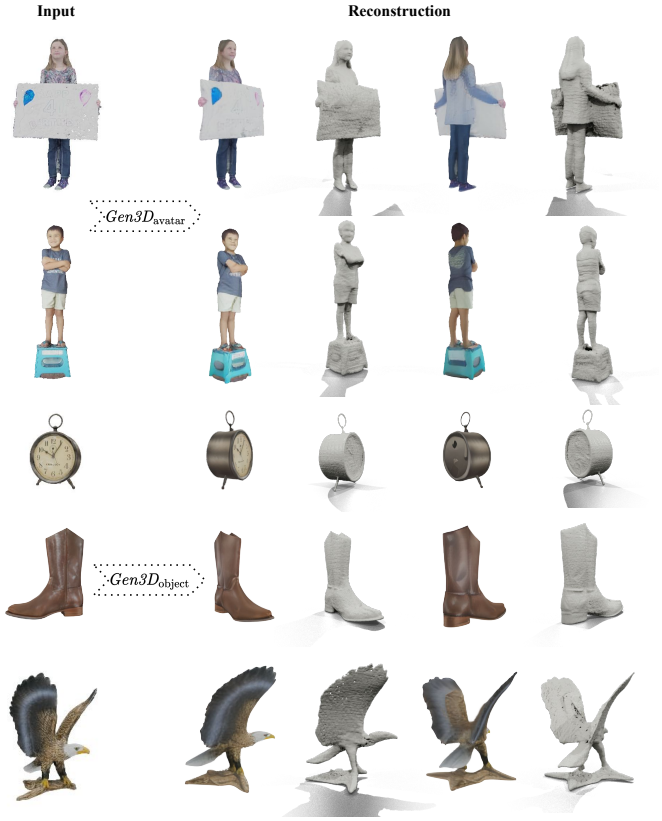


Fig. 11. Gallery of 3D reconstruction from single RGB images using our $Gen3D_{\text{avatar}}$ and $Gen3D_{\text{object}}$.

ACKNOWLEDGMENTS

This work is made possible by funding from the Carl Zeiss Foundation. This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (EmmyNoether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Y.Xue. R. Marin has been supported by the innovation program under Marie Skłodowska-Curie grant agreement No 101109330. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

REFERENCES

[1] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan, "LRM: large reconstruction model for single image to 3d," *CoRR*, vol. abs/2311.04400, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.04400

[2] Z. Zou, Z. Yu, Y. Guo, Y. Li, D. Liang, Y. Cao, and S. Zhang, "Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers," *CoRR*, vol. abs/2312.09147, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.09147

[3] H. Ho, J. Song, and O. Hilliges, "Sith: Single-view textured human reconstruction with image-conditioned diffusion," *CoRR*, vol. abs/2311.15855, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.15855

[4] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black, "ICON: implicit clothed humans obtained from normals," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 13 286–13 296. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01294

[5] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, "ECON: explicit clothed humans optimized via normal integration," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 512–523. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.00057

[6] Z. Zhang, Z. Yang, and Y. Yang, "SIFU: side-view conditioned implicit function for real-world usable clothed human reconstruction," *CoRR*, vol. abs/2312.06704, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.06704

[7] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: a skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818013

[8] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 9264–9275. [Online]. Available: https://doi.org/10.1109/ICCV51070.2023.00853

[9] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su, "Zero123++: a single image to consistent multi-view diffusion base model," *CoRR*, vol. abs/2310.15110, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.15110

[10] P. Wang and Y. Shi, "Imagedream: Image-prompt multi-view diffusion for 3d generation," *CoRR*, vol. abs/2312.02201, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.02201

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 10 674–10 685. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01042

[12] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 13 142–13 153. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.01263

[13] T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, J. Wang, C. Qian, D. Lin, and Z. Liu, "Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 803–814. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.00084

[14] X. Yu, M. Xu, Y. Zhang, H. Liu, C. Ye, Y. Wu, Z. Yan, C. Zhu, Z. Xiong, T. Liang, G. Chen, S. Cui, and X. Han, "Mvimgnet: A large-scale dataset of multi-view images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 9150–9161. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.00883

[15] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov, and B. Ghanem, "Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors," *CoRR*, vol. abs/2306.17843, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.17843

[16] M. Liu, C. Xu, H. Jin, L. Chen, M. V. T, Z. Xu, and H. Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/4683beb6bab325650db13afd05d1a14a-Abstract-Conference.html

[17] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu, "LGM: large multi-view gaussian model for high-resolution 3d content creation," *CoRR*, vol. abs/2402.05054, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.05054

[18] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139:1–139:14, 2023. [Online]. Available: https://doi.org/10.1145/3592433

[19] S. Saito, Z. Huang, R. Natsume, S. Morishima, H. Li, and A. Kanazawa, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 2304–2314. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00239

[20] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2022. [Online]. Available: https://doi.org/10.1145/3503250

[21] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530127

[22] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision (ECCV)*, 2022.

[23] X. Long, Y. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S. Zhang, M. Habermann, C. Theobalt, and W. Wang, "Wonder3d: Single image to 3d using cross-domain diffusion," *CoRR*, vol. abs/2310.15008, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.15008

[24] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," *CoRR*, vol. abs/2309.03453, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.03453

[25] K. V. Alwala, A. Gupta, and S. Tulsiani, "Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3773–3782.

[26] A. Thai, S. Stojanov, V. Upadhya, and J. M. Rehg, "3d reconstruction of novel object shapes from single images," 2020.

[27] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. T. Freeman, and J. B. Tenenbaum, "Learning 3D Shape Priors for Shape Completion and Reconstruction," in *European Conference on Computer Vision (ECCV)*, 2018.

[28] Y. Xian, J. Chibane, B. L. Bhatnagar, B. Schiele, Z. Akata, and G. Pons-Moll, "Any-shot gin: Generalizing implicit networks for reconstructing novel classes," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022.

[29] X. Zhang, Z. Zhang, C. Zhang, J. B. Tenenbaum, W. T. Freeman, and J. Wu, "Learning to Reconstruct Shapes From Unseen Classes," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[30] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 4578–4587. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Yu_pixelNeRF_Neural_Radiance_Fields_From_One_or_Few_Images_CVPR_2021_paper.html

[31] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, "Pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 19 457–19 467. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.01840

[32] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, T. Karras,

and G. Wetzstein, "Efficient geometry-aware 3d generative adversarial networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 16 102–16 112. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01565

[33] D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y. Cao, "Triposr: Fast 3d object reconstruction from a single image," *CoRR*, vol. abs/2403.02151, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.02151

[34] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," *arXiv preprint arXiv:2404.07191*, 2024.

[35] Y. Xu, H. Tan, F. Luan, S. Bi, P. Wang, J. Li, Z. Shi, K. Sunkavalli, G. Wetzstein, Z. Xu, and K. Zhang, "DMV3D: denoising multi-view diffusion using 3d large reconstruction model," *CoRR*, vol. abs/2311.09217, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.09217

[36] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," *CoRR*, vol. abs/2308.16512, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2308.16512

[37] M. Liu, R. Shi, L. Chen, Z. Zhang, C. Xu, X. Wei, H. Chen, C. Zeng, J. Gu, and H. Su, "One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion," *CoRR*, vol. abs/2311.07885, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.07885

[38] Y. Xu, Z. Shi, W. Yifan, S. Peng, C. Yang, Y. Shen, and W. Gordon, "Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation," *arxiv: 2403.14621*, 2024.

[39] B. Kabadayi, W. Zielonka, B. L. Bhatnagar, G. Pons-Moll, and J. Thies, "Gan-avatar: Controllable personalized gan-based human head avatar," in *International Conference on 3D Vision (3DV)*, March 2024.

[40] K. Youwang, T.-H. Oh, and G. Pons-Moll, "Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[41] G. Tiwari, N. Sarafianos, T. Tung, and G. Pons-Moll, "Neural-gif: Neural generalized implicit functions for animating people in clothing," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 11 688–11 698. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01150

[42] Y. Xue, H. Li, S. Leutenegger, and J. Stueckler, "Event-based non-rigid reconstruction from contours," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022, p. 78. [Online]. Available: https://bmvc2022.mpi-inf.mpg.de/78/

[43] Y. Xue, B. L. Bhatnagar, R. Marin, N. Sarafianos, Y. Xu, G. Pons-Moll, and T. Tung, "NSF: neural surface fields for human modeling from monocular depth," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 15 004–15 014. [Online]. Available: https://doi.org/10.1109/ICCV51070.2023.01382

[44] Y. Xue, H. Li, S. Leutenegger, and J. Stückler, "Event-based non-rigid reconstruction of low-rank parametrized deformations from contours," in *International Journal of Computer Vision (IJCV)*. Springer Science and Business Media LLC, Feb. 2024. [Online]. Available: http://dx.doi.org/10.1007/s11263-024-02011-z

[45] B. L. Bhatnagar, X. Xie, I. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Behave: Dataset and method for tracking human object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.

[46] I. A. Petrov, R. Marin, J. Chibane, and G. Pons-Moll, "Object pop-up: Can we infer 3d objects and their poses from human interactions alone?" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 4726–4736. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.00458

[47] X. Xie, B. L. Bhatnagar, and G. Pons-Moll, "CHORE: contact, human and object reconstruction from a single RGB image," *CoRR*, vol. abs/2204.02445, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.02445

[48] X. Xie, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll, "Template free reconstruction of human-object interaction with procedural interaction generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[49] X. Xie, B. L. Bhatnagar, and G. Pons-Moll, "Visibility aware human-object interaction tracking from single RGB camera," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 4757–4768. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.00461

[50] Y. Li, H.-y. Chen, E. Larionov, N. Sarafianos, W. Matusik, and T. Stuyck, "DiffAvatar: Simulation-ready garment optimization with differentiable simulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[51] Z. Liu, Y. Feng, Y. Xiu, W. Liu, L. Paull, M. J. Black, and B. Schölkopf, "Ghost on the shell: An expressive representation of general 3d shapes," *arXiv preprint arXiv:2310.15168*, 2023.

[52] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human poseitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors," *CoRR*, vol. abs/2103.17265, 2021. [Online]. Available: https://arxiv.org/abs/2103.17265

[53] X. Zhang, B. L. Bhatnagar, S. Starke, V. Guzov, and G. Pons-Moll, "COUCH: towards controllable human-chair interactions," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13665. Springer, 2022, pp. 518–535. [Online]. Available: https://doi.org/10.1007/978-3-031-20065-6_30

[54] X. Zhang, B. L. Bhatnagar, S. Starke, I. Petrov, V. Guzov, H. Dhamo, E. Pérez-Pellitero, and G. Pons-Moll, "FORCE: dataset and method for intuitive physics guided human-object interaction," *CoRR*, vol. abs/2403.11237, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.11237

[55] T. Jiang, X. Chen, J. Song, and O. Hilliges, "Instantavatar: Learning avatars from monocular video in 60 seconds," *arXiv*, 2022.

[56] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, "HumanNeRF: Free-viewpoint rendering of moving people from monocular video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 210–16 220.

[57] Z. Dong, C. Guo, J. Song, X. Chen, A. Geiger, and O. Hilliges, "Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence," *arXiv*, 2022.

[58] S. Saito, T. Simon, J. M. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 81–90. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Saito_PIFuHD_Multi-Level_Pixel-Aligned_Implicit_Function_for_High-Resolution_3D_Human_Digitization_CVPR_2020_paper.html

[59] A. Sengupta, T. Alldieck, N. Kolotouros, E. Corona, A. Zanfir, and C. Sminchisescu, "DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans," Mar. 2024, arXiv:2404.00485 [cs]. [Online]. Available: http://arxiv.org/abs/2404.00485

[60] X. Yang, Y. Luo, Y. Xiu, W. Wang, H. Xu, and Z. Fan, "D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 9088–9098. [Online]. Available: https://ieeexplore.ieee.org/document/10377954/

[61] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," *CoRR*, vol. abs/1904.05866, 2019. [Online]. Available: http://arxiv.org/abs/1904.05866

[62] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[63] P. Dhariwal and A. Q. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang,

and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=AAWuCvzaVt

[64] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. [Online]. Available: https://openreview.net/forum?id=qw8AKxfYbI

[65] S. Tang, J. Chen, D. Wang, C. Tang, F. Zhang, Y. Fan, V. Chandra, Y. Furukawa, and R. Ranjan, "Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction," *CoRR*, vol. abs/2402.12712, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.12712

[66] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: an open large-scale dataset for training next generation image-text models," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

[67] T. Anciukevicius, Z. Xu, M. Fisher, P. Henderson, H. Bilen, N. J. Mitra, and P. Guerrero, "Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 12 608–12 618. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.01213

[68] A. Tewari, T. Yin, G. Cazenavette, S. Rezchikov, J. Tenenbaum, F. Durand, B. Freeman, and V. Sitzmann, "Diffusion with forward models: Solving stochastic inverse problems without direct supervision," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/28e4ee96c94e31b2d040b4521d2b299e-Abstract-Conference.html

[69] A. Karnewar, A. Vedaldi, D. Novotny, and N. Mitra, "Holodiffusion: Training a 3D diffusion model using 2D images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[70] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[71] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, "2d gaussian splatting for geometrically accurate radiance fields," *CoRR*, vol. abs/2403.17888, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.17888

[72] Z. Yu, T. Sattler, and A. Geiger, "Gaussian opacity fields: Efficient high-quality compact surface reconstruction in unbounded scenes," *arXiv:2404.10772*, 2024.

[73] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*. IEEE, 2022, pp. 2553–2560. [Online]. Available: https://doi.org/10.1109/ICRA46639.2022.9811809

[74] X. Kong, S. Liu, X. Lyu, M. Taher, X. Qi, and A. J. Davison, "Eschernet: A generative model for scalable view synthesis," *CoRR*, vol. abs/2402.03908, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.03908

[75] "Axyz," Nov 2023. [Online]. Available: https://secure.axyz-design.com

[76] "Treedy," Nov 2023. [Online]. Available: https://treedys.com/

[77] "Twindom," Nov 2023. [Online]. Available: https://web.twindom.com/

[78] "Renderpeople," Nov 2023. [Online]. Available: https://renderpeople.com/

[79] S.-H. Han, M.-G. Park, J. H. Yoon, J.-M. Kang, Y.-J. Park, and H.-G. Jeon, "High-fidelity 3d human digitization from single 2k resolution images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2023)*, June 2023.

[80] H. Hsuan-I, X. Lixin, S. Jie, and H. Otmar, "Learning locally editable virtual humans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[81] Z. Su, T. Yu, Y. Wang, and Y. Liu, "Deepcloth: Neural garment representation for shape and style editing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1581–1593, 2023.

[82] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu, "Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021.

[83] G. Tiwari, B. L. Bhatnagar, T. Tung, and G. Pons-Moll, "SIZER: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12348. Springer, 2020, pp. 1–18. [Online]. Available: https://doi.org/10.1007/978-3-030-58580-8_1

[84] S. S. Jinka, A. Srivastava, C. Pokhariya, A. Sharma, and P. J. Narayanan, "SHARP: shape-aware reconstruction of people in loose clothing," *Int. J. Comput. Vis.*, vol. 131, no. 4, pp. 918–937, 2023. [Online]. Available: https://doi.org/10.1007/s11263-022-01736-z

[85] Q. Ma, S. Tang, S. Pujades, G. Pons-Moll, A. Ranjan, and M. J. Black, "Dressing 3d humans using a conditional mesh-vae-gan," *CoRR*, vol. abs/1907.13615, 2019. [Online]. Available: http://arxiv.org/abs/1907.13615

[86] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[87] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[88] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *CVPR*, 2017.

[89] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402 Vol.2.

[90] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 586–595. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html

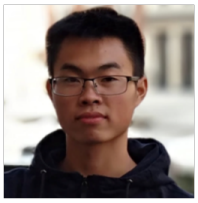
[91] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6626–6637. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html

[92] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, "What do single-view 3d reconstruction networks learn?" in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 3405–3414. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Tatarchenko_What_Do_Single-View_3D_Reconstruction_Networks_Learn_CVPR_2019_paper.html

[93] V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani, "SV3D: Novel Multi-view

Synthesis and 3D Generation from a Single Image using Latent Video Diffusion," Mar. 2024.
[94] Z. He and T. Wang, "Openlrm: Open-source large reconstruction models," https://github.com/3DTopia/OpenLRM, 2023.
[95] Q. Feng, Y. Liu, Y.-K. Lai, ingyu Yang, and K. Li, "Fof: Learning fourier occupancy field for monocular real-time human reconstruction," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
[96] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3d scan sequences," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5484–5493. [Online]. Available: https://doi.org/10.1109/CVPR.2017.582
[97] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "Clothcap: seamless 4d clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 73:1–73:15, 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073711
[98] R. Gao*, A. Holynski*, P. Henzler, A. Brussee, R. Martin-Brualla, P. P. Srinivasan, J. T. Barron, and B. Poole*, "Cat3d: Create anything in 3d with multi-view diffusion models," *arXiv*, 2024.

Yuxuan Xue is currently pursuing Ph.D. degree in the Real Virtual Human (RVH) group at the Univerisity Tübingen, supervised by Prof. Dr. Gerard Pons-Moll. Before that, he received B.Sc. (2016) and M.Sc. (2022) from Technical University of Munich (TUM). His research interests lie on perceiving human from real world and modelling into metaverse. He has published at top conferences and journal in machine learning and vision (ICCV, NeurIPS, ICLR, IJCV). His work got Best Student Paper Award at BMVC 2022.



Xianghui Xie is currently pursuing Ph.D. degree in the Real Virtual Human (RVH) group at the Univerisity Tübingen and Max Planck Institute for Informatics, supervised by Prof. Dr. Gerard Pons-Moll. Before that, he obtained B.Sc. from KU Leuven and M.Sc from Saarland university . His research interests lie on modelling interaction between human and objects. He has published at top machine learning and vision conferences (ECCV, CVPR, NeurIPS).

Riccardo Marin received the PhD degree in Computer Science from the University of Verona, Italy, in 2021. After that, he was a post-doc researcher and Adjunct Professor at the Sapienza University of Rome as part of the GLADIA lab. In 2022, he joined the University of Tuebingen as a post-doc researcher in the Real Virtual Humans group, funded by a Humboldt Foundation Research Fellowship and a Marie Skłodowska-Curie Post-Doctoral Fellowship. He is currently a post-doc researcher at the Technical University of Munich (TUM) as part of the Computer Vision Group. His research interests include 3D Shape Analysis, Matching and Registration, Geometric Deep Learning, and Virtual Humans.

Gerard Pons-Moll is Professor at the University of Tuebingen, head of the Emmy Noether independent research group "Real Virtual Humans", senior researcher at the Max Planck for Informatics (MPII) in Saarbrucken, Germany. His research lies at the intersection of computer vision, computer graphics and machine learning – with special focus on analyzing people in videos, and creating virtual human models by "looking" at real ones. His work has received several awards including the prestigious Emmy Noether Grant (2018), a Google Faculty Research Award (2019), a Facebook Reality Labs Faculty Award (2018), and the German Pattern Recognition Award (2019). His work got Best Papers Awards and nomminations at CVPR'20, CVPR'21, ECCV'22. He serves regularly as area chair for the top conferences in vision and graphics (CVPR, ICCV, ECCV, Siggraph).