Supplementary Material for
Personalized Face Inpainting with Diffusion Models by Parallel Visual Attention

1. Additional Implementation Details

1.1. Training of PVA Pathway

We trained the PVA modules, the identity encoder, and the embedding of the special token. The attention matrices of the PVA module were first initialized from the text attention matrices $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$ and then trained. We used 1.6×10^{-5} learning rate for the PVA modules and the identity encoder, but 10^{-3} learning rate for the special token. Notice that we used the same token across all identities. If we use a different token for each identity, then the special token for an unseen identity will still be unknown in inference.

We used a two-stage training strategy. First, we trained the PVA modules, the transformer, and the special token while keeping the FaceNet frozen. Second, we trained all these components together. The two stages took 100K iterations each. The reason for doing this was that the transformer and PVA were not trained at first but FaceNet was a pretrained model. Allowing the early-stage gradients to flow into FaceNet might ruin the pretrained model.

To encourage PVA to incorporate a flexible number of reference images, we randomly sampled a subset of reference images in each training step. Specifically, we uniformly sampled from $\{1, 2, 3, 4, 5\}$ as the number of reference images, and randomly sampled from the 5 reference images. Also, with 0.5 probability, we replicate one of the reference images using its horizontally flipped version. This was to align with the settings of finetuning. See Sec. 1.2 for explanations.

All images were augmented with only random horizontal flips. We used random masks from the CelebAHQ-IDI dataset in training. In each batch, we merged the random masks with randomly selected semantic rectangular masks.

The training was parallelized on 4 RTX A4500 GPUs and took around 70 hours.

1.2. Finetuning

Given N_{ref} images of identity p, $\mathcal{R}_p = \{\mathbf{x}_r\}_{i=1}^{N_p^{\text{ref}}}$, we finetuned the PVA and cross attention modules so that the diffusion models could adapt to this identity better. Notice that in finetuning, we needed to use one image for inpainting and other images as reference, resulting in only $N_p^{\text{ref}} - 1$ reference images. In inference, we still used the N_p^{ref} images as reference, which would be an inconsistency between training and inference. Therefore, we chose to "pad" the length of reference images to N_p^{ref} by replicating an image. Specifically, we randomly select one image from the $N_p^{\text{ref}} - 1$ references and reflected it horizontally. When $N_p^{\text{ref}} = 1$, however, we did not have any reference image. In this case, we use the reflected version of the inference image as the reference. As the finetuning was only 40 steps, we found that this did not lead to the trivial solution of copy-pasting the reference directly to the inpainted image.

We used a stratified sampling approach to reduce the variance of the gradients. Specifically, we replicated the same batch of images m times, each with a different time, $\{t_i\}_{i=1}^m$. The conventional sampling approach is to sample each t_i from the same uniform distribution $\mathcal{U}[0, 1]$. In stratified sampling, we sample t_i from $\mathcal{U}[\frac{i-1}{m}, \frac{i}{m}]$. The diffusion models at different time steps will have drastically different behavior, *e.g.*, imagining new structures when t is large and refining the details when t is small. The stratified sampling technique ensured that different time steps could be covered evenly with a small number of batch sizes, *e.g.*, 4.

1.3. Reproduction of Baselines

We used the pretrained model of SDI and Paint by Example as-is. The SDI model was obtained from the tag "stabilityai/stable-diffusion-2-inpainting" in diffusers. The Paint by Example model was obtained from their official release[1].

Textual Inversion, Custom Diffusion, and MyStyle were fine-tuned on each identity of the test set separately. Both Textual Inversion and Custom Diffusion used a batch size of 8 and AdamW with 10^{-2} weight decay. Textual Inversion was trained for 5K iterations using an effective learning rate of 10^{-2}. For Custom Diffusion, we trained the cross-attention modules with 8×10^{-6} learning rate for 1K iterations. MyStyle was trained for 1K iterations using Adam [2] with a learning rate of 3×10^{-3}. In inference, MyStyle projected the image to be inpainted onto the latent space of the finetuned model, which took around 1 minute for each image.

Textual Inversion and Custom Diffusion were trained on 4 RTX A4000 GPUs, which took around 1.2 hours and 1 hour for each identity. MyStyle was trained on a single RTX A4000 GPU and took around 15 minutes.

1.4. Inference

We used a slightly different setting in the classifier-free guidance. The conventional setting of classifier-free guidance used "photo of a person" as the positive condition and

[1] https://github.com/Fantasy-Studio/Paint-by-Example

Task	Methods	Positive Condition	Negative Condition
Inpainting-Only	PVA	Photo of a person & $\mathbf{E}_I(\{\mathbf{x}_r\})$	Photo of a person
	Default	Photo of a person	$\varnothing$
Controlled-Inpainting	PVA	Photo of a person, smiling & $\mathbf{E}_I(\{\mathbf{x}_r\})$	Photo of a person & $\mathbf{E}_I(\{\mathbf{x}_r\})$
	Default	Photo of a person, smiling	$\varnothing$

Supplementary Table 1. Comparisons between PVA and the default method on the conditions used in classifier-free guidance.

Type	Name	Prompt
Expression	Laughing	Photo of a person, laughing
	Serious	Photo of a person, serious
	Smile	Photo of a person, smiling
	Sad	Photo of a person, looking sad
	Angry	Photo of a person, angry
	Surprised	Photo of a person, surprised
Makeup	Makeup	Photo of a person, with heavy makeup
	Beard	Photo of a person, has beard
	Lipstick	Photo of a person, wearing lipstick
Action	Funny	Photo of a person, making a funny face
	Tongue	Photo of a person, putting the tongue out
	Singing	Photo of a person, singing with a microphone
	Cigarette	Photo of a person, smoking, has a cigarette
Accessory	Eyeglass	Photo of a person, wearing eyeglasses
	Sunglasses	Photo of a person, wearing sunglasses

Supplementary Table 2. The full list of prompts used in the language-controllable inpainting experiment.

$\varnothing$ as the negative condition. We also used this setting in Textual Inversion and Custom Diffusion.

However, the PVA was different in that the condition had extra visual features, and the identity-related information was mostly contained in the visual component. In light of this, we could keep the text features the same and contrast the visual features. In the inpainting-only task, we used the "photo of a person" with visual features as the positive condition and used the prompt without the visual features as negative ones. In the language-controlled inpainting task, we used the controlling prompt with visual features as the positive condition and the neutral prompt with visual features as the negative condition. The differences are summarized in Supplementary Table 1.

1.5. Evaluation

We evaluated the inpainting performance on four different types of semantic regions, lower face, eye & brow, whole face, and random. As different semantic regions might have different characteristics, we calculated the metrics for every region separately and averaged the results across all four regions. The results per region are described in Sec. 3.

The full list of prompts used in the language-based controlling experiment is listed in Supplementary Table 2.

2. Construction Pipeline of CelebAHQ-IDI

Preprocessing. We first checked for duplicate images, including horizontally reflected duplicates. We filtered out 403 duplicate images in total, which consisted of 199 pairs and 5 triplets. Then we detected the facial landmarks using dlib [1]. Two images that failed in detection were also discarded.

Mask generation. We constructed rectangular masks that covered several semantic regions of the face, including "eye and brow", "lower face", "whole face", *etc*. Each mask was the bounding box of the landmarks of the corresponding semantic region and was diluted 20% in both width and height. We also generated random masks following the protocol of LaMa [3] and merged them with rectangular masks. Specifically, we used the "configs/data_gen/random_thick_512.yaml" configurations in the LaMa [3] code base[2] for generating the random masks. We sampled 30K masks and stored them and directly sampled from these 30K masks as random masks in training.

Dataset split. We filtered all identities with images less than or equal to the reference number. For each remaining identity, we randomly chose reference images and left the rest as inference images. Finally, we randomly split the dataset into training, validation, and testing with ratios of 0.6, 0.1, and 0.3 based on identities.

3. Per Region Evaluation Results

The identity similarity, FID, and KID per region for all methods are presented in Supplementary Tables 3, 4, and 5. We used the "Mean" results in the paper. We observed that the eye & brow region is the easiest region for inpainting and the whole face region is the hardest region.

References

[1] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 2

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[3] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky.

[2] https://github.com/advimman/lama

Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2

Method	Lower Face	Eye & Brow	Whole Face	Random	Mean
SDI	0.444 ± 0.105	0.613 ± 0.086	0.094 ± 0.088	0.283 ± 0.210	0.359
PbE	0.500 ± 0.098	0.638 ± 0.084	0.217 ± 0.100	0.363 ± 0.180	0.430
MyStyle	0.696 ± 0.134	0.786 ± 0.100	0.639 ± 0.133	0.661 ± 0.146	0.696
TI-1	0.639 ± 0.091	0.759 ± 0.066	0.401 ± 0.132	0.528 ± 0.176	0.582
TI-6	0.686 ± 0.098	0.789 ± 0.072	0.504 ± 0.153	0.597 ± 0.171	0.644
CD-1	0.735 ± 0.089	0.823 ± 0.062	0.580 ± 0.131	0.659 ± 0.146	0.700
CD-6	0.757 ± 0.094	0.832 ± 0.068	0.635 ± 0.136	0.694 ± 0.141	0.729
PVA-1	0.634 ± 0.087	0.776 ± 0.064	0.418 ± 0.115	0.532 ± 0.165	0.590
PVA-2	0.670 ± 0.084	0.797 ± 0.063	0.494 ± 0.114	0.585 ± 0.150	0.637
PVA-4	0.671 ± 0.084	0.796 ± 0.064	0.505 ± 0.112	0.587 ± 0.144	0.640
PVA-6	0.657 ± 0.091	0.788 ± 0.074	0.496 ± 0.116	0.586 ± 0.144	0.632
PVA-FT-1	0.772 ± 0.082	0.856 ± 0.057	0.668 ± 0.116	0.716 ± 0.122	0.753
PVA-FT-6	0.789 ± 0.096	0.858 ± 0.064	0.707 ± 0.127	0.740 ± 0.120	0.773

Supplementary Table 3. Comparisons of identity similarity per masked region on CelebAHQ-IDI-5 dataset. Numbers after "±" indicate the standard deviation. The "-1" and "-6" denote the classifier-free guidance strength. "FT" denotes finetuning on each identity for 40 iterations.

Method	Lower Face	Eye & Brow	Whole Face	Random	Mean
SDI2	7.039	4.244	12.301	9.383	8.242
PbE	10.092	5.866	15.081	13.682	11.180
MyStyle	29.221	8.993	34.754	37.755	27.681
CD-1	6.041	3.709	7.262	7.288	6.075
CD-6	9.540	4.569	12.969	11.829	9.727
TI-1	6.770	3.746	9.335	8.821	7.168
TI-6	19.370	5.403	29.811	23.759	19.586
PVA-1	8.613	6.296	9.615	9.766	8.572
PVA-2	10.501	7.179	11.923	11.792	10.349
PVA-4	10.289	6.968	12.129	12.243	10.407
PVA-6	25.377	14.185	30.156	30.342	25.015
PVA-1	8.323	6.061	9.243	9.240	8.217
PVA-6	20.18	12.3	23.2	23.65	19.8

Supplementary Table 4. FID per masked region for all methods on CelebAHQ-IDI-5 dataset. For the notations in the table, please refer to the Supplementary Table 3.

Method	Lower Face	Eye	Whole Face	Random	Mean
SDI	1.782	1.793	4.486	2.808	2.717
PbE	4.750	2.901	9.220	7.486	6.089
MyStyle	4.620	1.169	6.351	7.977	5.029
CD-1	5.188	5.629	5.299	5.637	5.438
CD-6	5.279	5.454	6.120	6.628	5.870
TI-1	5.071	4.775	5.587	5.151	5.146
TI-6	6.735	5.399	11.487	9.993	8.404
PVA-1	4.433	4.090	4.314	4.637	4.369
PVA-2	5.415	4.784	5.873	6.224	5.574
PVA-4	5.729	4.883	6.737	7.273	6.155
PVA-6	8.551	7.122	10.295	10.634	9.151
PVA-FT40-1	4.359	4.008	4.368	4.421	4.289
PVA-FT40-6	4.600	4.929	5.282	4.874	4.921

Supplementary Table 5. KID ($\times 10^{-3}$) per masked region for all methods on CelebAHQ-IDI-5 dataset. For the notations in the table, please refer to the Supplementary Table 3.