Zero-1-to-A: Zero-Shot One Image to Animatable Head Avatars Using Video Diffusion

Supplementary Material

7. Introduction

In this supplementary material, we provide additional details and insights into the work presented in the paper.

- Sec. 8 details the image preprocessing steps and explores potential applications for generating avatars.
- Sec. 9 discusses related solutions, highlighting their limitations, the differences from our approach, and the advantages these differences bring.
- Sec. 10 presents comprehensive experimental results, including qualitative comparisons with image-to-3D methods, quantitative evaluations against video diffusion models, and various ablation studies.

Furthermore, we visualize the spatial and temporal inconsistencies in video diffusion models and demonstrate the improvements introduced by our method (Sec. 9.2). In the ablation study, we evaluate the adaptability of our method to different video diffusion models and its applicability to in-the-wild datasets.

8. Additional Implementation Details

Pre-process. Based on [15, 17, 61], the preprocessor removes the background and estimates the pose and FLAME parameters of input portrait. As shown in Fig. 11, the preprocess of each image contains two steps: 1) Background Removal: Given a portrait image, we first use Rembg [61] to remove the background and only retain the foreground portrait; 2) Pose Estimation: Then, we use MICA [92] and EMOCA [17] to estimate the FLAME shape, expression and pose parameters of portrait. In training, we use the carved image as the input of video diffusion and initialize the learnable shape parameter with estimated FLAME shape.

Application. Once optimized, the parameters of the animatable Gaussian head are fixed, enabling real-time animation and rendering of the avatar using motion and camera sequences. These motion sequences incorporate expression and pose parameters from FLAME 2020 [40]. Following HeadStudio [90], we employ advanced models such as face-to-FLAME [17, 19, 92], speech-to-FLAME [83], and text-to-speech [1] to convert video, speech, and text into FLAME animation inputs. As a result, the generated avatar supports multimodal control, demonstrating practical applications in real-world scenarios.

Figure 11. **Pre-process.** Given a portrait image, we first remove its background and then estimate the FLAME parameter.

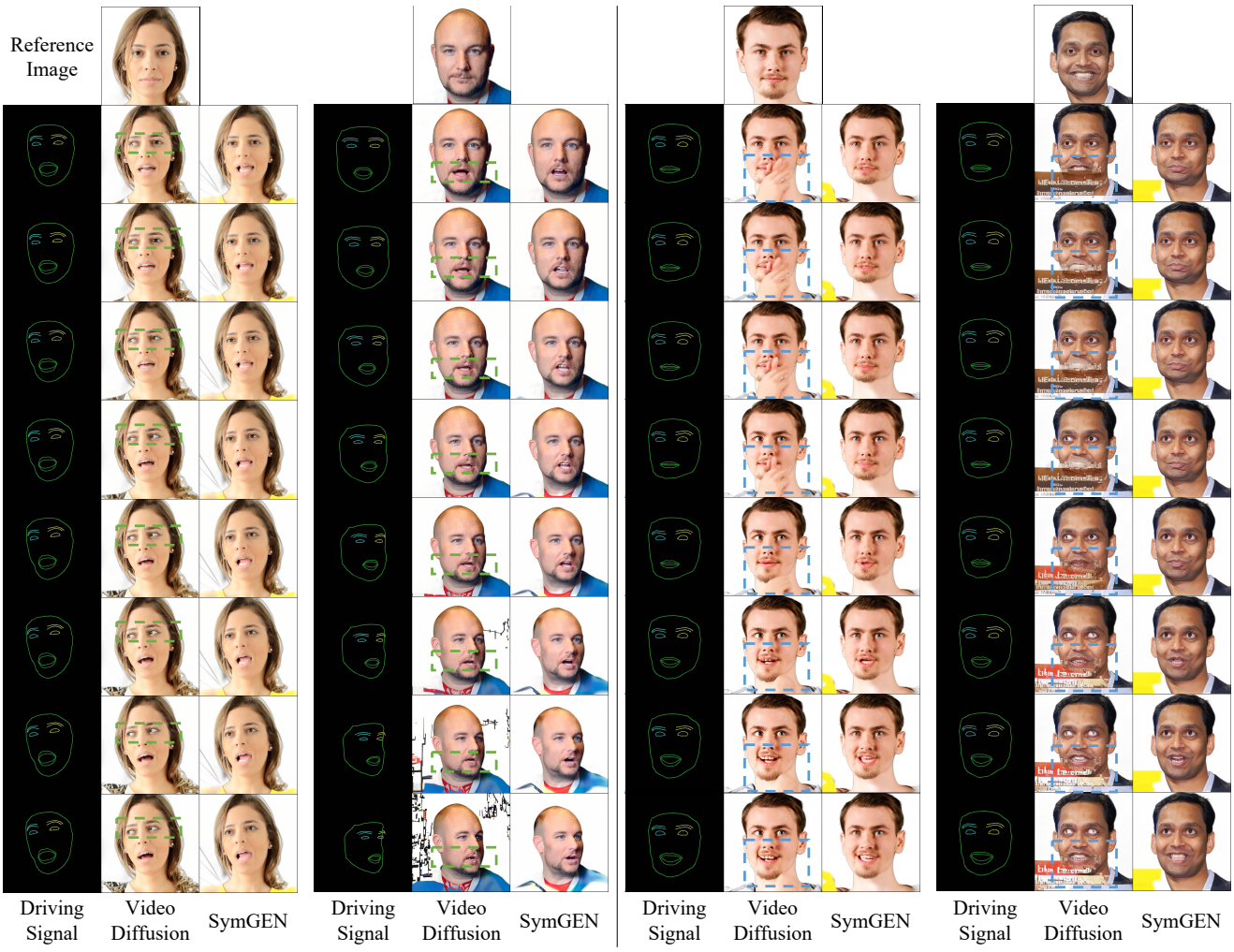
Figure 12. **2D Image Generation with SDS-based Loss.** From left to right: reference image, SDS [58], ISM [42], NFSD [35] and video diffusion generation [73].

9. Discussion

9.1. Discussion with Related Solutions

DreamFusion v.s. Zero-1-to-A. In Fig. 12, we compare 2D image generation results using SDS [58], ISM [42], and NFSD [35]. Notably, NFSD employs negative prompts to suppress unwanted noise in the diffusion score. We implement this by treating reference image with data augmentation (e.g., blur, brightness adjustment, Gaussian noise) as negative prompts. Compared to video diffusion generation, results from SDS-based loss exhibit issues of over-smoothing and over-saturation. We attribute this to additional temporal modules introduced in portrait video diffusion, which may adversely affect score distillation. This limitation motivates us to explore alternative solutions.

Instruct-NeRF2NeRF (IN2N) v.s. Zero-1-to-A. IN2N [25] introduces a 3D editing method called iterative dataset update, which alternates between editing the ground-truth dataset and optimizing the 3D scene. In contrast, Zero-1-to-A is a 4D generation method that progressively builds a pseudo ground-truth dataset while optimizing the 4D avatar. Different from the editing task, the lack of consistent input in the generation task creates a negative cycle in iterative dataset update, leading to incorrect convergence (*e.g.*, misaligned eyes and inability to open the mouth, as shown in the fifth column of Fig. 8). To



(a) Spatial Inconsistency

(b) Temporal Inconsistency

Figure 13. **Visualization of Spatial and Temporal Inconsistencies in Video Diffusion Models.** Portrait video diffusion exhibits spatial inconsistencies, such as incorrect eye positioning in side views (green boxes), and temporal inconsistencies, evident in significant changes triggered by minor facial expressions (blue boxes).

Table 2. **Quantitative Evaluation of Avatar Animation.** We evaluate ID consistency, temporal smoothness, and rendering speed, demonstrating that our method is able to enhance the performance of portrait video diffusion.

Face Animation	ID ↑	Motion ↑	Speed ↑
AniPortrait [73]	**0.5081**	0.8410	0.52 FPS
Follow-Your-Emoji [53]	0.4988	0.8934	0.56 FPS
Ours (w. [53])	0.5000	**0.9187**	**71 FPS**

address this, we propose a simple-to-complex progressive learning strategy that breaks this cycle and significantly improves generation performance.

9.2. Discussion on the Motivation

In Fig. 13, we show the spatial and temporal inconsistencies in video diffusion models and demonstrate the improvements achieved by our method. On the left of Fig. 13, spatial inconsistencies are shown by fixing the expression and varying the camera pose. Ideally, the portrait's expression should remain unchanged. However, as the camera pose shifts, the iris incorrectly looks left, and teeth that were initially absent appear (highlighted in green boxes). On the right, temporal inconsistencies are illustrated by fixing the camera pose and varying the expression. Ideally, the portrait should deform smoothly and accurately. Instead, even with minor changes, such as gradually opening the mouth, the generated video exhibits abrupt and incorrect variations (highlighted in blue boxes). With SymGEN, we achieve improvements in video generation under large pose changes and exaggerated expressions, resulting in a spatially and temporally consistent pseudo-ground truth dataset.

In summary, video diffusion models [53, 73] suffer from

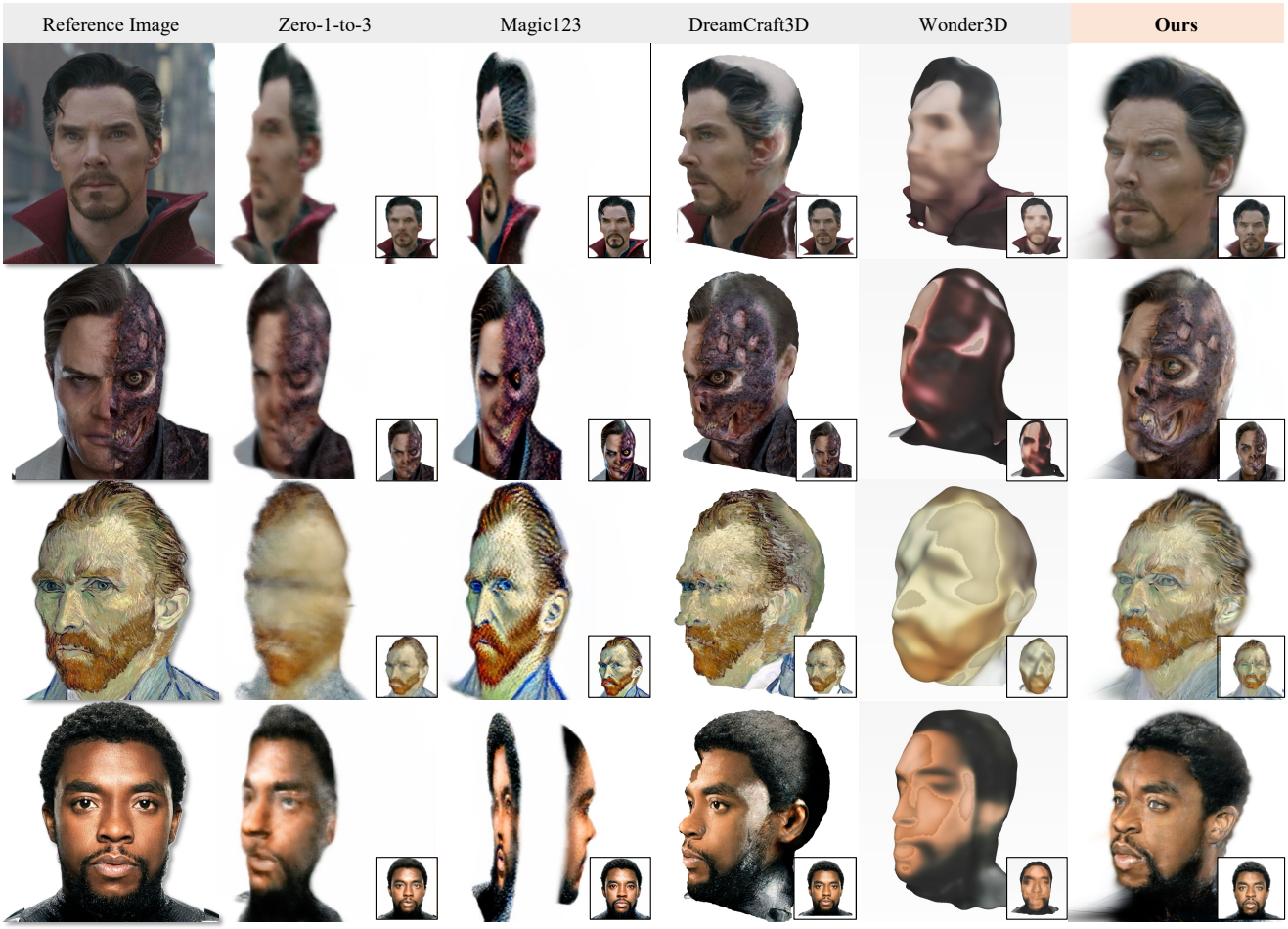


Figure 14. **Comparisons with Image-to-3D Methods.** Our method delivers comparable performance in texture reconstruction while achieving superior 3D consistency.

Figure 15. **Comparisons with Portrait3D [24].** Our method matches the performance of Portrait3D while providing animatable avatars, enabling a wider range of applications.

severe spatial and temporal inconsistencies, making them unsuitable for direct 4D avatar reconstruction. Our proposed SymGEN framework iteratively constructs a consistent dataset, enabling the reconstruction of 4D avatars.

10. Additional Experiments

10.1. Comparisons with Image-to-3D Methods

In Fig. 14, we compare our method with diffusion-based image-to-3D approaches, including Zero-1-to-3 [46], Magic123 [59], DreamCraft3D [67], and Wonder3D [49]. We reproduce Zero-1-to-3, Magic123, and DreamCraft3D using threestudio[†] and implement Wonder3D with NeuS following the official guidelines[†]. The results show that

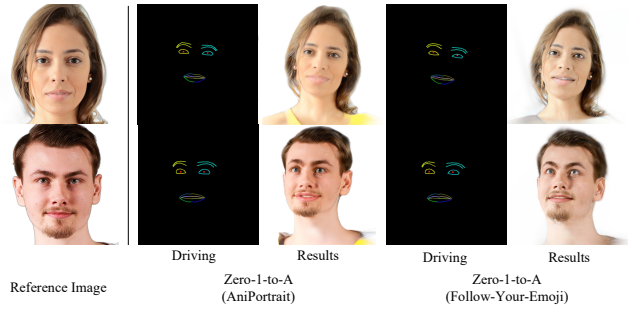


Figure 16. **Evaluation on Different Video Diffusion Models.** Our method demonstrates its effectiveness by seamlessly adapting to various video diffusion models.

[†]https://github.com/threestudio-project/threestudio
[†]https://github.com/xxlong0/Wonder3D

our Zero-1-to-A delivers comparable texture fidelity and superior geometry reconstruction, leveraging a head prior model.

Comparisons with Portrait3D. Portrait3D [24] is a diffusion-based image-to-avatar method. However, as the code is not yet open-sourced, we could not reproduce its results for the avatar generation benchmark [44, 90]. In Fig. 15, we compare our method with Portrait3D using results captured from its official project[†]. Our method achieves comparable performance while offering animatable avatars, enabling broader applications than Portrait3D.

10.2. Comparisons with Video Diffusion Methods.

In Tab. 2, we quantitatively evaluate ID consistency, temporal smoothness, and rendering speed. ID consistency (ID) is measured using cosine similarity of identity embeddings, while temporal smoothness (Motion) is evaluated using a stability score based on frequency analysis of estimated 2D motion. Higher low-frequency energy indicates greater video stability (details in [47]). The evaluation is conducted on 18 samples and 300 frames from real-world portrait videos [83]. Our method demonstrates improved ID consistency, temporal smoothness, and rendering speed, highlighting the effectiveness of Zero-1-to-A.

10.3. Additional Ablations

Evaluation on Different Video Diffusion Models. In Fig. 16, we compare results using different video diffusion models (AniPortrait [73] and Follow-Your-Emoji [53]). Notably, using AniPortrait achieves better color fidelity to the reference image than using Follow-Your-Emoji. Our method adapts seamlessly to various video diffusion models, effectively generating animatable avatars and demonstrating robust performance.

Evaluation on in-the-wild Cases. We further evaluate our method on a wide range of in-the-wild cases. Specifically, we sampled multiple portraits from the FFHQ dataset [34], with results shown in Fig. 17. The results demonstrate that our approach is broadly applicable across genders and diverse ethnic groups.

[†]https://jinkun-hao.github.io/Portrait3D/

Figure 17. **Evaluation on in-the-wild Cases.**