Supplementary Document for PFAvatar: Pose-Fusion 3D Personalized Avatar Reconstruction from Real-World Outfit-of-the-Day Photos

Implementation Details

Training of Pose-Aware Diffusion Models. The foundational stable diffusion model utilized in our method is Stable Diffusion V1.5 (Rombach et al. 2021). We fine-tuned the pose-aware diffusion model using the PyTorch framework in conjunction with Diffusers (von Platen et al. 2022). Our OOTD training dataset, consisting of 5–20 images per subject, was used for fine-tuning over 600 iterations. The learning rate was established at 1×10^{-6}, and the batch size was set to 1. Upon completion of the fine-tuning, we achieved a pose-aware diffusion model. All training and inference procedures were executed on a singular NVIDIA A6000 GPU. The ControlBooth stage requires approximately 5 minutes for fine-tuning, with the total training process across both stages taking around 50 minutes.

Text-Augmentation Optimization. We augmented the text utilized in our approach to achieve a more authentic quality. The view-prompts were categorized into "front view, left side view, right side view, back view, overhead view, and bottom view." Furthermore, the human form was divided into four segments: body, head, hand, and foot.

Progressive Optimization Details. During the BoothAvatar phase, we employ a progressive Multi-resolution sampling strategy for efficient optimization, whereby the rendering resolution escalates from 64×64 to 512×512 as iterations advance. In the training stages, we randomly sample the timestep from a uniform distribution within the interval of $[0.02, 0.98]$, while the classifier-free guidance scale is established at 50.0. Furthermore, the weight term $w(t)$ for the 3D-SDS loss is set to 1.0.

Camera Settings. For each iteration, the camera view is randomly sampled in spherical coordinates, with the radius, azimuth angle, elevation angle, and field of view (FoV) drawn from the ranges of [1.0, 2.0], [0, 360], [60, 120], and [40, 70], respectively. Over the course of 10,000 iterations, we assigned a sampling probability of 0.7 to the entirety of the body, while designating 0.12 for each of the head, hand, and foot segments.

GPT-4V Queried Prompt For ControlBooth Stage. PuzzleAvatar(Xiu et al. 2024) provides an excellent paradigm for prompts. To further investigate GPT's capability for interpreting images, we have implemented several modifications: (a) we requested an analysis of the character's current viewpoint, (b) a description of their coloration to enhance appearance control, and (c) the inclusion of descriptions for anime characters. Below is the specific prompt have formulated: *Analyzes the provided images, each representing a person or an anime character. Analyzes only the features of the specified parts. Identifies and describes the individual's gender, facial features (including hair), hairstyle, and specific clothing items such as shirts, hats, pants, shoes, dresses, skirts, scarves, etc. Returns the results in a dictionary format with keys being "gender", "face", "hairstyle", and each clothing type. The value should provide 1-3 adjectives or nouns describing the topological or geometric features of the hair, such as length (e.g. short, long, medium, super short, knee-length, floor-length, ankle-length, hip-length, calf-length, etc.), shape (e.g. oval, round, square, heart-shaped, diamond-shaped, rectangular, shaggy, razor-cut, messy, layered, unkempt, etc.), tightness (e.g. skinny, comfortable, fitted, tight, loose, skinny, close-fitting, etc.), style (e.g. modern, casual, sporty, classic, formal, vintage, bohemian, edgy, etc.), or type of hairstyle (e.g. long, short, wavy, straight, curly, bald, medium-length, ponytail, bun, braid, beard, sideburns, dreadlocks, goatee, etc.), referencing color or texture pattern. Include accessories, but do not include any clothing in another description. Omit any keys where clothing does not appear or the description is empty. Also gives the current direction, including 6 cases, "front", "left", "back", "right", "top", "bottom". The response should be just a dictionary, without any other sentences, explanations or markdown symbols (such as { }). For example: { "gender": "Male", "face": "oval", "hairstyle": "green short braid", "shirt": "black short tight crew neck", "pants": "red fitted pants", "shoes": "blue sneakers", "jacket": "hoodie", "view": "front" } If it is an anime character, add the key "anime character": true.*

Additional Analysis and Results

Data Quality Analysis. We conduct an ablation study to evaluate the robustness of our method's two stages—ControlBooth (Stage 1) and BoothAvatar (Stage 2)—under varying data quality conditions. The results are presented in Table 1 and Figure 1, demonstrating the effectiveness and stability of our approach under diverse

Method	CLIP-I↑		DINO↑		CLIP-T↑	
	stage1	stage2	stage1	stage2	stage1	stage2
PFAvatar(Ours)	0.8984	0.8382	0.7416	0.6667	0.3348	0.3168
lack full-body	0.9068	0.8443	0.7441	0.6637	0.3271	0.3230
inaccurate pose	0.8918	0.8163	0.7095	0.6612	0.3259	0.3145
50% train data	0.8756	0.8211	0.7351	0.6525	0.3231	0.3097
10% train data	0.8601	0.8093	0.7316	0.6516	0.2896	0.3023

Table 1: Quantitative result of ablation study on our model's robustness across varying data quality conditions.

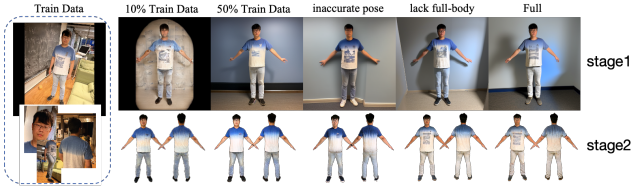

Figure 1: Qualitative result of ablation study on our model's robustness across varying data quality conditions.

data quality scenarios. First, the absence of full-body images in the album dataset presents a considerable challenge for reconstructing a complete full-body model. However, experimental results indicate that our two-stage approach remains effective, yielding promising outcomes even in the absence of such images. Additionally, we conduct an ablation study in which certain images are randomly swapped with others depicting different poses, thereby introducing inaccuracies in pose matching within the training dataset. The results demonstrate that our method continues to generate reasonable outputs, primarily due to the CPPL design. Finally, we observe only a marginal decline in performance when the training set is extremely limited (e.g., 10% of the dataset), further underscoring the robustness of our approach.

User Study. To assess the perceptual quality of our method against existing state-of-the-art approaches, we conducted a comprehensive user study involving 25 participants. Each participant was presented with 20 randomly selected examples and asked to evaluate the results across four key criteria: (1) **3D Consistency**, (2) **Subject Fidelity**, (3) **Prompt Fidelity**, and (4) **Face Fidelity**.

- **3D Consistency:** Assesses the geometric coherence of the reconstructed avatar under 360-degree rotations, ensuring structural integrity from all viewpoints.
- **Subject Fidelity:** Evaluates how well the generated avatar preserves the identity and appearance of the subject as depicted in the reference image.
- **Prompt Fidelity:** Measures the degree to which the generated 3D geometry aligns with the semantic content of the input prompt.
- **Face Fidelity:** Focuses on the visual realism and detail of the facial region, particularly in close-up views.

Method	(1)	(2)	(3)	(4)
PFAvatar (ours)	68.9%	68.2%	66.1%	64.9%
PuzzleAvatar	20.0%	18.4%	21.0%	21.1%
AvatarBooth	11.1%	13.4%	12.9%	14.0%

Table 2: **User Study.** Users show a marked preference for our PFAvatar over PuzzleAvatar and AvatarBooth regarding (1) 3D consistency, (2) subject fidelity, (3) face fidelity, and (4) prompt fidelity.

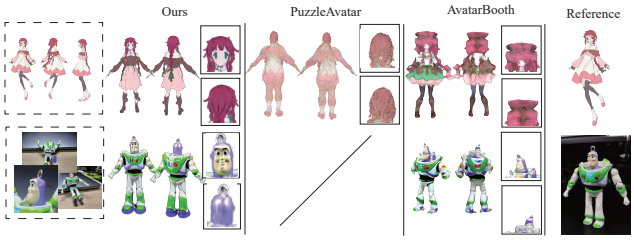

Figure 2: **More Qualitative Comparison: Anime Dataset.** PFAvatar(Ours) method not only accommodates real-person data but also extends its applicability to processing Anime OOTD data.

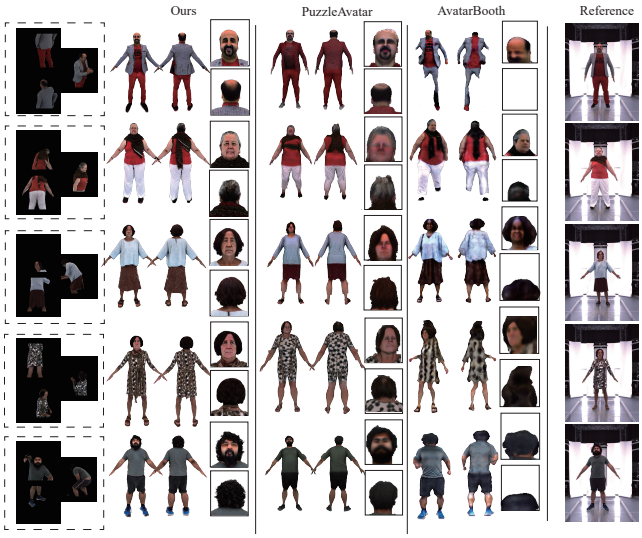

Figure 3: **More Qualitative Comparison: PuzzleIOI Dataset.** We compare our method with PuzzleAvatar and AvatarBooth for appearance-customized reconstruction on the PuzzleIOI Dataset. Our method consistently achieves superior reconstruction quality and more accurate subject fidelity compared to all other approaches.

Participants were instructed to select the method they perceived as best in each category for every example. As summarized in Table 2, the results show a clear preference for **PFAvatar** over baseline methods across all dimensions, particularly in 3D consistency, subject identity preservation, prompt alignment, and facial realism. These findings highlight the overall superiority of our method in both structural and perceptual quality.

Figure 4: **Faliure Case.** For avatars with complex clothing and poses, relying solely on the SDS method may lead to the generation of hallucinations.

Qualitative Result in the Anime Dataset. As demonstrated in Figure 2, we conducted a comparative analysis of our method against PuzzleAvatar and AvatarBooth for appearance-customized reconstruction within the Anime Dataset. The efficacy of PuzzleAvatar is diminished due to the inherent inseparability of clothing within the Anime Data. Our method exhibits a capacity for generalization, enabling it to process anime data OOTD effectively. The results unequivocally demonstrate that our approach consistently attains superior reconstruction quality and exhibits greater subject fidelity when compared to all other methodologies.

Qualitative Result in PuzzleIOI Dataset. As illustrated in Figure 3, our method further demonstrates its robustness under this benchmark. Notably, PFAvatar (ours), PuzzleAvatar (Xiu et al. 2024), and AvatarBooth (Zeng et al. 2023) utilize only cropped OOTD datasets without a reference image, and we have achieved superior results compared to these methods.

Application.

The avatars generated by our method enable a myriad of downstream applications(Huang et al. 2024), including (a) text-guided editing, (b) 3D animation, (c) expressive animation production, and (d) human video reenactment. See more application demos in our video demo.

Limitations and future Work

Limitations. PFAvatar still has several limitations. Since it relies on 2D generative models for guidance, it inevitably inherits some issues, such as imprecise appearance control, leading to certain hallucinations. As shown in Figure 4, the text is missing on the back of the character. These issues could be addressed by further controlling the personalized model.

Future Work. While PFAvatar has achieved superior results in the OOTD reconstruction task, there remain numerous areas for enhancement. The reliance on NeRF as a 3D representation has led to challenges in deriving high-quality geometry from implicit expressions; thus, the future incorporation of hybrid 3D representations (Gao et al. 2020) or the utilization of StableNormal (Ye et al. 2024) as a geometric prior will foster the generation of high-quality explicit geometry. Furthermore, although we demonstrate clearer representation of facial details, maintaining facial consistency in scenarios where the OOTD dataset presents multiple images with varying expressions necessitates further exploration.

References

Gao, J.; Chen, W.; Xiang, T.; Tsang, C. F.; Jacobson, A.; McGuire, M.; and Fidler, S. 2020. Learning Deformable Tetrahedral Meshes for 3D Reconstruction. arXiv:2011.01437.

Huang, Y.; Wang, J.; Zeng, A.; Zha, Z.-J.; Zhang, L.; and Liu, X. 2024. DreamWaltz-G: Expressive 3D Gaussian Avatars from Skeleton-Guided 2D Diffusion. arXiv:2409.17145.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; Nair, D.; Paul, S.; Berman, W.; Xu, Y.; Liu, S.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Xiu, Y.; Ye, Y.; Liu, Z.; Tzionas, D.; and Black, M. J. 2024. PuzzleAvatar: Assembling 3D Avatars from Personal Albums. arXiv:2405.14869.

Ye, C.; Qiu, L.; Gu, X.; Zuo, Q.; Wu, Y.; Dong, Z.; Bo, L.; Xiu, Y.; and Han, X. 2024. StableNormal: Reducing Diffusion Variance for Stable and Sharp Normal. arXiv:2406.16864.

Zeng, Y.; Lu, Y.; Ji, X.; Yao, Y.; Zhu, H.; and Cao, X. 2023. AvatarBooth: High-Quality and Customizable 3D Human Avatar Generation. arXiv:2306.09864.