FEEDFACE: EFFICIENT INFERENCE-BASED FACE PERSONALIZATION VIA DIFFUSION MODELS

Chendong Xiang[1,2] *, **Armando Fortes**[1] *, **Khang Hui Chua**[1], **Hang Su**[1,3], **Jun Zhu**[1,2,3]
[1]Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center, THBI Lab
[1]Tsinghua-Bosch Joint ML Center, Tsinghua University, Beijing, China
[2]ShengShu, Beijing, China [3]Pazhou Laboratory (Huangpu), Guangzhou, China
{xcd19,fmq22,ckh20}@mails.tsinghua.edu.cn
{suhangss,dcszj}@tsinghua.edu.cn

ABSTRACT

We introduce *FeedFace*, a novel inference-based method designed to augment text-to-image diffusion models with face-based conditional generation. Trained on a thoroughly curated and annotated dataset of diverse human faces, FeedFace operates without additional training for new facial conditions during generation. Our method can create images that are not only true to the textual descriptions but also exhibit remarkable facial faithfulness in seconds. Our model supports using multiple faces as input conditions, leveraging extra facial information to improve facial consistency. A key strength of our method lies in its efficiency. Through our experiments, we demonstrate that FeedFace can produce face-conditioned samples with comparable quality to leading industry methods, using only **0.4%** of their data volume and fewer than **5%** of the samples seen by these methods during training.

1 INTRODUCTION

Advancements in text-to-image (T2I) generation, particularly diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020), have led to the creation of high-quality and diverse images (Rombach et al., 2021; Ramesh et al., 2022; Bao et al., 2023). However, the challenge persists in generating images with highly consistent facial features. Traditional approaches predominantly rely on optimization-based methods with a limited dataset for training, enabling generation that aligns with the samples (Ruiz et al., 2022; Gal et al., 2022). Yet, this approach requires individual training and storage of parameters, resulting in substantial computational and memory costs.

In our research, we have developed a method that facilitates the generation of highly consistent human images without additional training for each new subject. Remarkably, our pre-training process is considerably more efficient compared to existing industry standards (e.g. Face0 (Valevski et al., 2023)), requiring only 41K data samples and 4.8M training steps. This represents a significant reduction in training costs, paving the way for more efficient and scalable solutions in the field.

2 METHOD

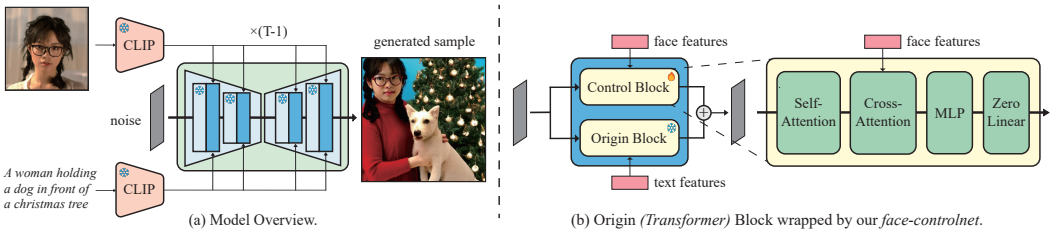
To seamlessly integrate new conditions in T2I models with efficient training, our approach focuses on two key aspects: model architectural design and strategic training framework formulation.

2.1 ARCHITECTURE DESIGN

2.1.1 FACE IMAGE ENCODER

We use the pre-trained CLIP image encoder[1] as the face feature extractor. CLIP (Radford et al., 2021), a multi-modal model trained on a vast array of image-text pairs, is adept at extracting rich

*Equal contribution.
[1]https://huggingface.co/openai/clip-vit-base-patch32



(a) Model Overview. (b) Origin *(Transformer)* Block wrapped by our *face-controlnet*.

Figure 1: Our *face-controlnet* architecture design.

semantic content due to its well-aligned image and text feature spaces. This is complemented by a simple, trainable linear layer to align the face features with the model's latent space.

2.1.2 FACE-CONTROLNET

We propose a *face-controlnet*, designed to enable our model to integrate additional conditional inputs. Illustrated in Figure 1, our architecture combines a *control block* and an *origin block*, both using the same image latent as input. The control block comprises a self-attention block, a cross-attention block, an MLP, and a linear layer initialized at zero. The cross-attention block is specifically used for processing facial condition information, while the zero-initialized linear layer ensures the initial model predictions remain consistent with the original model for a gradual optimization process and prevention of training disruptions. The origin block represents the standard transformer blocks in the U-Net/U-ViT and stays frozen during training. The outputs of these blocks are added together to produce the final result.

2.2 TRAINING STRATEGY DESIGN

We solely optimize the control blocks and the linear layers added to the CLIP image encoder. To aid the model learn the relevant parts of the image condition, we employ a regularization mask loss:

$$loss = \left\| mask \cdot (\epsilon - \epsilon_\theta(x_t, t, c)) + (1 - mask) \cdot (\epsilon_{\hat{\theta}} - \epsilon_\theta(x_t, t, c)) \right\|_2^2,$$

where $mask$ represents the facial mask, ϵ the ground truth noise, $\epsilon_{\hat{\theta}}$ the noise predicted by the original model, and ϵ_θ the noise in the face-conditioned model predictions. Details on training and inference can be found in Appendix B, including the hyperparameters used in our experiments.

3 EXPERIMENTS & RESULTS

Our training uses a refined collection of 41K samples (0.4% of the 10M used by Face0) from FFHQ (Karras et al., 2018), described in detail in Appendix A. During train, our model uses a total of 4.8M image-text pairs for optimization (3.75% of the 128M used by Face0).

Methods	Face Align. ↑		Text Align. ↑		Overall ↑	Time ↓
	Insightface	CLIP	ImageReward	CLIP		
DreamBooth	0.176	0.718	**0.878**	**0.316**	2.09	72min
FeedFace (Ours)	**0.293**	**0.818**	0.760	0.298	**2.17**	**22min**

Table 1: Results against DreamBooth (Ruiz et al., 2022) on face and text alignment, and time usage.

We build a benchmark with 36 identity cases. Specifics about metric design and computation are detailed in Appendix C. Summarized in Table 1, our results demonstrate the efficiency and effectiveness of our approach against DreamBooth (Ruiz et al., 2022), a widely recognized method. Generation samples and further results are available in Appendix E.

4 CONCLUSION

Our work successfully extends the capabilities of T2I diffusion models to support face-conditioned generation, demonstrating the efficacy and efficiency of our method. Discussion on limitations and future work is available in Appendix D

URM STATEMENT

All authors of this work meet the URM criteria of the ICLR 2024 Tiny Papers Track.

REFERENCES

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shiliang Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, 2023.

Jiankang Deng, J. Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2018.

Jiankang Deng, J. Guo, Evangelos Ververas, Irene Kotsia, Stefanos Zafeiriou, and InsightFace Face-Soft. Retinaface: Single-shot multi-level face localisation in the wild. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5202–5211, 2020.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ArXiv*, abs/2208.01618, 2022.

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2018.

Jie Liang, Huiyu Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 653–661, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2022.

Jascha Narain Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585, 2015.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020.

Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. *ArXiv*, abs/2306.06638, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *ArXiv*, abs/2304.05977, 2023.

(a) Standard diffusion loss (b) Mask diffusion loss (c) Regularization mask loss

Figure 2: Visualization of different loss function strategies with a mask. Higher brightness means the constraint of that given part of the image is closer to ground truth. Accordingly, the black parts mean there is no constraint.

A Data Processing Details

We use the FFHQ dataset (Karras et al., 2018), which includes 70K in-the-wild images, as our primary training dataset. The processing of these images included several critical steps to ensure the quality and relevance of the data for training:

- **Image Filtering:** Removal of images containing multiple or unrecognizable faces.
- **Image Captioning:** Utilization of the LLaVA visual-language model (Liu et al., 2023) for generating descriptive captions for the images.
- **Face Detection and Cropping:** Implementation of Insightface (Deng et al., 2018) for face detection. We expanded the bounding box around each face by 1.1 times, then cropped it into a square based on the longer side to ensure the entire face was included.
- **Face Mask Generation:** Create high-quality face masks using Facer (Deng et al., 2020).

This process yielded a refined dataset of 41K instances. Each instance comprises four elements: the original image, its corresponding caption for the text encoder, the detected face image for the image encoder, and the face mask for computing the regularization mask loss.

B Training and Inference Detail

B.1 Design of regularization mask loss

We illustrate various loss computation strategies using a mask in Figure 2. The mathematical formulations for each method are as follows:

$$\text{Standard diffusion loss}: \|(\epsilon - \epsilon_\theta(x_t, t, c))\|_2^2,$$

$$\text{Mask diffusion loss}: \|mask \cdot (\epsilon - \epsilon_\theta(x_t, t, c))\|_2^2,$$

$$\text{Regularization mask loss}: \left\|mask \cdot (\epsilon - \epsilon_\theta(x_t, t, c)) + (1 - mask) \cdot (\epsilon_{\hat{\theta}} - \epsilon_\theta(x_t, t, c))\right\|_2^2,$$

where $mask$ represents the facial mask, ϵ the ground truth noise, $\epsilon_{\hat{\theta}}$ the noise predicted by the original model, and ϵ_θ the noise in the face-conditioned model predictions, t is the noise level, c is the condition. The standard diffusion loss applies to all pixels in the image, which could lead the model to fit less important areas, especially when the facial region is small. The mask diffusion loss

Figure 3: Generated images with blurred and low-detail backgrounds due to the mask diffusion loss.

Methods	Training Data Volume ↓	# Samples Seen During Training ↓
Face0	10M	128M
FeedFace (Ours)	**41K**	**4.8M**

Table 2: Results against Face0 (Valevski et al., 2023) on training efficiency.

focuses solely on the facial pixels; however, without constraints on the background, this can result in a sacrificed background quality, often leading to low details or blurring, as shown in Figure 3. The regularization mask loss employs ground truth noise for constraints on the face, enabling the model to learn a more accurate mapping between facial features and generated images. For the background, it uses the prediction noise of the original model for constraints, ensuring that the *face-controlnet* does not impact unimportant areas. This approach aligns with the zero-linear aspect of the *face-controlnet*, allowing the model's initial predictions to match the original model. As a result, the background part loss is initially zero, with the optimization led primarily by the face part loss, thereby facilitating faster convergence.

B.2 Training detail

The overall framework is optimized with AdamW (Loshchilov & Hutter, 2019) with a learning rate of $2e^{-5}$. During training, the face condition feature is set to a zero vector 10% of the time to enable classifier-free guidance on the direction of the face. Additionally, also at a ratio of 10%, we use the standard diffusion loss to reduce the *face-paste* appearance in the generated images.

The training procedure was conducted on 6 NVIDIA A100 GPUs (80GB), with a total batch size of 96 (i.e., with a batch size of 16 per GPU) and for 50k steps. The process took approximately one day and a half to complete, with the model training on 4.8M image-text pairs. Head-to-head comparison with Face0 (Valevski et al., 2023) is presented in Table 2.

B.3 Inference Details

B.3.1 Multiple Image Condition

For batch image conditioning, we first detect faces in each image, then encode these using the CLIP image encoder to obtain a batch of image features. These features are concatenated along the sequence length dimension and then fed into the control block as conditions.

B.3.2 Classifier-free guidance

Classifier-free guidance is implemented with a scale factor of 5 for both face and text directions. For conditional noise prediction, input text and faces serve as conditions, while for unconditional noise prediction, an empty text prompt and a zero vector as the face condition are used. The final noise

prediction equation is:

$$\epsilon = \epsilon_c + scale \times (\epsilon_c - \epsilon_{uc}),$$

where ϵ_c is noise prediction on both text and face, and ϵ_{uc} is noise prediction on empty text and face.

C EVALUATION DETAIL

We use the PPR10K dataset (Liang et al., 2021), which consists of 11K high-quality photos organized into 1,681 identity-consistent groups, to construct our benchmark. After filtering out images with multiple faces, we select 36 identity-consistent groups as reference images for our final evaluation. For each group, 5-8 captions are randomly chosen from our caption buffer (captions for other datasets). This results in 36 test cases, each with 5-8 prompts, amounting to approximately 200 image-text pairs in total.

Our evaluation focuses on three key metrics: face and text alignment, and time usage. Face features are extracted using Insightface (Deng et al., 2018), and the mean cosine similarity between reference and generated faces is calculated to assess face similarity. Textual consistency is evaluated using ImageReward (Xu et al., 2023) to generate a corresponding score. Furthermore, we evaluate our method based on CLIP image and text similarity metrics, commonly used by related works in the field. Here, for the face alignment, we detect and crop the faces into squares based on the longer side to ensure the entire face is included and use the CLIP image embedding cosine similarity between reference and generated faces. Time usage is measured as the total time taken to process all 36 cases.

It is important to note that these scores depend on the specific prompts and identities. The scores for face and textual alignment not based on CLIP were normalized as follows:

$$score_{norm} = \frac{score - score_{min}}{score_{max} - score_{min}}.$$

A preliminary step involved generating 20 images for each image-text pair within the test case using the original model to determine reasonable maximum and minimum scores for the alignment metrics. These, expected to show high text consistency but low face similarity, serve as base images.

The maximum face similarity score was established based on the self-similarity of reference images. In contrast, the respective minimum score was derived from evaluating the similarity between reference images and the base images. Similarly, the maximum textual consistency score was obtained from the scores of base images, and the respective minimum score was calculated using the scores of reference images.

D LIMITATIONS AND FUTURE WORK

Our work successfully demonstrates the efficient generation of visually consistent facial images using face-conditioned T2I diffusion models, however, several challenges persist. In the context of facial generation, nuanced details such as facial expressions and orientation still pose difficulties, many times resulting in a pasting-like artifact. Moreover, despite the proficiency of our model in producing high-quality and consistent facial images, there are noticeable trade-offs in terms of the semantic alignment with the textual descriptions and overall image quality. Addressing these issues not only underscores the current limitations but also points towards potential avenues for future research. Additionally, in the realm of conditional generation, further future developments might involve handling more complex conditions (e.g., simultaneous generation of multiple identities), and extending our work to the open-domain conditional generation task.

E ADDITIONAL RESULTS

In Figures 4 and 5, we present some generated samples from the UniDiffuser-based implementation of our method.

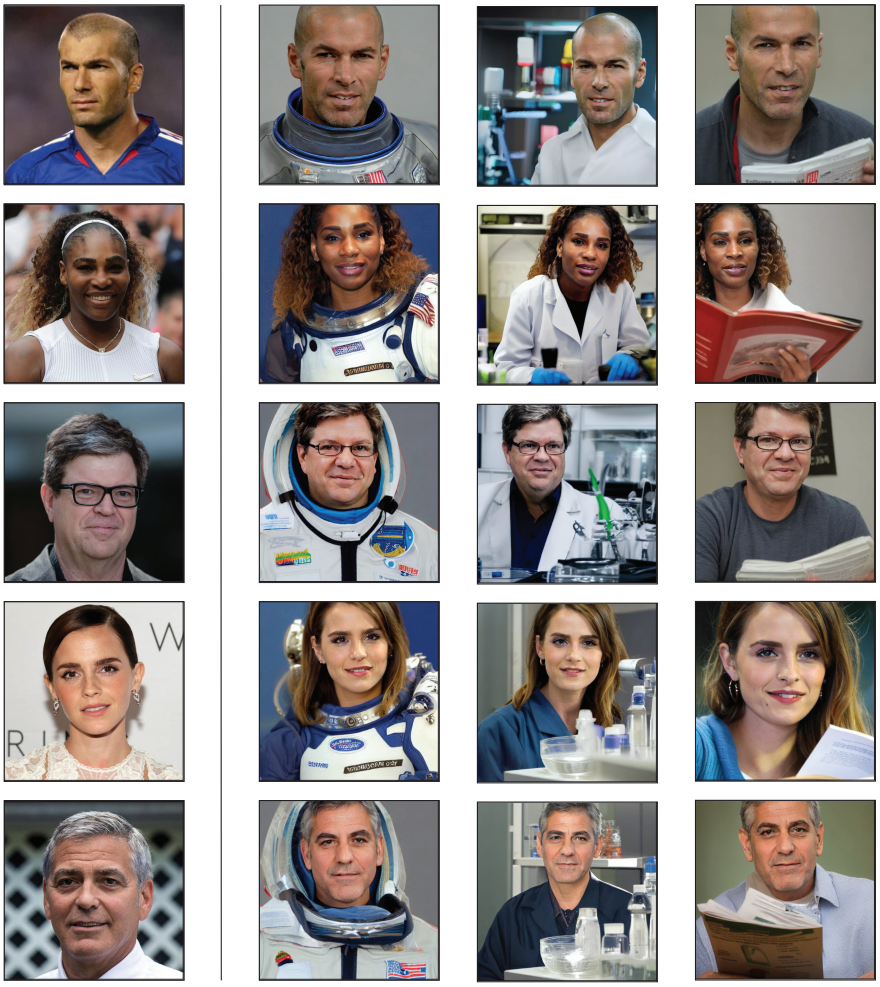


Figure 4: Additional generated samples from our method. Takes several images as input but we only show one image here for simplicity.

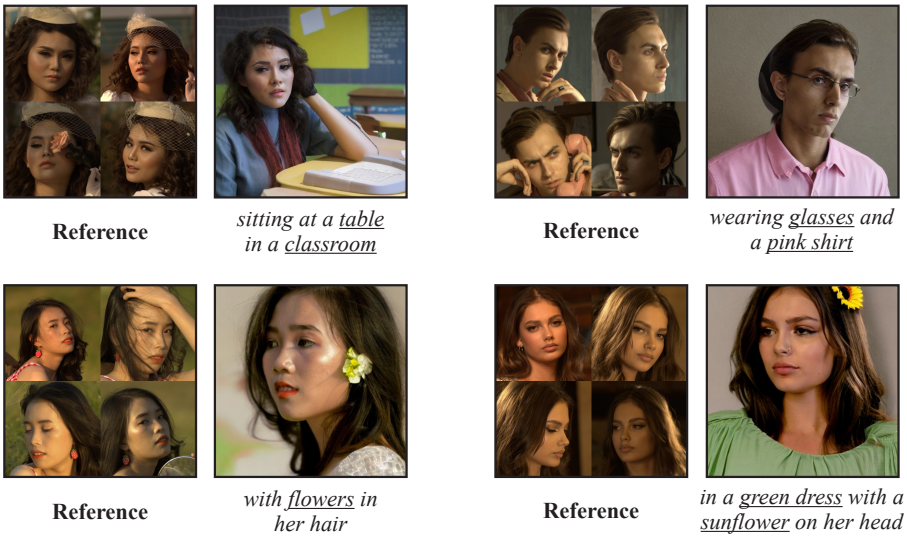


Figure 5: Additional generated samples from our method with emphasis on text alignment. The image references are taken from the PPR10K dataset (Liang et al., 2021).