Supplementary Materials of
DivAvatar: Diverse 3D Avatar Generation with a Single Prompt

A. More Qualitative Comparisons

We provide additional examples on the qualitative comparisons between DivAvatar, Stable Dreamfusion, AvatarCraft and AvatarVerse. For each prompt, we obtain five different samples for each method. The comparison is shown in Figs. 3 and 4 below. Our method demonstrates significantly higher level of diversity as compared to Stable Dreamfusion, AvatarCraft and AvatarVerse.

While the code for the general text-to-3D approach Prolific Dreamer [3] has been made public, its capability for diverse generation is limited to the early, coarse stage of the process. This stage can generate up to four samples at a time, but the results are intermediate avatars that, although diverse, are blurry and of lower quality. Additionally, the approach lacks support for incorporating SMPL geometry prior, leading to avatars with incomplete body shapes. The refining stage of the model is further constrained to refining of only a single sample from the coarse stage. Given these limitations, it is not suitable to directly compare the initial, diverse but coarse results from Prolific Dreamer with those from the baselines and DivAvatar. Therefore, we have included the results from Prolific Dreamer in the Supplementary Materials here in Fig. 5 for two specific prompts. In contrast, DivAvatar allows for the generation of an unlimited number of diverse samples by sampling noises during inference.

B. More Implementation Details

In the process of finetuning θ , we add the SDS loss and the feature-based depth loss to the generator only. The output rendering resolution of the human generative model is 512x256. We retrieve and pad the rendered front and back view of each generated human to 512x512 for img2img refinement [1].

We use the publicly available code [2] for DMTet finetune, where the output rendered image is shape 1024x1024. To ensure that the diverse appearances from the finetuned generative model are not lost, we use the refined front and back images as image conditions. We alternate between SDS loss and the MSE loss at each iteration. We use the default value 1000 and 1 for MSE loss and SDS loss weights respectively.

The finetune of the human generative model process takes around 2 hours for one prompt, involving 5000 iterations. The inference process to obtain one sample is around 30 seconds. The mesh optimization process of each sample takes an average of 30 minutes, involving img2img refinement and 5000 iterations of DMTet finetune.

B.1. Semantic Zoom details

As mentioned in the main paper, we define 6 semantic regions: upper body and lower body with their respective back views, left hand and right hand. We provide further explanation on the mechanism here. We define the front view regions as encompassing the upper body, lower body, left hand, and right hand. Conversely, the back views include the upper body backview and lower body backview. The upper body is composed of the head and torso, while the lower body comprises the pelvic region and legs. During the training process, the focus shifts to specific body parts with varying frequency to capture necessary detail: every three iterations for the upper body, every four for hands, and every five for the lower body. Additionally, there is a 0.4 chance during each upper or lower body iteration to switch the rendering viewpoint to a backview, ensuring both front and back perspectives are adequately generated. We employ SDS loss alongside adapted text prompts tailored to the zoomed-in regions. For example, at iteration 2, the text prompt will describe "a full body rendering of a farmer." By iteration 3, this adjusts to either "an upper body rendering of a farmer" with a 0.6 probability, or a "backview of the upper body rendering of a farmer" with a 0.4 probability. By iteration 4, the focus will shift to the hands, with an equal chance of modifying the prompt to "right hand of a farmer" or "left hand of a farmer."

C. More Quantitative Comparisons

C.1. Comparison of duration

Based on the aforementioned implementation details, we provide comparisons of computation requirements between the baselines in Tab. 1 below. We report duration for a single sample, and for multiple samples generation that can

highlight the use case of our method. All experiments are conducted in a single Tesla V100 32GB GPU.

In the provided comparison of computation requirements detailed in Tab. 1, it's evident that our method, DivAvatar, excels particularly in scenarios requiring the generation of a large number of diverse samples. DivAvatar's initial coarse generation phase takes 120 minutes, during which it can produce an unlimited number of coarse samples. Subsequently, each sample requires just an additional 30 minutes for mesh optimization. Notably, the time for coarse generation remains constant at 120 minutes regardless of the number of samples produced, with only the mesh optimization time scaling linearly with the number of samples.

In contrast, the Prolific Dreamer method [3], is structured in three stages. It begins by generating up to four coarse diverse samples within 168 minutes. However, its second stage—refining the geometry—takes 85 minutes and is limited to just one of these samples. The final texturing stage also focuses solely on the refined sample and requires an additional 165 minutes. Other methods like Stable Dreamfusion, AvatarCraft, and AvatarVerse do not support the generation of diverse results, and Prolific Dreamer itself restricts diversity by only finalizing one refined result from a batch. Overall, DivAvatar stands out for its ability to consistently generate an unlimited number of distinct samples efficiently, making it especially suitable for applications that demand high diversity without necessitating multiple training loops.

Table 1. Comparison of computational duration with existing baselines. Values are in minutes.

	1 Sample	4 Samples	8 Samples
Stable Dreamfusion	30	120	240
AvatarCraft	220	880	1760
AvatarVerse	200	800	1600
ProlificDreamer	420	420	840
DivAvatar	150	240	360

C.2. Comparison of memory usage

We show memory usage and render resolution of the methods in Tab. 2.

D. Additional studies

D.1. Broader range of p values.

The image in Fig. 2 showcases five randomly retrieved samples during inference. We extend the analysis of p values beyond the 1.0 and 0.1 discussed in the main paper, as detailed below. Our noise sampling strategy introduces significantly greater diversity in both the coarse and the refined appearances. Specifically, when p=0.1, the highest level of diversity is observed. This diversity persists at p=0.3, though the variations in appearance are less pronounced compared to p=0.1.

Table 2. Comparison of computational usage with existing baselines.

	Usage (GB)	Resolution
Stable Dreamfusion	8	512x512
AvatarCraft	22	128x128
AvatarVerse	20	512x512
ProlificDreamer	27	512x512
DivAvatar	28	512x256

At p=0.5, diversity is present but restricted, primarily manifesting in different accessories across the coarse and refined appearances. The batch, however, tends to look quite uniform. As p increases to 0.7 and further to 1.0, the differences between coarse results become minimal. Nevertheless, slight variations remain in the clothing prints of the refined results, attributed to minor alterations introduced by the img2img pipeline during mesh optimization. This demonstrates that the main source of diversity is from our strategic sampling itself, while the subsequent mesh optimization contributes only minor variations that do not significantly impact the overall diversity of results.

In summary, utilizing a fixed noise source throughout the majority of iterations ensures diversity. Conversely, as the likelihood of random noise increases (by increasing p), observed diversity diminishes. Our method, which proposes keeping the noise primarily constant (keeping a low value of p) , proves capable of generating diverse samples in real-time during inference.

D.2. More diverse poses.

In Fig. 1, we showcase a variety of poses that go beyond the typical casual human stances often observed. Sampling diverse poses during inference does not compromise the efficiency or memory consumption of our method. We demonstrate the ability to generate varied gestures, primarily controlled by the SMPL parameter θ, as evidenced by the unique poses in Samples 4 and 5. Additionally, variation in body shape is influenced by the SMPL parameter β, illustrated by the differences in height: Samples 3 to 5 feature avatars of a dwarfed stature, whereas Samples 1 and 2 depict avatars of normal height.

E. Limitations and Future Work

Even though we utilise a GAN model for diverse appearances in inference times, the output texture lacks photorealistic details, necessitating additional mesh optimiza-

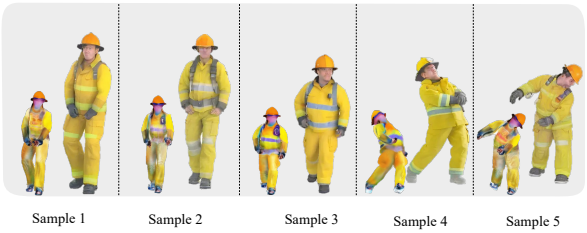

Figure 1. Failure case example. Text prompt: 'A firefighter'. The samples were obtained from model trained with p=0.1 which should have generated diverse appearances.

tion for each sample. Enhancing texture quality for direct high-quality avatar generation without further refinement remains for future work. Additionally, our model shows limited diversity in specific uniforms like nurses or firefighters (in Fig. 1 below), likely due to reliance on appearance priors from the EVA3D's training on the DeepFashion dataset. If prompt subject is not in the training dataset, the main source of appearance will be dependent on the SDS loss which tends to converge. Even though we integrate an open-world text-to-image model (Stable Diffusion) as an additional prior, this issue persists because such models also face challenges with underrepresented categories like specific uniforms. Furthermore, we encounter ongoing challenges inherent to EVA3D, such as watermark artifacts and inability to generate loose clothes.

F. Ethical Considerations

While our technique is capable of creating lifelike 3D human representations, we are aware of the potential for its misuse. Such technology, in the wrong hands, could contribute to the creation of deepfakes that are indistinguishable from real humans, potentially being used to fabricate misleading or harmful media content. We emphasize our ethical obligation to promote the correct application of this technology, which we are confident can benefit both the research community and various industry sectors.

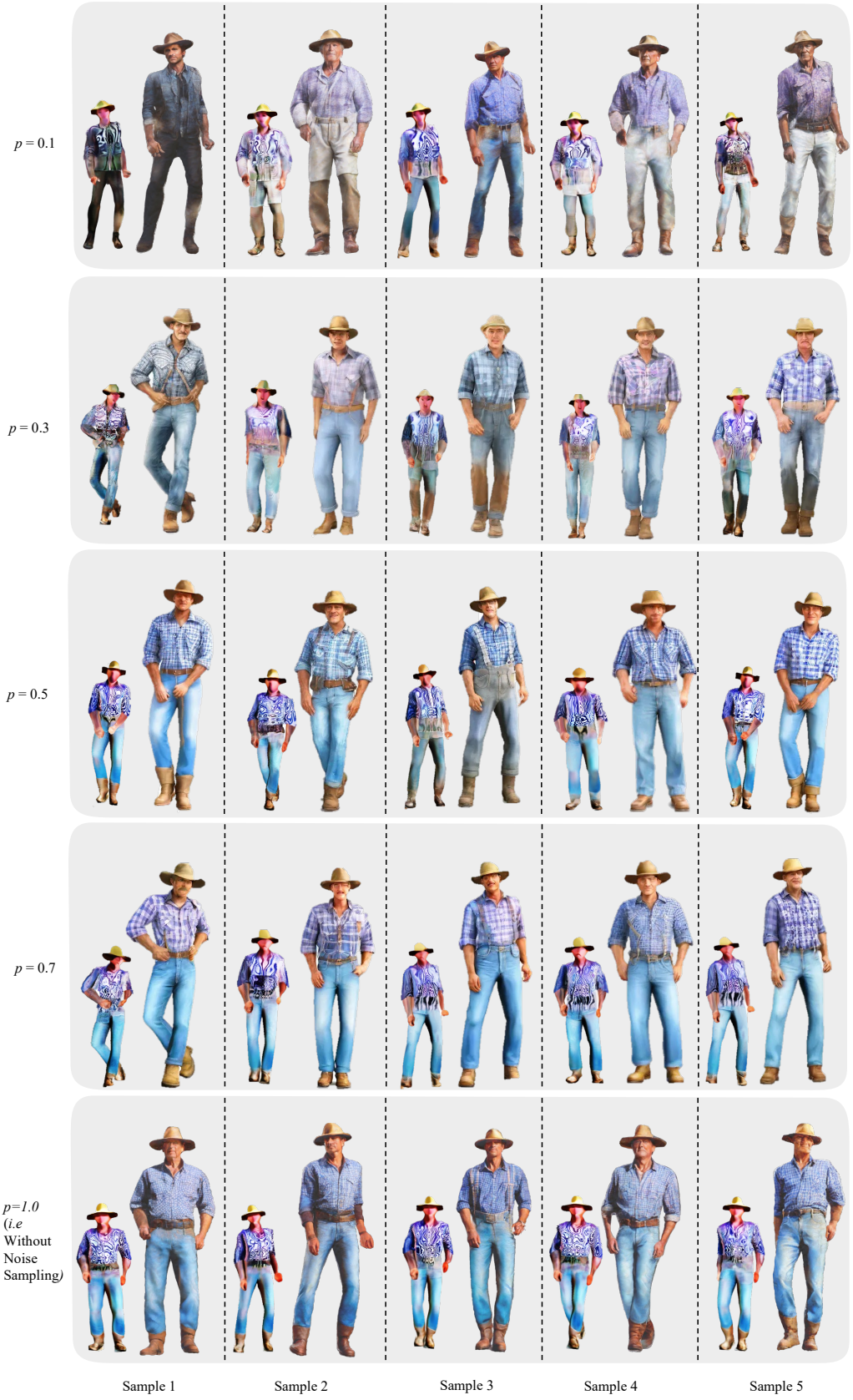

Figure 2. A wide range of p values.

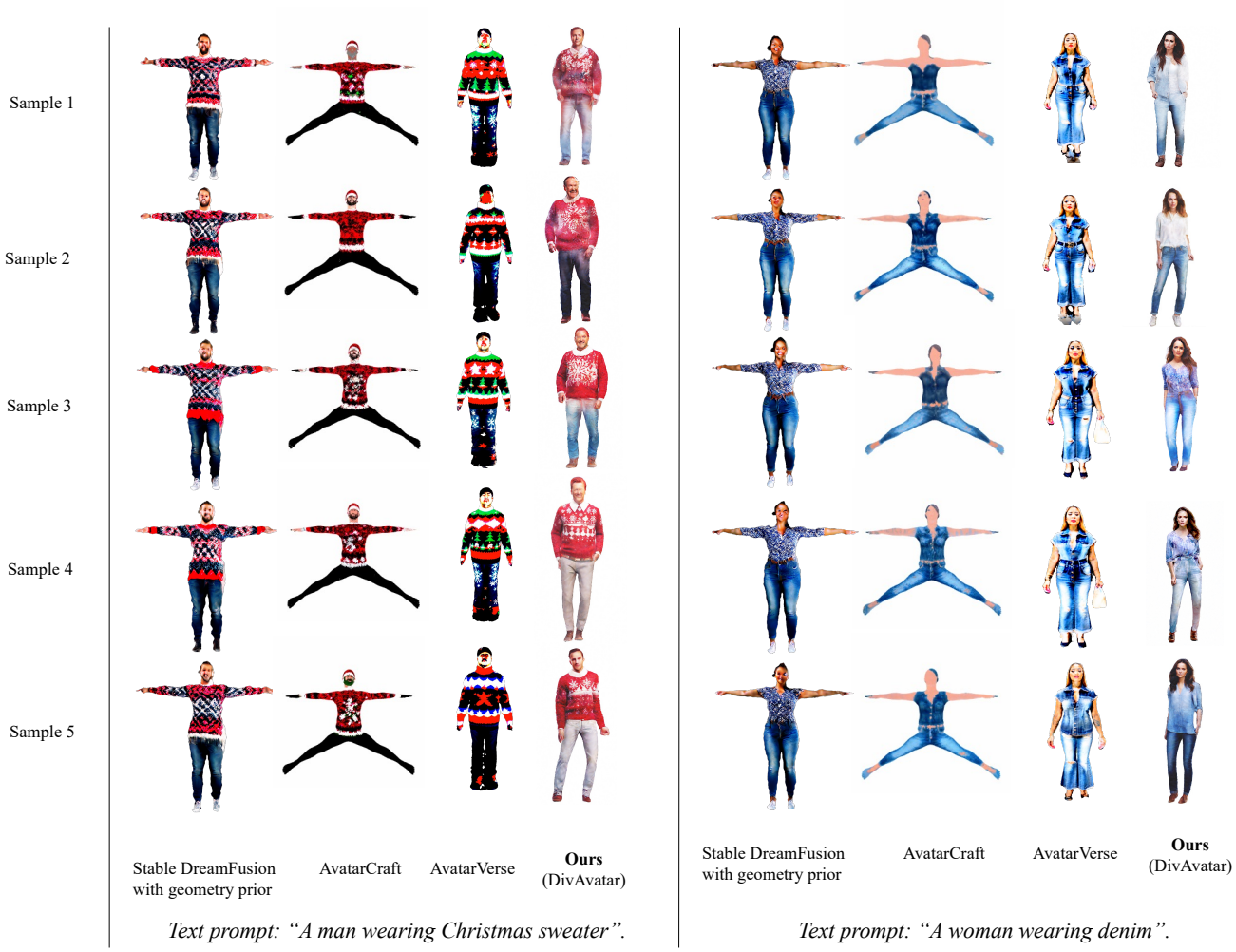

Figure 3. Comparisons on text prompt: 'A man wearing Christmas sweater' and 'A woman wearing denim'.

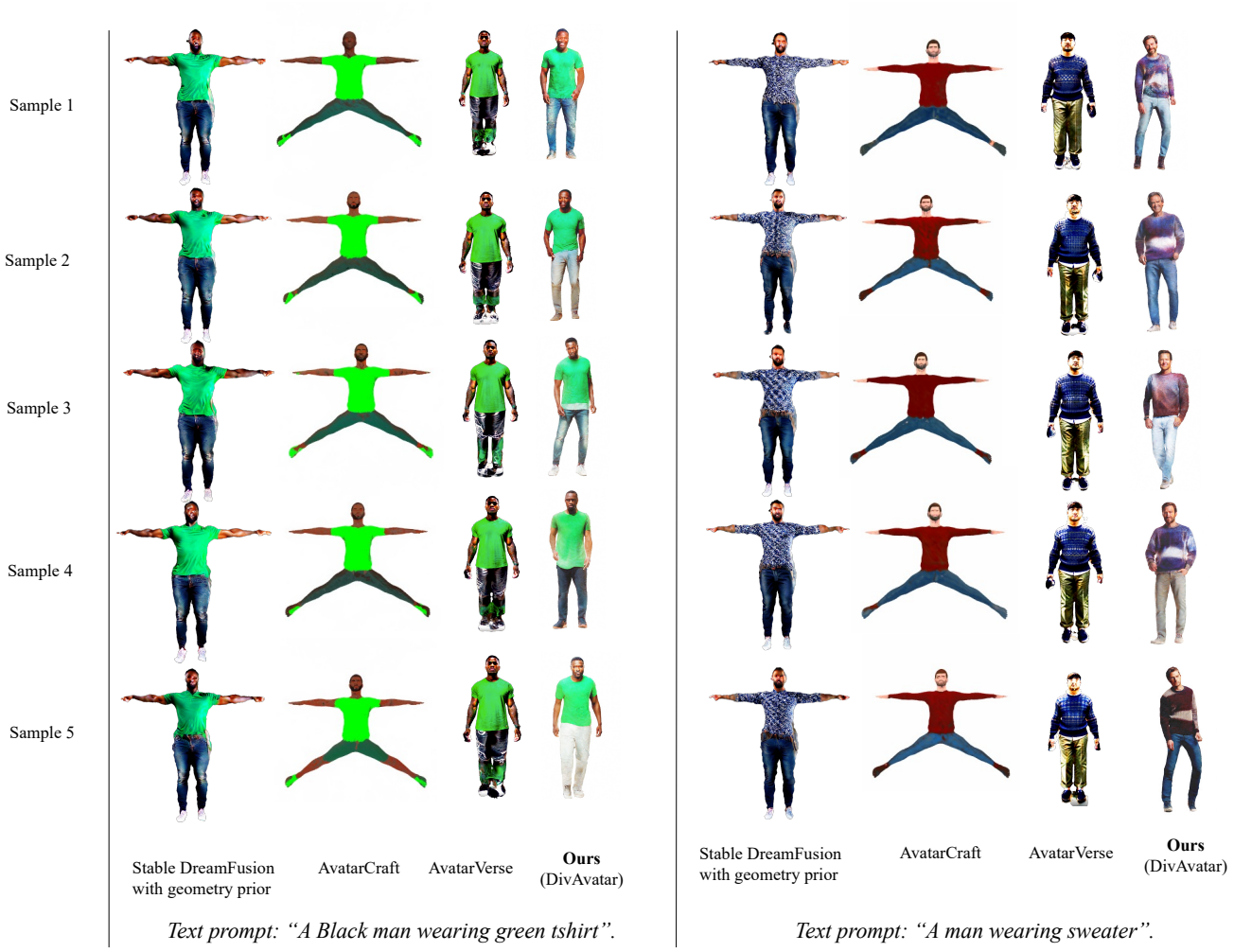

Figure 4. Comparisons on text prompt: 'A Black man wearing green tshirt' and 'A man wearing sweater'.

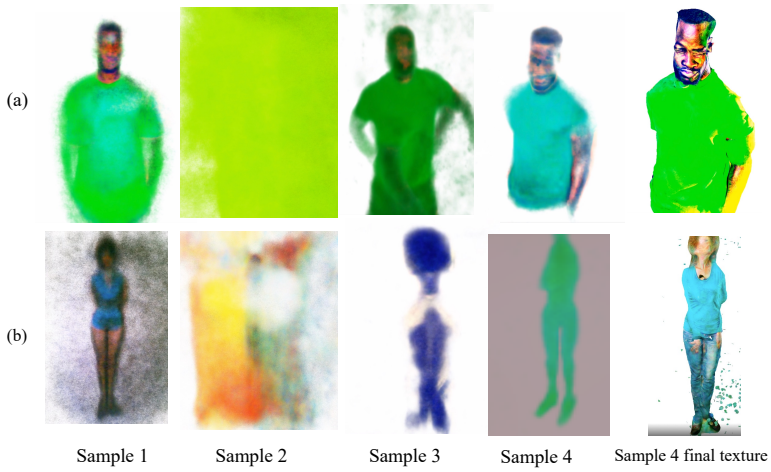

Figure 5. Results from Prolific Dreamer on selected text prompts. Text prompt of (a): 'A Black man wearing green tshirt'. Text prompt of (b): 'A woman'. The method can only generate up to four diverse samples in the coarse stage, and refinement is limited to one sample.

References

[1] Stable diffusion xl 1.0 image to image pipeline cpu. https://huggingface.co/spaces/Manjushri/SDXL-1.0-Img2Img-CPU. Accessed: 2023-11-15. 1

[2] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion. 1

[3] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 1, 2