CVPR #8804

DiffStyle360: Diffusion-Based 360° Head Stylization via Style Fusion Attention

Supplementary Material

Figure 1. Dataset pairs used in our training.

1. Dataset Pair Examples

Fig. 1 presents an example triplet from the *Joker* style domain, consisting of a content image generated by a base 3D-GAN generator [1], a cross-identity style exemplar, and the corresponding ground-truth stylized output produced by a domain-adapted 3D-GAN model [2]. Note that the style exemplar itself is also generated using the same domain-adapted model. Using a style image from a different identity encourages a clearer separation between identity and style, enabling more effective disentanglement of content and stylistic attributes during training. For our dataset, we generate 150 samples for each of the six style domains, yielding a total of 900 training pairs. A representative subset of these samples is shown in Fig. 2.

2. Implementation Details

2.1. Training Setup

We fine-tune the Style Appearance and View Consistency modules of our model on a synthetic 3D-GAN–generated stylized dataset for 800 iterations. The dataset includes six style categories: Joker, Pixar, Sketch, Statue, Werewolf, and Zombie. We use the AdamW optimizer with a learning rate of 10^{-5} and apply Classifier-Free Guidance (CFG) with a weight of 3.0. The batch size is 16, corresponding to the number of generated views per generation. All experiments are conducted on a single NVIDIA A100 GPU. We use $\tau = 1.05$ as the default key scaling value.

2.2. Baselines

For diffusion-based stylization methods (IP2P and InstantID), we first stylize the input content image using a text-driven style prompt. The resulting stylized image is then passed through DiffPortrait360 [6] to synthesize multi-view stylized outputs.

We train the latent mapper of StyleCLIP [8] using the PanoHead generator. For StyleGAN-NADA [5] and StyleGANFusion [9], we rely on the official implementation of StyleGANFusion and adapt the generator backbone to PanoHead. In the case of StyleGANFusion, we use their EG3D configuration for PanoHead, and for StyleGAN-NADA [5] we incorporate adaptive layer selection as described in the original method. DiffusionGAN3D [7] is implemented directly from the paper, as no official code is available. We preserve each baseline's original hyperparameters, such as denoiser settings, noise schedule, learning rate, and optimizer, unless training instability requires adjustments.

2.3. Quantitative Scores

To evaluate our approach, we generate three reference image distributions using the Stable Diffusion pipeline. The first distribution is created by adding noise at timestep $t = 25$ to the input images and denoising them for 50 steps with the corresponding style prompt using each baseline's diffusion checkpoints. The resulting edited images serve as ground-truth targets for computing FID and CLIP similarity. The second distribution consists of the stylized outputs produced by the 3D head stylization methods, which we compare against the first distribution for FID and against matched pairs for CLIP scores. The third distribution contains the original, unedited images and is used to evaluate identity preservation (ID) and ΔD. For ID preservation, we compute ArcFace-based [4] identity similarity between the stylized outputs and their corresponding input images. For depth consistency, ΔD is measured using depth maps estimated by the DepthAnything model [11]. These metrics are computed by comparing images from the second and third distributions. Fig. 3 shows representative examples from all three distributions.

3. User Study

We conducted a user study with **15 participants**, each of whom completed **18 comparison questions**. Options were randomly permuted to ensure fairness. In each question, the participant was shown:

- the original content image,
- a style exemplar image
- stylized outputs produced by different methods.

Participants were asked to select the image that provides the **best balance of identity preservation and stylization**, i.e., the image that both retains recognizable identity and reflects the target style. We did not ask separate questions for identity and stylization; participants provided a single holistic judgment for each comparison.

The stylization results are provided in gif format.

As shown in Table 1, participants preferred our method in **78.2%** of all comparisons, indicating that our approach

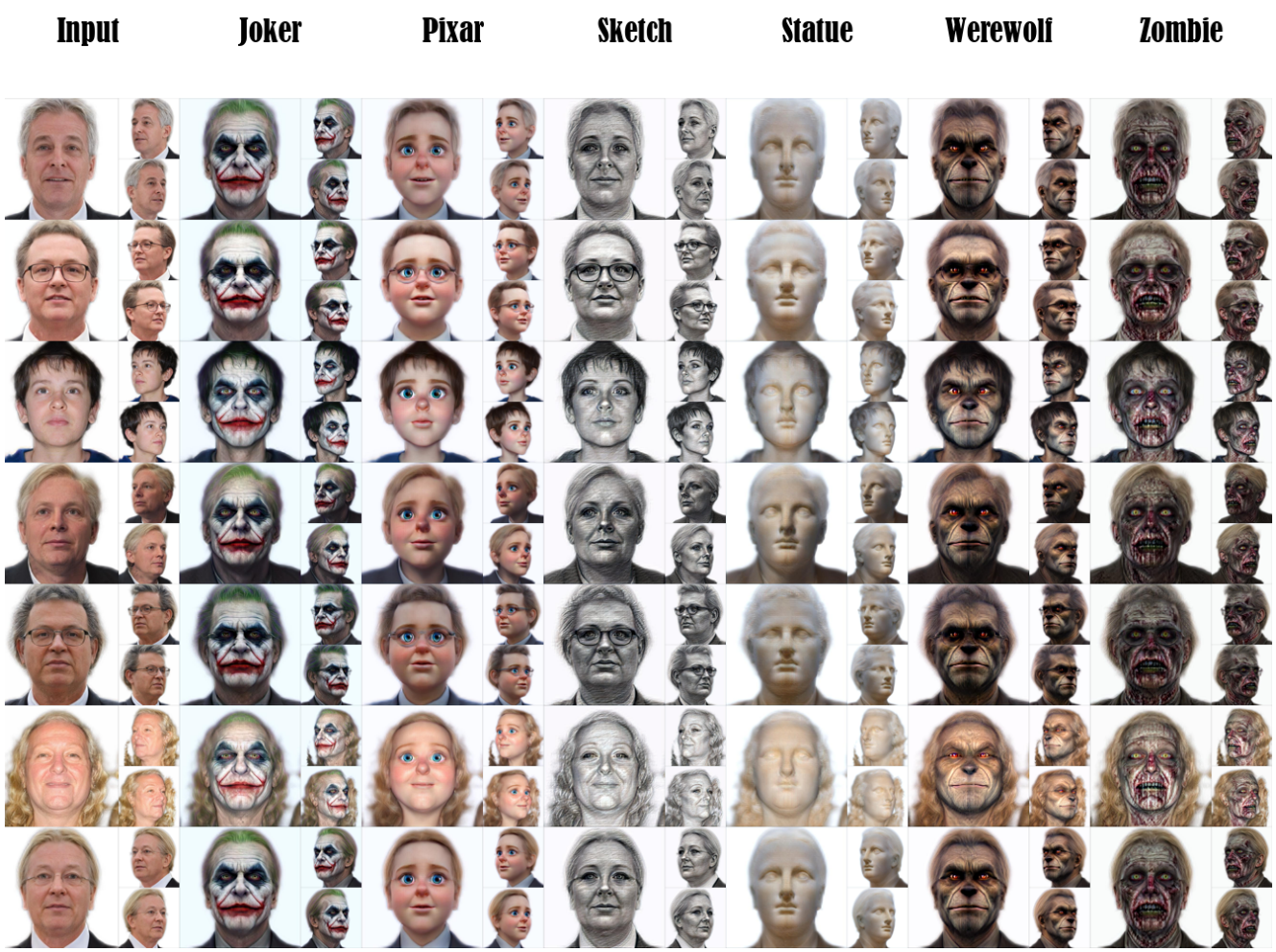


Figure 2. Dataset pair examples.

Method	User Study (%)
IP2P [3]	6.7%
StyleGAN-Fusion [9]	2.4%
Bilecen *et al.* [2]	12.7%
Ours	**78.2%**

Table 1. User study preference on the FFHQ dataset, aggregated over 270 responses (15 participants × 18 questions). Higher values indicate stronger user preference.

achieves the most favorable perceptual trade-off between identity retention and stylistic fidelity.

4. Stylization-Identity Trade-Off Analysis

As shown in Tab. 2 and Fig. 6, the quantitative trends reveal a clear stylization–identity trade-off among baseline methods. InstantID [10] consistently reports high identity similarity scores across all styles; however, these metrics reflect its generally weak stylization capability rather than true identity preservation under meaningful edits. Because InstantID applies only subtle tonal changes, the output remains close to the input image, inflating ID scores while failing to capture the intended style.

IP2P [3] shows more inconsistent behavior across domains. For styles such as *Sketch*, *Statue*, and *Werewolf*, IP2P performs filter-like transformations that preserve most facial structure, leading to high identity scores in Tab. 2. In contrast, for semantically demanding styles such as *Joker* and *Pixar*, IP2P introduces stronger but less controlled changes that reduce identity similarity while still failing to deliver faithful stylization.

These quantitative observations are clearly reflected in the qualitative comparisons in Fig. 6. When integrated into the DiffPortrait360 pipeline for novel-view synthesis, both

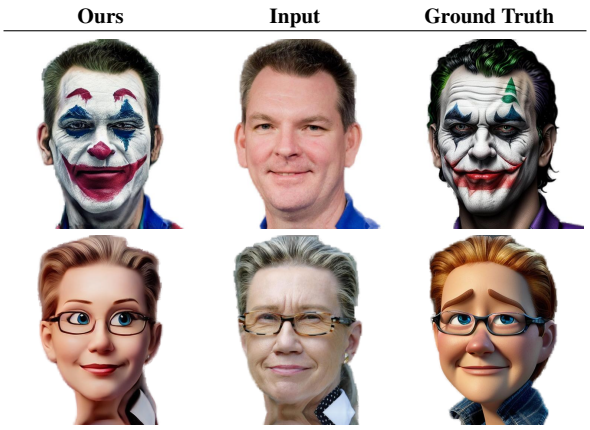


Figure 3. From left to right: a stylized image from the second distribution (outputs of our model), the corresponding unedited input image from the third distribution (used for evaluating identity preservation), and the edited result from the first distribution obtained by using the Stable Diffusion pipeline (used as the ground truth for FID and CLIP evaluations).

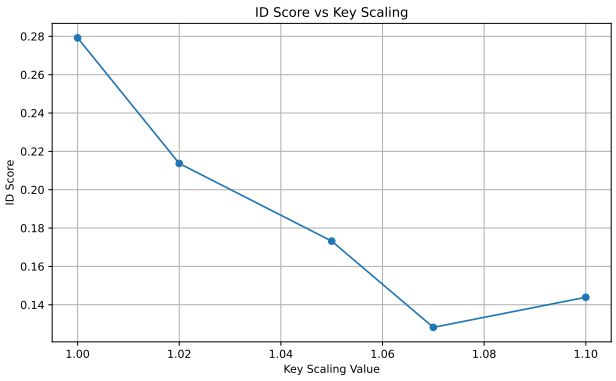
IP2P and InstantID struggle to maintain consistency across viewpoints. Their limited or unstable stylization does not generalize well to unseen views, frequently resulting in degraded back-view quality, texture collapse, and Janus-like artifacts where facial features incorrectly appear on the back of the head. These multi-view failures further indicate that high ID scores in Tab. 2 do not correspond to robust stylization, but rather to insufficient or inconsistent transformations.

In contrast, our approach delivers semantically aligned, domain-consistent stylizations that remain coherent across all viewpoints. As evidenced by both Tab. 2 and Fig. 6, our method achieves a more favorable balance, producing strong stylization while maintaining significantly better identity retention and multiview consistency than prior methods.

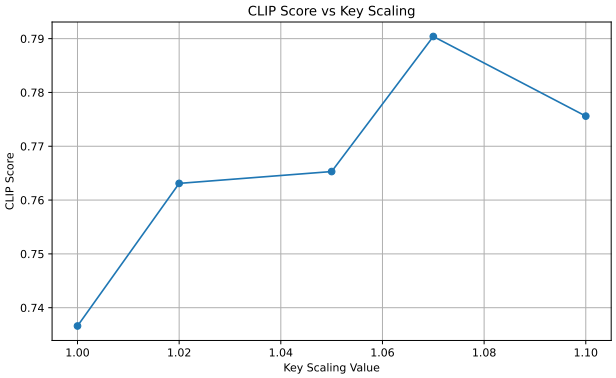
4.1. Effect of Key Scaling on Stylization and Identity Preservation

To analyze the impact of the proposed Key Scaling mechanism, we evaluate identity similarity and CLIP-based style alignment across different scaling values. As shown in Fig. 4a, increasing the scaling factor consistently lowers identity similarity: the ID score drops from 0.28 at $\tau = 1.0$ to 0.13 at $\tau = 1.07$, indicating stronger deviation from the input identity as stylistic cues are amplified in attention. Conversely, Fig. 4b shows that CLIP alignment improves with larger scaling values, peaking at $\tau = 1.07$ before slightly declining at $\tau = 1.10$.

Overall, Key Scaling provides a controllable trade-off between stylization strength and identity preservation. Smaller values retain identity but weaken stylization, whereas larger values yield stronger style fidelity at the cost



(a) Identity score vs. Key Scaling.



(b) CLIP score vs. Key Scaling

Figure 4. Analysis of the effect of Key Scaling on stylization and identity preservation for Statue style on the RenderMe360 dataset. Increasing the scaling factor τ enhances stylistic alignment (higher CLIP) but leads to reduced identity similarity (lower ID), demonstrating a clear stylization–identity trade-off.

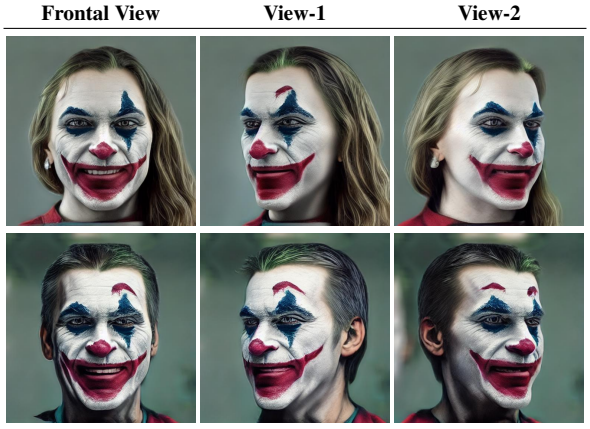


Figure 5. Limitations of our method. In rare cases, certain style elements may not remain fully consistent across views. As shown, the Joker eyebrow appearance exhibits slight inconsistency between angles.

Method	Joker		Pixar		Sketch		Statue		Werewolf	
	CLIP ↑	ID ↑	CLIP ↑	ID ↑	CLIP ↑	ID ↑	CLIP ↑	ID ↑	CLIP ↑	ID ↑
IP2P [3]	0.8832	0.2237	0.8096	0.2046	0.6592	0.7334	0.7693	0.5829	0.7782	0.4554
InstantID [10]	0.6997	0.4662	0.6621	0.4857	0.6994	0.4946	0.7627	0.4706	0.6589	0.4915
StyleCLIP [8]	0.7821	0.1239	0.7481	0.1753	0.7413	0.1845	0.7123	0.2059	0.6634	0.1170
StyleGAN-Fusion [9]	0.8452	0.1105	0.7813	0.2211	0.7652	0.1636	0.8232	0.1385	0.8008	0.1477
StyleGANNADA [5]	0.8035	0.1728	0.7710	0.1873	0.7375	0.1666	0.7302	0.0934	0.7611	0.1329
DiffusionGAN3D [7]	0.7884	0.2298	0.8077	0.1545	0.6745	0.1707	0.8450	0.1552	0.7856	0.2036
Bilecen *et al.* [2]	0.8617	0.1877	0.8211	0.2336	0.7577	0.2752	0.8197	0.1952	0.8444	0.1755
Ours	0.8334	0.3132	0.7868	0.3500	0.7578	0.3428	0.8090	0.3529	0.7865	0.3186
Ours + Key Scaling	0.8649	0.1843	0.8116	0.2598	0.7466	0.1945	0.8182	0.2031	0.8545	0.1611

Table 2. CLIP and identity similarity scores across **Joker**, **Pixar**, **Sketch**, **Statue**, and **Werewolf** style domains on the FFHQ dataset.

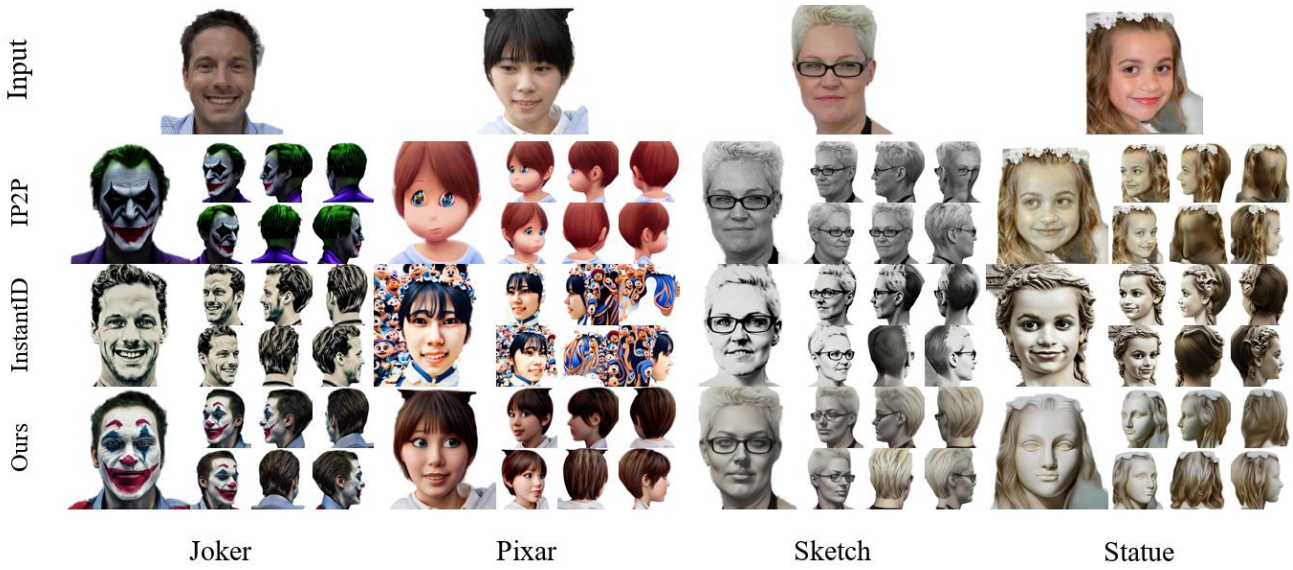


Figure 6. Qualitative illustration of the stylization–identity trade-off and view-consistency behavior across methods for Joker, Pixar, Sketch, and Statue domains. IP2P applies filter-like adjustments for some domains such as Sketch and Statue (leading to inflated ID scores) but introduces uncontrolled changes for others (e.g., Joker and Pixar), causing identity drift. When combined with DiffPortrait360 for novel-view synthesis, both methods frequently fail to maintain consistent 3D structure, showing back-view degradation and Janus-like artifacts. InstantID produces only subtle tonal modifications, resulting in weak stylization but high identity scores as reflected in Tab. 2. Our method achieves strong, semantically aligned stylization with stable identity across all viewpoints and robust multi-view geometry.

of identity drift. We select $\tau = 1.05$ as a balanced setting that offers improved stylization with manageable identity loss.

5. Limitations

Although our method achieves strong multi-view consistency in the vast majority of cases, certain challenging styles can still introduce localized inconsistencies across viewpoints. As illustrated in Fig. 5, the stylized Joker eyebrows appear coherent in some views but exhibit slight deformation or misalignment when rendered from other angles. These artifacts typically arise when the target style contains highly exaggerated or fine-grained facial elements that place strong demands on cross-view geometric alignment.

Importantly, such failures occur only in rare cases and do not reflect the general behavior of our system. Across most styles and identities, our approach maintains stable geometry, consistent textures, and coherent stylization across all novel views.

6. Additional Results

Figs. 7 and 8 present the style reference images for a diverse set of artistic domains and the corresponding outputs produced by our model for multiple input identities.

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360 degrees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20950–20959, 2023. 1

[2] Bahri Batuhan Bilecen, Ahmet Berke Gökmen, Furkan Güzelant, and Ayşegül Dündar. Identity preserving 3d head stylization with multiview score distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. Poster. 1, 2, 4

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. 2, 4

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1

[5] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4), 2022. 1, 4

[6] Yuming Gu, Phong Tran, Yujian Zheng, Hongyi Xu, Heyuan Li, Adilbek Karmanov, and Hao Li. Diffportrait360: Consistent portrait diffusion for 360 view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26263–26273, 2025. 1

[7] Biwen Lei, Kai Yu, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. Diffusiongan3d: Boosting text-guided 3d generation and domain adaptation by combining 3d gans and diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10487–10497, 2024. 1, 4

[8] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. 1, 4

[9] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 1, 2, 4

[10] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 4

[11] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, pages 21875–21911. Curran Associates, Inc., 2024. 1

Figure 7. Additional qualitative results across a wide range of identities and style domains.

Figure 8. Additional qualitative results across a wide range of identities and style domains.