Towards High-Fidelity 3D Portrait Generation with Rich Details by Cross-View Prior-Aware Diffusion

Haoran Wei[1*], **Wencheng Han**[1*], **Xingping Dong**[2], **Jianbing Shen**[1†]

[1]SKL-IOTSC, CIS, University of Macau
[2]School of Computer Science, Wuhan University
{hr.wei1998, wenchenghan, xingping.dong}@gmail.com

Abstract

Recent diffusion-based Single-image 3D portrait generation methods typically employ 2D diffusion models to provide multi-view knowledge, which is then distilled into 3D representations. However, these methods usually struggle to produce high-fidelity 3D models, frequently yielding excessively blurred textures. We attribute this issue to the insufficient consideration of cross-view consistency during the diffusion process, resulting in significant disparities between different views and ultimately leading to blurred 3D representations. In this paper, we address this issue by comprehensively exploiting multi-view priors in both the conditioning and diffusion procedures to produce consistent, detail-rich portraits. From the conditioning standpoint, we propose a Hybrid Priors Diffusion model, which explicitly and implicitly incorporates multi-view priors as conditions to enhance the status consistency of the generated multi-view portraits. From the diffusion perspective, considering the significant impact of the diffusion noise distribution on detailed texture generation, we propose a Multi-View Noise Resampling Strategy integrated within the optimization process leveraging cross-view priors to enhance representation consistency. Extensive experiments show that our method produces 3D portraits with accurate geometry and rich details from a single image.

Code — https://haoran-wei.github.io/Portrait-Diffusion

Introduction

The generation of realistic 3D portraits from a single image (Deng, Wang, and Wang 2024a; Xiang et al. 2020; Doukas, Zafeiriou, and Sharmanska 2021; Wu et al. 2023; Ma et al. 2023) has become an important focus in computer vision and graphics, with broad applications in virtual reality (Jiang et al. 2024; Li et al. 2024b; Ye et al. 2024).

Recent advancements forego GAN-based methods (Yin et al. 2023a; An et al. 2023), which rely on costly large-scale training (Deng, Wang, and Wang 2024a,b), and instead adopt text-to-image diffusion priors for stronger generalization and higher-quality generation of diverse full-head portraits (Yang, Chen, and Liao 2023; Xie et al. 2023; Corneanu, Gadde, and Martinez 2024). These approaches incorporate additional priors, such as reference image latents (Xie et al. 2024; Zhang et al. 2024), ID features (Shao et al. 2024; Hao et al. 2024), and view embeddings (Shao et al. 2024), to enhance the consistency between new perspectives and the primary viewpoint. Subsequently, they commonly employ Score Distilling Sampling (SDS) loss (Poole et al. 2022b) to distill these 2D priors into 3D representations.

However, in single-image 3D portrait generation, these methods still face challenges: generated portraits often appear over-smoothed and fail to capture detailed textures like hair strands, as illustrated in fig. 1, limiting their practical applications. We attribute this issue to their insufficient consideration of cross-view consistency during the diffusion process, which leads to significant disparities across different views and ultimately produces blurred 3D outputs under SDS optimization. Although these methods seek to enhance consistency by integrating additional priors, they rely solely on diffusion attentions to implicitly convey these priors. This results in a lack of explicit constraints, leading to status inconsistent across different viewpoints.

Additionally, the diffusion procedure is inherently stochastic: even under identical conditions, randomly sampled noise can yield divergent representations. By using view-independent procedures with purely random diffusion noise, these methods overlook the how this randomness undermines representation consistency. Consequently, both inconsistencies in status and representation jointly result in over-smoothed 3D models under the SDS optimization, which enforces 3D consistency and continuity in sacrifice of texture details.

To address these issues, we propose fully exploiting cross-view priors in both the conditioning and diffusion procedures to enhance multi-view consistency, thus yielding detail-rich 3D portraits, as showcased in fig. 1.

From a conditioning perspective, we propose Hybrid Priors Diffusion Model (HPDM). Our approach seeks to transfer and utilize cross-view prior information in both explicit and implicit ways to control the novel view generation. In an explicit manner, we begin by employing geometric priors to map pixels from the current view to the next, providing an explicit reference to dominate the generation process. In an implicit manner, given that this reference encompasses only a limited overlapping region and contains artifacts introduced through perspective transformations, we

[*]These authors contributed equally.
[†]Correspondence author: *Jianbin Shen*.

Figure 1: Our proposed **Portrait Diffusion** framework can generate high-quality detail-rich 3D portraits from a single reference portrait image. In comparison to SOTA methods **Era3D** (Li et al. 2024a) and **Portrait3D** (Wu et al. 2024), our approach achieves clearer and more detailed textures.

further propose to utilize the robust modeling capabilities of attention mechanisms to mitigate these deficiencies. These mechanisms capture finer texture and geometry priors and implicitly transfer these priors into the control conditions, ensuring a more comprehensive and precise guidance for the portrait status of novel viewpoint.

From a diffusion procedure perspective, our goal is to manage randomness in adjacent viewpoints so that they can share detailed, consistent representations. We introduce a Multi-View Noise Resampling Strategy (MV-NRS) integrated into the SDS loss, which manages each view's noise distribution by passing cross-view priors. MV-NRS consists of two main components: first, a shared anchor noise initialization that leverages geometric priors to establish a preliminary representation; and second, an anchor noise optimization phase, where we resample and update the anchor noise based on denoising gradient consistency prior to progressively align the representations during the SDS optimization. To summarize, our main contributions are as follows:

- We developed a Portrait Diffusion pipeline consisting of Portrait Geometry Restoration and Multi-view Diffusion Refinement modules to generate rich-detail 3D portraits.
- We designed a Hybrid Priors Diffusion Model that emphasizes both explicit and implicit integration of multi-view priors to impose conditions, aiming to enhance the consistency of multi-view status.
- We introduced a Multi-View Noise Resampling Strategy integrated within the SDS loss to manage randomness across different views through the transfer of cross-view priors to achieve fine-grained consistent representations.
- Through extensive experiments, we show that our proposed pipeline successfully achieves high-fidelity 3D full portrait generation with rich details.

Related Work

One-shot 3D Generation. Recent 3D GANs (Yin et al. 2023b; Dundar et al. 2023; Reddy, Elezi, and Deng 2024) greatly advance one-shot 3D object generation in quality and efficiency, but limited training data restrict their priors (Deng et al. 2022; Gao et al. 2022; Chan et al. 2022; Xiang et al. 2022). Leveraging 2D diffusion priors (Mirzaei et al. 2023; Wang et al. 2024), recent works achieve better texture fidelity and multi-view coherence. DreamFusion (Poole et al. 2022a) distills guidance from diffusion models, while DreamCraft3D (Sun et al. 2024) enhances view consistency via Bootstrapped Score Distillation. Make-It-3D (Tang et al. 2023) and Make-It-Vivid (Tang et al. 2024) refine texture realism through perceptual supervision and UV-space diffusion.

One-shot 3D Portrait Generation In 3D portrait synthesis, Yin et al. (An et al. 2023) generates 360° portraits using a two-phase registration strategy and tri-mesh neural volumetric representations. Leveraging diffusion priors, recent methods enable zero-shot full-head generation with high fidelity. For instance, Portrait3D (Wu et al. 2024) employs 3DPortraitGAN to produce 360° canonical portraits, alleviating "grid-like" artifacts through a pyramidal tri-grid representation and refining details with fractional distillation sampling. DiffusionAvatars (Kirschstein, Giebenhain, and Nießner 2024) combines a diffusion-based renderer with neural head models and uses cross-attention to ensure consistent expressions across angles. Another ID-Sculpt system (Hao et al. 2024) focuses on identity preservation through geometry initialization, sculpting, and texture generation, applying ID-aware techniques at each stage.

Method

Problem Formulation

Existing diffusion-based methods for generating 3D objects predominantly utilize Score Distillation Sampling (SDS) loss (Poole et al. 2022b) to distill 2D diffusion priors into 3D representations. This process can be formulated as follows:

$$\Phi^* = \arg\min_{\Phi}(\mathcal{L}_{\text{SDS}}(\Phi;\theta) + \mathcal{L}_{\text{ref}}(\Phi; I^{\text{ref}})) \quad (1)$$

where Φ denotes the parameters of a 3D model, $\mathcal{L}_{\text{SDS}}(\Phi;\theta)$ represents the SDS loss using a diffusion model paramter-

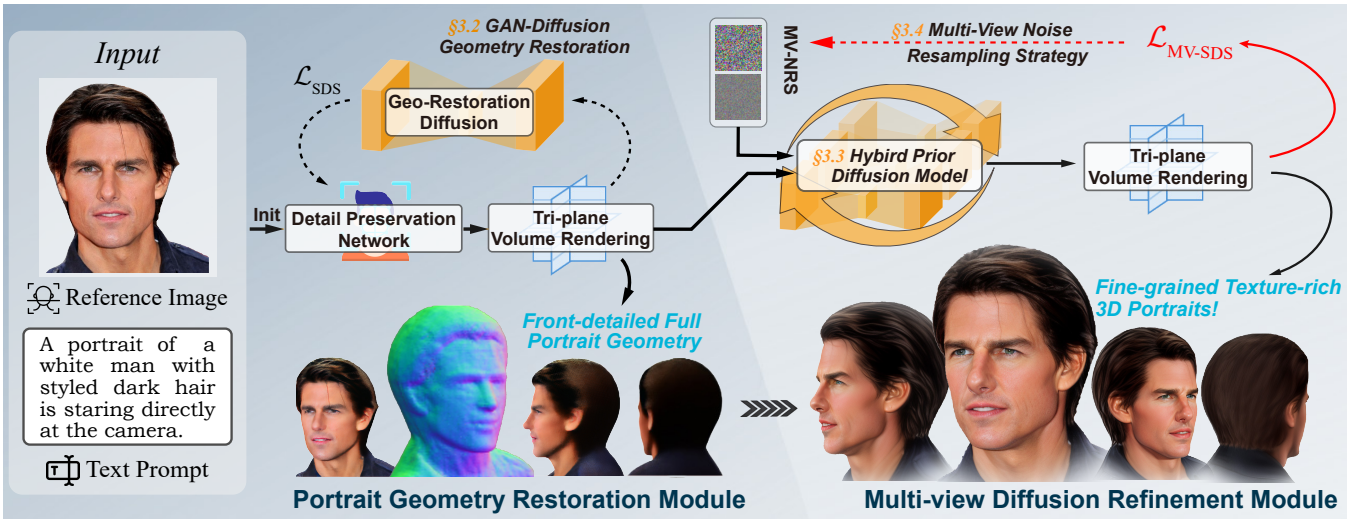

Figure 2: **The Portrait Diffusion Framework.** This framework comprises two modules. *GAN-Diffusion Geometry Restoration Module* incorporate GAN and diffusion priors to derive full-head portrait geometry from given images. *Multi-view Diffusion Texture Refinement* transforms over-smoothed textures into detailed representations.

ized by θ, and $\mathcal{L}_{\text{ref}}(\Phi; I^{\text{ref}})$ is a loss computed from reference image I^{ref}. The SDS loss can be formulated as:

$$\nabla_{\Phi}\mathcal{L}_{\text{SDS}} = \mathbf{E}_{t,v,\epsilon}\Big[w_t\left(\epsilon_\theta(z_{t,v}, t, c) - \epsilon\right)\cdot\nabla_{\Phi}\mathcal{R}_{\Phi}(v)\Big] \\ z_{t,v} = \sqrt{\alpha_t}z_v(\Phi) + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon\sim\mathcal{N}(0,\mathbf{I}), \tag{2}$$

where $z_{t,v}$ is a noisy latent representation obtained by combining the image latents $z_v(\Phi)$, which is rendered from viewpoint v by Φ, with random noise ϵ; $\epsilon_\theta(z_{t,v}, t, c)$ is a diffusion UNet model that predicts the noise component at each time step t, conditioned on c. w_t and α_t are weights, and $\mathcal{R}$ is rendering function.

From (2), the SDS loss aggregates the denoising gradients from all v to the 3D model parameters Φ. When the denosing distributions across viewpoints are inconsistent, the SDS loss will produce over-smoothed representations to minimize the overall loss by averaging conflicting gradients, sacrificing the details of each perspective. The denoising function $\boldsymbol{\epsilon_\theta}$ is influenced by both the conditions $\boldsymbol{c}$ and the distribution of noise $\boldsymbol{\epsilon}$ from each viewpoint, making them essential for the quality of the 3D representation.

Previous methods failed to fully exploit multi-view priors to jointly control the condition $\boldsymbol{c}$ and noise $\boldsymbol{\epsilon}$, resulting in inconsistent multi-view denoising and limited 3D texture details. Although some incorporate auxiliary priors such as ID features, they rely solely on implicit transfer through embeddings and attention without explicit guidance, leading to weak portrait consistency across views. Moreover, ignoring the influence of $\boldsymbol{\epsilon}$ causes misaligned denoising gradients and loss of fine-grained geometry.

Detail-Rich Portrait Diffusion Pipeline

Our Portrait Diffusion framework for high-fidelity detail-rich 3D portrait generation is illustrated in fig. 2. It consists of two modules:

GAN-Diffusion Portrait Geometry Restoration Module. We propose a GAN-Diffusion hybrid module to reconstruct complete 360° head geometry from a single image, achieving higher geometric fidelity and faster inference. A pretrained GAN serves as the baseline; however, its representation capacity is confined by training data priors, often leading to incomplete 360° shapes and the Janus problem. To overcome these limitations, we introduce diffusion-model priors via an SDS loss to finetune the generator, enabling artifact-free 360° facial geometry synthesis. Directly finetuning the entire GAN, however, may degrade its learned priors and produce coarse geometry. To preserve these priors, we add a *Detail Preservation Block* to the generator and fine-tune only this block during restoration. The optimization is formulated as:

$$\phi^* = \arg\min_{\phi}\sum_{I\in\mathcal{D}} L_{\text{SDS}}\big(\mathcal{G}_{\phi,\psi}(I);\,\psi,\,\theta_{\text{SD, Zero-123}}\big), \tag{3}$$

$$T_I^* = \mathcal{G}_{\phi^*,\psi}(I), \\ \Psi_{0,I} = \{T_I^*,\,\psi_{\text{dec, sr}}\}, \tag{4}$$

where ϕ and ψ denote the parameters of the Detail Preservation Block and the pretrained GAN, respectively, $\theta_{\text{SD, Zero-123}}$ are parameters of the Stable Diffusion and Zero-123 models, $\Psi_{0,I}$ represents the initialized NeRF parameters for image I, and $\psi_{\text{dec, sr}}$ corresponds to NeRF rendering parameters.

While this module refines portrait geometry, it causes some texture detail loss from multi-view inconsistency. We address this with a Multi-view Diffusion Refinement Module built on the geometric prior.

Multi-view Diffusion Refinement Module generates fine-grained 3D texture based on the reconstructed geometry through our Hybrid Priors Diffusion Model and Multi-View Noise Resampling Strategy, as shown in fig. 2 (c).

This method is designed to thoroughly utilize various priors from both conditioning and diffusion procedure perspec-

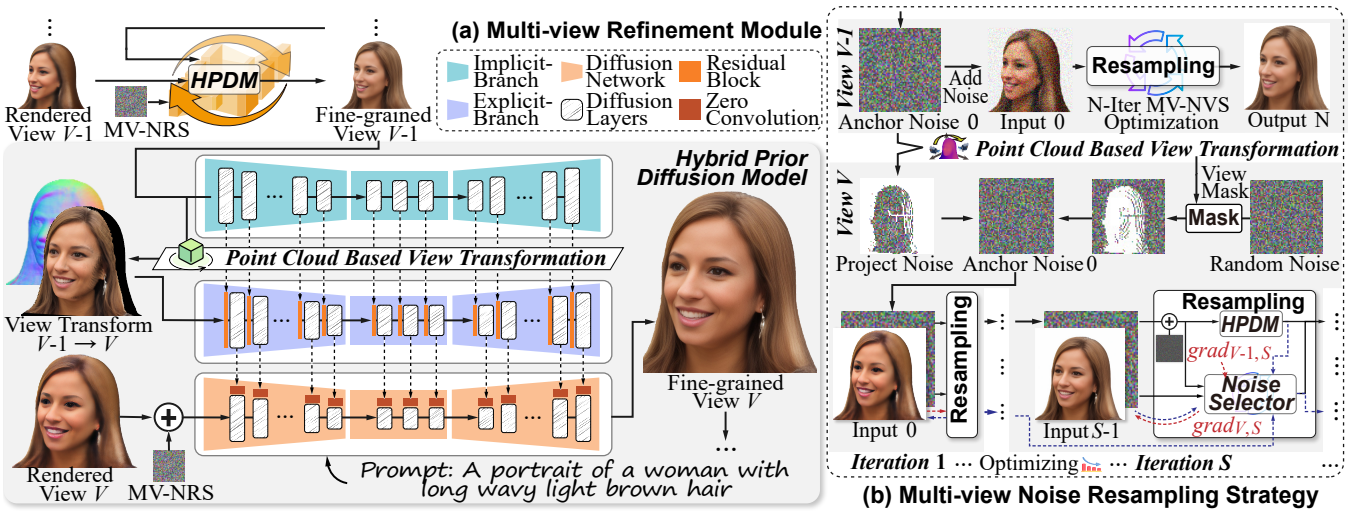

Figure 3: The presentations of our proposed **Hybrid Priors Portrait Diffusion model** (a) and **Multi-View Noise Resampling Strategy** (b). HPDM is designed to leverage various multi-view priors in a hybrid manner to condition the new view synthetic process for more consistent status. NV-NRS is designed to transfer corss-view priors to control the diffusion noise distribution for representations alignment.

tives to improve consistency. From a conditioning perspective, the Hybrid Priors Diffusion Model effectively leverages and transmits multi-view priors both explicitly and implicitly—utilizing additional conditioning branches parameterized by θ^{Ex} and θ^{Im}—to enhance the consistency of novel viewpoints. From a diffusion procedure perspective, we acknowledge the role of noise in conveying detailed multi-view priors and devised a Multi-View Noise Resampling Strategy integrated within the SDS loss ($\mathcal{L}_{\text{SDS}}^{\textbf{MV-NRS}}$), which adjusts the distribution of resampled diffusion noise ϵ^{Rs} for fine-grained representations alignment. Through the $\mathcal{L}_{\text{SDS}}^{\textbf{MV-NRS}}$ and $\theta^{\text{Ex}}, \theta^{\text{Im}}$, we can generate detail-rich portraits:

$$\Psi_I^* = \arg\min_{\Psi} \big(\mathcal{L}_{\text{SDS}}^{\textbf{MV-NRS}}(\Psi, \epsilon^{\text{Rs}}; \{\theta, \theta^{\textbf{Ex}}, \theta^{\textbf{Im}}\}) + \mathcal{L}_{\text{ref}}(\Psi; I) \big), \quad \text{s.t. } \Psi(0) = \Psi_{0,I} \tag{5}$$

Multi-view Status Consistency

fig. 3 (a) presents our **H**ybrid **P**rior **D**iffusion **M**odel (**HPDM**), which focuses on leveraging multi-view priors in a hybrid manner to condition the novel view synthesis process for more consistent portrait status. Initially, we leverage explicit priors by providing reference images to dominate the generation process, offering direct control and constraints. Thus we introduce our *Explicit-Branch*, which takes the image projected from the driving view to the target view as an explicit reference and extends it to fill in the invisible areas.

To generate this reference, we convert the driven view image I_{v_i} into a colored 3D point cloud P_{v_i} using the NeRF-rendered depth map D_{v_i}. Then, a reference image of target view is rendered from this colored point cloud:

$$I_{v_{i+1}}^{\text{Proj}} = \mathcal{R}_{P_{v_i}}(v_{i+1}, I_{v_i}) \tag{6}$$

Besides, the segmentation mask S_{v_i} is similarly rendered onto the target view as an auxiliary mask condition $S_{v_{i+1}}^{\text{Proj}}$. The $z_{v_{i+1}}^{\text{Proj}}$, obtained by encoding $I_{v_{i+1}}^{\text{Proj}}$ from an VAE, along with the $S_{v_{i+1}}^{\text{Proj}}$ are fed into the diffusion UNet.

The diffusion UNet is an adapted version of a pretrained inpainting diffusion UNet (Ju et al. 2024). In this adaptation, the cross-attention components are removed to focus entirely on the reference. Features from this UNet are injected into the frozen layers of the original diffusion UNet layer by layer with zero convs, allowing for dense, pixel-level control over the generation process:

$$\epsilon_\theta(z_t,\ t,\ y)_l = \epsilon_\theta(z_t,\ t,\ y)_l + w^{\text{Ex}} \cdot \mathcal{Z}(\epsilon_\theta^{\text{Ex}}([z^{\text{Proj}},\ S^{\text{Proj}},\ z_t],\ t)_l), \tag{7}$$

where l denotes the l layer of the UNet, $\mathcal{Z}$ denotes zero conv and w^{Ex} denotes control weight.

However, since the reference cannot guide all areas and the degraded priors during the view transformation, relying solely on such explicit priors transfer would introduce noises into control signals.

To address this, we aim to implicitly leverage priors to compensate for these deficiencies. To enhance texture priors, we integrated a second branch, *Implicit-Branch*, for loss-less texture understanding, and designed a res-block to semi-explicitly pass this understanding to the Explicit-Branch. In detail, this branch is a copy of a Explicit-Branch that directly takes the driving image $I_{v_{i-1}}$ as input. To ensure effective transfer of these priors to the Explict-Branch, we first explicitly rendering Implicit-Branch latents into the target view and then implicitly integrating them into the Explicit-Branch through res-blocks with zero-convs. We opt for simple res-blocks rather than complex transformers, benefiting from the spatial prior alignment provided by explicit geo-

metric projection. This process can be expressed as:

$$\begin{aligned}&\epsilon_\theta^{\text{Ex}}([z_{v_i}^{\text{Proj}},\ S_{v_i}^{\text{Proj}},\ z_{t,v_i}],\ t)_l \\ &= \text{Res}\big[\epsilon_\theta^{\text{Ex}}([z_{v_i}^{\text{Proj}},\ S_{v_i}^{\text{Proj}},\ z_{t,v_i}],\ t)_l, \\ &\quad \mathcal{R}_{P_{v_{i-1}}}(v_i, \epsilon_\theta^{\text{Im}}([z_{v_{i-1}},\ z_{t,v_i}],\ t)_l)\big]\end{aligned} \tag{8}$$

Additionally, to compensate for the geometric artifacts in the explicit reference, we incorporate the geometric prior of the current view. To ensure that the generation aligns with the current geometry, the rendered coarse image I^R and normal map N^R are included as additional conditions through a Geo-Block within the Explicit-Branch. To enhance the geometry without overshadowing the reference texture, we once again employ a res-block to implicitly merge these conditions with the reference in latent space:

$$z^{\text{Proj}} = \text{Res}(z^{\text{Proj}}, \text{VAE}(N^R, I^R)) \tag{9}$$

Multi-View Representation Consistency

While the aforementioned conditions help maintain generation consistency, the stochastic nature of the diffusion process still causes multi-view misalignment. To address this, we explicitly control the noise sampling distribution across views by leveraging cross-view priors, ensuring that generated representations remain spatially consistent and visually coherent. Specifically, we introduce a **M**ulti-**V**iew **N**oise **R**esampling **S**trategy (**MV-NRS**) within the SDS loss, comprising two stages: *anchor noise initialization* and *anchor noise optimization*, as illustrated in Figure 3(b).

We first establish an *anchor noise set* $\epsilon_{v_1:v_N}^{\text{Ac}}$ across N views whose generated results exhibit initial alignment. Subsequently, we perform resampling in the local vicinity of this anchor noise to refine inter-view consistency. The resampled noise ϵ^{Rs} follows a Gaussian distribution with small variance σ^2:

$$\epsilon^{\text{Rs}} \sim \mathcal{N}\big(\sqrt{1-\sigma^2}\,\epsilon^{\text{Ac}},\ \sigma^2\mathbf{I}\big). \tag{10}$$

We recognize that the output of the 2D diffusion model demonstrates both invariance to linear transformations and robustness to small-scale nonlinear transformations. Consequently, it exhibits a degree of invariance to small-range viewpoint changes, which can be considered as local linear transformations. Therefore, by aligning the inputs according to the viewpoints, we can ensure that the outputs align as well. Since the input of the UNet consists of a combination of rendered image latents and noises, we only need to align the noises. To achieve this, we just lift the driven view noise into a point cloud and render it onto the target views:

$$\begin{aligned}&\epsilon_{v_{i+1},0}^{\text{Ac}} = \mathcal{R}_{P_{v_i,0}}(\epsilon_{v_i,0}^{\text{Ac}}, v_{i+1}) + \epsilon^{\text{rand}} \odot M_{v_{i+1},0}^{\text{void}} \\ &\epsilon_{v_0,0}^{\text{Ac}} \sim \mathcal{N}(0,\mathbf{I}), \quad \epsilon^{\text{rand}} \sim \mathcal{N}(0,\mathbf{I}),\end{aligned} \tag{11}$$

where $s = 0$ denotes the initial training iteration and M is a mask that indicates the locations of voids in the rendered noise. Compared to random noise initialization, this method uses cross-view priors to build noise, enabling the capture of some small-scale noise distributions that are almost impossible to obtain through pure random sampling, particularly when the multi-view generative distributions are far apart.

Next, since the initial representations may not be perfectly aligned, we utilize multi-view gradient consistency to gradually fine-tune the anchor noises during the SDS training. Specifically, we have designed a Resampling Retention Strategy:

In each training iteration s, we first resampled a noise $\epsilon_{v_i,s}^{\text{Rs}}$ according to (10). Then, we decide whether to keep the resampled noise for updating the anchor noise by utilizing the multi-view gradient consistency score. The key idea is to compute the gradients obtained from both the resampled noise and the anchor noise, and then assess their similarity with the gradients from the driven viewpoint. By comparing these similarities, we can determine whether to retain the resampled noise.

In detail, for $\epsilon_{v_i,s}^{\text{Rs}}$, we first compute the loss between a rendered image $I_{v_i,s}^R$ and denoised image $I_{v_i,s}^D$ from $\epsilon_{v_i,s}^{\text{Rs}}$, then backpropagate it to get the gradient $grad_{v_i,s}^D$:

$$\mathcal{L}_{v_i,s}^D = \mathcal{L}_I(I_{v_i,s}^R, I_{v_i,s}^D) \tag{12}$$

$$grad_{v_i,s}^D = \mathcal{BP}_{\Psi_{s-1}}(\mathcal{L}_{v_i,s}^D) \tag{13}$$

The cosine similarity is used to evaluate the consistency between this gradient and the applied gradient of the driven view $grad_{v_{i-1},s}$:

$$S_{v_i,s}^D = \frac{grad_{v_i,s}^D \cdot grad_{v_{i-1},s}}{\|grad_{v_i,s}^D\|\|grad_{v_{i-1},s}\|} \tag{14}$$

Similarly, for anchor noise $\epsilon_{v_i,s-1}^{\text{Ac}}$, we direct use the $I_{v_i,s-1}$ of previous training iteration to compute the gradient $grad_{v_i,s}^P$ and the corresponding score $S_{v_i,s}^P$.

If $S_{v_i,s}^D > S_{v_i,s}^P$, the resampled noise is superior over the anchor noise in terms of gradient consistency. Therefore, we update the anchor noise and treat $I_{v_i,s}^D$ as the current target image, $grad_{v_i,s}^D$ as the current applied gradient. Conversely, we retain the anchor noise and target image, using the corresponding gradient instead:

$$\begin{aligned}&I_{v_i,s}, \epsilon_{v_i,s}^{\text{Ac}}, grad_{v_i,s} = \\ &\quad \begin{cases} I_{v_i,s},\ \epsilon_{v_i,s}^{\text{Rs}},\ grad_{v_i,s}^D & \text{if } S_{v_i,s}^D > S_{v_i,s}^P \\ I_{v_i,s-1},\ \epsilon_{v_i,s-1}^{\text{Ac}},\ grad_{v_i,s}^P & \text{otherwise} \end{cases}\end{aligned} \tag{15}$$

After calculations for all viewpoints, we aggregate the gradients across all views, denoted as $Grad_s$. Finally, we update the 3D model in the current training iteration:

$$Grad_s = \sum_v grad_{v_i,s} \tag{16}$$

$$\Psi_s = \Psi_{s-1} + \alpha_s \cdot Grad_s \tag{17}$$

Experiments

Implementation Details

We use 128×128 triplane features with 96 channels to render 512×512 images. For geometry restoration, we sample 7-9 azimuth views for SDS training; texture restoration uses fixed viewpoints and an additional image-supervised multi-step denoising SDS loss for enhanced consistency.

Figure 4: Qualitative comparison to SOTA approaches: Era3D (Li et al. 2024a), DreamCraft3D (Sun et al. 2024), Portrait3D (Wu et al. 2024), ID-Sculpt (Hao et al. 2024), and Arc2Avatar (Gerogiannis et al. 2025). Our *Portrait Diffusion* produces photorealistic 3D portraits with intricate textures in facial features and individual hair strands; ***zoom in to see our details!***.

Our HPDM is trained and tested on GAN-generated in-the-wild data, with a diffusion control model fine-tuned to generate outputs conditioned on a specified prior. Evaluation was conducted on 30 head images, including 8 celebrities and 25 in-the-wild samples with diverse poses and expressions. All experiments ran at A100 GPUs, with training and inference speeds comparable to Mip-NeRF.

Qualitative Results

We compare Portrait Diffusion with open-source SOTA methods: Portrait3D (Wu et al. 2024) (originally text-2-3D, but we bypass text-2-image step with the reference image as input for fair comparison), DreamCraft3D (Sun et al. 2024), Era3D (Li et al. 2024a), ID-Sculpt (Hao et al. 2024), and Arc2Avatar (Gerogiannis et al. 2025). As shown in fig. 4, Era3D yields overly smooth, toy-like textures; DreamCraft3D loses fine detail in profile views; Portrait3D and Arc2Avatar suffer from identity inconsistency, blurry textures, and hair artifacts; and ID-Sculpt, while richer in texture, introduces unnatural geometry artifacts. In contrast, our approach delivers the highest fidelity and texture detail.

Quantitative Evaluation

To comprehensively assess the quality of the generated head models, we employ three metrics: CLIP-I for global structural consistency, LPIPS for perceptual similarity, and ID for identity preservation. CLIP-I measures the cosine similarity between the CLIP feature embeddings of rendered and reference images. All metrics are computed over five rendered views—front, left-frontal, left, right-frontal, and right—relative to the input reference. Both variants, *Portrait Diffusion-Fast* (1h training) and *Portrait Diffusion-Detail* (3h training), obtain the highest CLIP-I scores, reflecting strong semantic coherence across multiple viewpoints. They also achieve the lowest LPIPS values, indicating superior perceptual alignment, while maintaining the highest ID accuracy for consistent identity reproduction.

Method	CLIP-I ↑	LPIPS ↓	ID ↑	Train
Era3D	0.9934	0.4053	0.2845	0.5h
Portrait3D	0.9956	0.4258	0.1899	1h
ID-Sculpt	0.9954	0.4425	0.3045	1h
Arc2Avatar	0.9948	0.4215	0.275	1.5h
DreamCraft3D	0.9969	0.4064	0.2314	6h
Portrait Diffusion	0.9982	0.3625	0.3438	1h
Portrait Diffusion-D	**0.9986**	**0.3616**	**0.3438**	3h

Table 1: Quantitative comparison to SOTA approaches.

Method	LPIPS↓	SSIM↑
Ex-Branch	0.4013	0.1008
+Geo-Block	0.0685	0.8024
+Im-Branch	0.0405	0.8732

(a) HPDM Ablation Study.

Method	NIQE↓	DS↑
Random noise	4.1552	0.6104
Resampled	4.1324	0.6412
MV-NRS	3.9143	0.7263

(b) MV-NRS Ablation Study

Table 2: Quantitative Ablation Study Comparison.

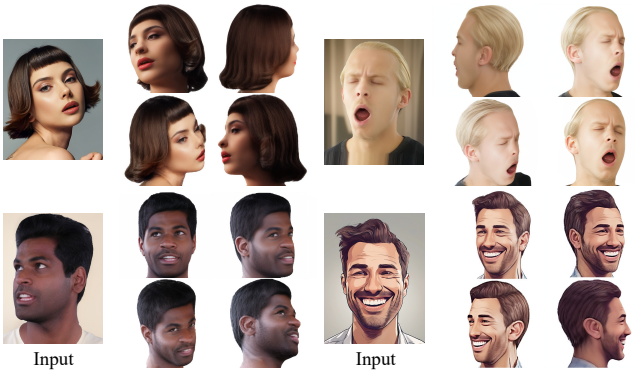

Figure 5: Robustness evaluation with more novel poses and views, exaggerated expressions and stylization.

Ablation Study

Effectiveness of Geo Reconstruction Modules. fig. 6 presents ablation results for the GAN Prior and Portrait Geometry Restoration modules. GAN priors yield a coarse head with awkward distributions and multi-face artifacts. Diffusion alleviates these issues, but results in over-smooth textures and rough surfaces, affecting depth projection and geometry priors. The Detail Preservation Network further refines surfaces, enabling more accurate downstream tasks.

Effectiveness of the Hybrid Priors Diffusion Model (HPDM). Figure 8 illustrates that relying solely on the Explicit Branch results in disordered geometry and stripe artifacts. Introducing the geo-block aligns geometric priors but residual stripes persist. When the Implicit Branch with texture priors is incorporated, these artifacts vanish, yielding coherent geometry and clean textures. The quantitative metrics further verify that HPDM effectively leverages complementary priors for high-fidelity reconstruction.

Effectiveness of Multi-View Noise Resampling (MV-NRS). Figure 7 demonstrates that standard random noise preserves multi-view consistency but produces blurred textures. Anchor Noise increases determinism yet introduces misalignment and view-specific artifacts. The proposed MV-NRS, which optimizes Anchor Noise under multi-view constraints, achieves sharper and more consistent textures. Quantitative evaluations uses NIQE and Discriminator Score (DS), confirm its advantage in fine-detail reconstruction.

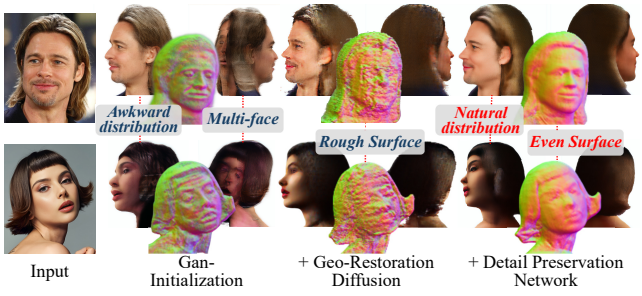

Figure 6: Visual Results for Ablation study on Geo Reconstruction Modules. ***Zoom in for details***.

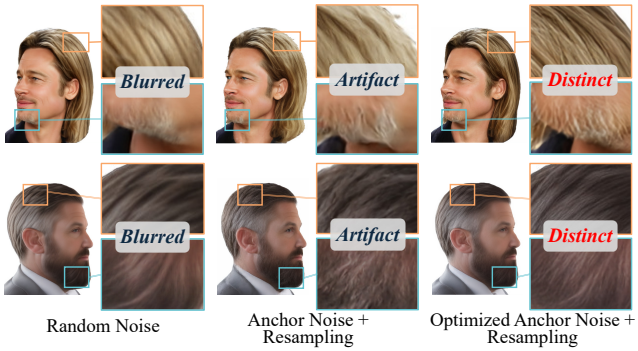

Figure 7: Visual Results for Ablation study on Multi-View Noise Resampling Strategy (MV-NRS). ***Zoom in for details***.

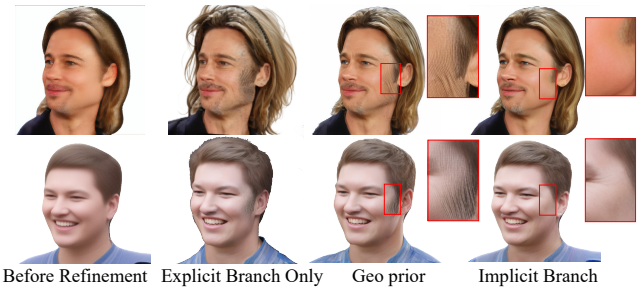

Figure 8: Visual Results for Ablation study on Hybrid Priors Diffusion Model (HPDM). ***Zoom in for details***.

Robustness Evaluation. Figure 5 presents results under challenging conditions such as non-frontal poses, exaggerated expressions, large vertical viewpoints, and stylized faces, where our method still maintains structural and textural fidelity.

Conclusion

In this paper, we proposed a Portrait Diffusion pipeline for detail-rich 3D full portrait generation. The pipeline comprises a Portrait Geometric Restoration module and a Multi-view Diffusion Refinement module. The latter incorporates a Hybrid Priors Diffusion model that effectively leverages multi-view priors for consistent status, and a Multi-View Noise Resampling Strategy to ensure consistent representations. Qualitative and quantitative assessments have shown that our pipeline exhibits superior detail and realism compared to SOTA alternatives. Overall, Portrait Diffusion sets a new benchmark for 3D portrait generation and opens new directions for computer vision and graphics research.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund of Macau SAR (FDCT) under grants 0102/2023/RIA2 and 0154/2022/A3 and 001/2024/SKL, and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC), and National Natural Science Foundation of China (Grant No. 62471342).

References

An, S.; Xu, H.; Shi, Y.; Song, G.; Ogras, U. Y.; and Luo, L. 2023. PanoHead: Geometry-aware 3D Fullhead Synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20950–20959.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; and Khamis, S. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Corneanu, C.; Gadde, R.; and Martinez, A. M. 2024. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4334–4343.

Deng, Y.; Wang, D.; and Wang, B. 2024a. Portrait4D: Learning One-Shot 4D Head Avatar Synthesis using Synthetic Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Deng, Y.; Wang, D.; and Wang, B. 2024b. Portrait4D-v2: Pseudo Multi-View Data Creates Better 4D Head Synthesizer. *arXiv preprint arXiv:2403.13570*.

Deng, Y.; Yang, J.; Xiang, J.; and Tong, X. 2022. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10673–10683.

Doukas, M. C.; Zafeiriou, S.; and Sharmanska, V. 2021. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, 14398–14407.

Dundar, A.; Gao, J.; Tao, A.; and Catanzaro, B. 2023. Progressive learning of 3d reconstruction network from 2d gan data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gao, J.; Shen, T.; Wang, Z.; Chen, W.; Yin, K.; Li, D.; Litany, O.; Gojcic, Z.; and Fidler, S. 2022. GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images. *arXiv preprint arXiv:2209.11163*.

Gerogiannis, D.; Papantoniou, F. P.; Potamias, R. A.; Lattas, A.; and Zafeiriou, S. 2025. Arc2avatar: Generating expressive 3d avatars from a single image via id guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10770–10782.

Hao, J.; Tang, J.; Zhang, J.; et al. 2024. Portrait3D: 3D Head Generation from Single In-the-wild Portrait Image. *arXiv preprint arXiv:2406.16710*.

Jiang, J.; Lin, G.; Rong, Z.; Liang, C.; Zhu, Y.; Yang, J.; and Zhong, T. 2024. MobilePortrait: Real-Time One-Shot Neural Head Avatars on Mobile Devices. *arXiv preprint arXiv:2407.05712*.

Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*.

Kirschstein, T.; Giebenhain, S.; and Nießner, M. 2024. DiffusionAvatars: Deferred Diffusion for High-fidelity 3D Head Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, P.; Liu, Y.; Long, X.; Zhang, F.; Lin, C.; Li, M.; Qi, X.; Zhang, S.; Luo, W.; Tan, P.; et al. 2024a. Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention. *arXiv preprint arXiv:2405.11616*.

Li, X.; De Mello, S.; Liu, S.; Nagano, K.; Iqbal, U.; and Kautz, J. 2024b. Generalizable one-shot 3D neural head avatar. *Advances in Neural Information Processing Systems*, 36.

Ma, Z.; Zhu, X.; Qi, G.-J.; Lei, Z.; and Zhang, L. 2023. Otavatar: One-shot talking face avatar with controllable triplane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16910.

Mirzaei, A.; Aumentado-Armstrong, T.; Derpanis, K. G.; Kelly, J.; Brubaker, M. A.; Gilitschenski, I.; and Levinshtein, A. 2023. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20669–20679.

Poole, B.; Jain, A.; Barron, J.; and Mildenhall, B. 2022a. Dreamfusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022b. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Reddy, P.; Elezi, I.; and Deng, J. 2024. G3DR: Generative 3D Reconstruction in ImageNet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9655–9665.

Shao, R.; Pang, Y.; Zheng, Z.; Sun, J.; and Liu, Y. 2024. Human4DiT: Free-view Human Video Generation with 4D Diffusion Transformer. *arXiv preprint arXiv:2405.17405*.

Sun, J.; Zhang, B.; Shao, R.; Wang, L.; Liu, W.; Xie, Z.; and Liu, Y. 2024. DreamCraft3D: Hierarchical 3D Generation with Bootstrapped Diffusion Prior. In *The Twelfth International Conference on Learning Representations*.

Tang, J.; Wang, T.; Zhang, B.; Zhang, T.; Yi, R.; Ma, L.; and Chen, D. 2023. Make-it-3D: High-fidelity 3D creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*.

Tang, J.; Zeng, Y.; Fan, K.; Wang, X.; Dai, B.; Chen, K.; and Ma, L. 2024. Make-it-vivid: Dressing your animatable biped cartoon characters from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6243–6253.

Wang, D.; Zhang, T.; Abboud, A.; and Süsstrunk, S. 2024. InNeRF360: Text-Guided 3D-Consistent Object Inpainting on 360-degree Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12677–12686.

Wu, Y.; Xu, H.; Tang, X.; Fu, H.; and Jin, X. 2023. 3DPortraitGAN: Learning One-Quarter Headshot 3D GANs from a Single-View Portrait Dataset with Diverse Body Poses. *arXiv preprint arXiv:2307.14770*.

Wu, Y.; et al. 2024. Portrait3D: Text-guided High-Quality 3D Portrait Generation Using Pyramid Representation and GANs Prior. *ACM Transactions on Graphics (TOG)*, 43(4): 1–12.

Xiang, J.; Yang, J.; Deng, Y.; and Tong, X. 2022. GRAM-HD: 3D-Consistent Image Generation at High Resolution with Generative Radiance Manifolds. *arXiv preprint arXiv:2206.07255*.

Xiang, S.; Gu, Y.; Xiang, P.; He, M.; Nagano, K.; Chen, H.; and Li, H. 2020. One-shot identity-preserving portrait reenactment. *arXiv preprint arXiv:2004.12452*.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

Xie, Y.; Xu, H.; Song, G.; Wang, C.; Shi, Y.; and Luo, L. 2024. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Yang, S.; Chen, X.; and Liao, J. 2023. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3190–3199.

Ye, Z.; Zhong, T.; Ren, Y.; Yang, J.; Li, W.; Huang, J.; Jiang, Z.; He, J.; Huang, R.; Liu, J.; et al. 2024. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*.

Yin, F.; Zhang, Y.; Wang, X.; Wang, T.; Li, X.; Gong, Y.; Fan, Y.; Cun, X.; Shan, Y.; and Oztireli, C. 2023a. 3D GAN Inversion with Facial Symmetry Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 342–351.

Yin, F.; Zhang, Y.; Wang, X.; Wang, T.; Li, X.; Gong, Y.; Fan, Y.; Cun, X.; Shan, Y.; Oztireli, C.; et al. 2023b. 3d gan inversion with facial symmetry prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 342–351.

Zhang, B.; Cheng, Y.; Wang, C.; Zhang, T.; Yang, J.; Tang, Y.; Zhao, F.; Chen, D.; and Guo, B. 2024. Rodinhd: High-fidelity 3d avatar generation with diffusion models. *arXiv preprint arXiv:2407.06938*.