Supplementary Material
Mixture of Global and Local Experts with Diffusion Transformer for Controllable Face Generation

Anonymous Author(s)
Submission Id: 3972

A Societal Impacts and Responsible AI

Our research focuses on controllable face generation, based on a diffusion transformer architecture combined with a mixture of global and local experts, aiming to support a variety of optimistic application scenarios. This technology is not intended to mislead or deceive. However, similar to other generative models, it may still be misused for impersonating individuals. We strongly oppose any behavior that produces deceptive or harmful facial content.

While acknowledging the potential risks of misuse, we also recognize the significant positive potential of this technology. Our method can be broadly applied in digital creativity, virtual human interaction, and personalized content generation. Additionally, it holds promise in public-interest applications such as generating portraits of missing children or criminal suspects (e.g., by reconstructing facial contours from the semantic mask and inferring other attributes from textual descriptions), thereby contributing to public safety and social welfare. We are committed to the responsible development of AI technologies that benefit humanity.

To mitigate potential misuse and provide necessary safeguards, we are also exploring the application of our method in advancing face forgery detection. Specifically, we use the generated facial images as training data to support the development of general-purpose face forgery detection models. Preliminary experiments show that incorporating data generated by our method improves the generalization ability of these models. We will continue to share our latest progress with the research community actively.

B Dataset Details

B.1 MM-CelebA-HQ

This dataset contains 30,000 high-resolution facial images, each annotated with a corresponding semantic segmentation map and ten natural language descriptions. The semantic segmentation maps label each pixel into one of 19 categories, including *background*, *face skin*, *nose*, *eyeglasses*, *left eye*, *right eye*, *left eyebrow*, *right eyebrow*, *left ear*, *right ear*, *inner mouth*, *upper lip*, *lower lip*, *hair*, *hat*, *earring*, *necklace*, *neck*, and *clothing*. These segmentation labels provide pixel-level structural and contextual information useful for supervised learning and evaluation in image generation tasks. Each image is also paired with 10 unique text descriptions that capture detailed visual characteristics, such as facial features, expressions, accessories, age, and gender. During training, one of the 10 descriptions is randomly selected, while during testing, the first description is always used to ensure consistency. The dataset serves as a comprehensive benchmark for text-to-image generation and multimodal learning, supporting tasks such as conditional image synthesis, semantic-guided generation, and multimodal learning.

B.2 MM-FFHQ-Female

FFHQ-Text [4] is a smaller-scale but highly specialized dataset that comprises 760 high-quality female face images from the FFHQ (Flickr-Faces-HQ) [1] dataset. Each image is paired with 9 distinct natural language descriptions, detailing fine-grained facial attributes such as makeup style, hairstyle, facial expression, skin tone, and accessories. These descriptions emphasize subtle details and variations, making the dataset particularly suitable for evaluating text-to-image generation and manipulation tasks that require high sensitivity to nuanced text cues. To produce semantic masks for FFHQ-Text, we leverage two pretrained facial parsing models—FaRL [3] and SegFace [2]. For masks with overall accuracy (OA) below 0.8, we conduct manual annotation, whereas for those achieving OA $\geq$ 0.8, we adopt a randomized sampling strategy, comprising 90% from FaRL and 10% from SegFace. A randomly sampled textual description for each image is used during zero-shot evaluation to ensure consistency across experiments. The combination of detailed textual annotations, semantic masks, and high-resolution facial images enables comprehensive studies on fine-level semantic alignment and learning in multimodal models. We will release this dataset to promote community development.

C More Visualization

C.1 Multimodal Face Generation

Fig. 1 demonstrates the comparative results of multimodal face generation between our method and several state-of-the-art multimodal generation methods. As shown, our method consistently produces more realistic and semantically faithful faces, effectively integrating multiple modalities such as text descriptions and segmentation masks. Noticeably, our generated faces exhibit finer facial details and more accurate modality alignment than other methods.

C.2 Mask-to-Face Generation

In Fig. 2, we present a comprehensive comparative visualization of mask-to-face generation performance. Although our method is not explicitly designed for the mask-to-face generation task, it still achieves commendable structural alignment with the input semantic masks, yielding compelling images of superior fidelity.

C.3 Text-to-Face Generation

Fig. 3 illustrates the comparative results of text-to-face generation methods. It is evident from the examples provided that our method surpasses previous techniques in capturing subtle textual cues and translating them accurately into visual facial features. Compared to other methods, our generated faces better reflect the described attributes, demonstrating a notable improvement in text consistency.

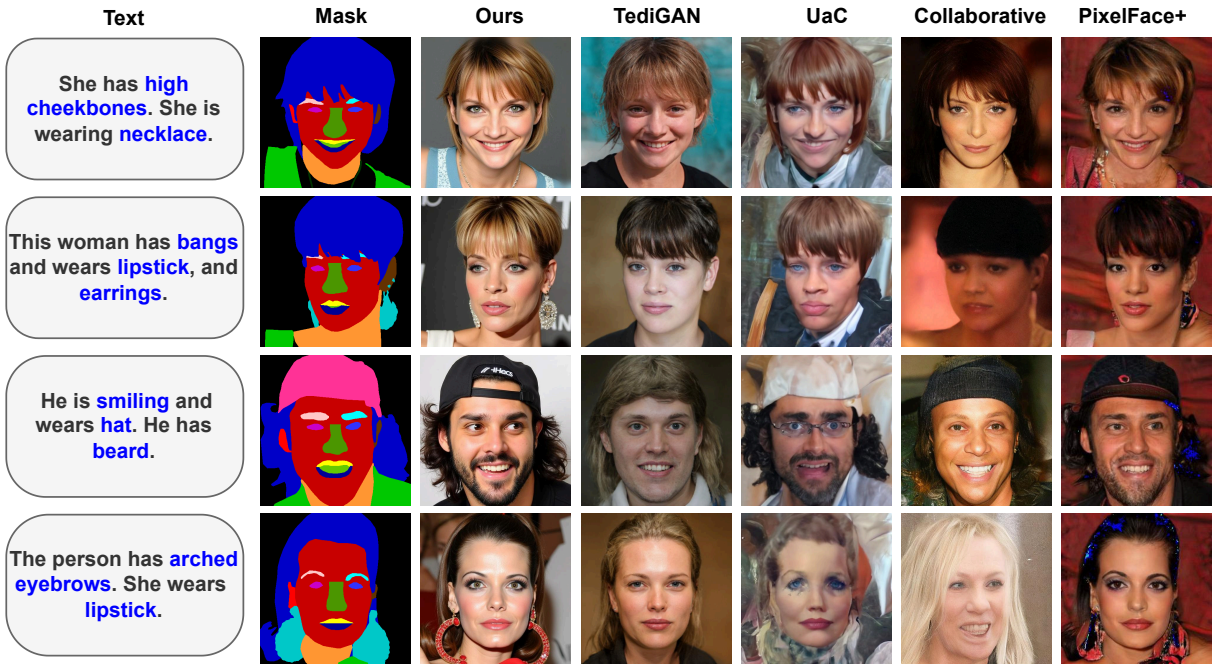

Figure 1: Comparative results of multimodal face generation on the MM-CelebA-HQ dataset.

C.4 Zero-shot Generalization

As shown in Fig. 4, our method produces significantly more faithful and realistic generations than baseline methods. The results demonstrate consistency with the input prompts while preserving fine-grained semantic structures with the mask.

C.5 Ablation Studies

Fig. 5 illustrates the visual results of our ablation studies. The figure includes variations such as *Only Global*, *Only Local*, *w/o Diffusion*, *Scalar Gating*, and our final best-performing method.

References

[1] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*. 4401–4410.
[2] Kartik Narayan, Vibashan VS, and Vishal M Patel. 2024. Segface: Face segmentation of long-tail classes. *arXiv preprint arXiv:2412.08647* (2024).
[3] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. 2022. General facial representation learning in a visual-linguistic manner. In *CVPR*. 18697–18709.
[4] Yutong Zhou. 2021. Generative adversarial network for text-to-face synthesis and manipulation. In *ACM FG*. 2940–2944.

Figure 2: Comparative results of mask-to-face generation on the MM-CelebA-HQ dataset.

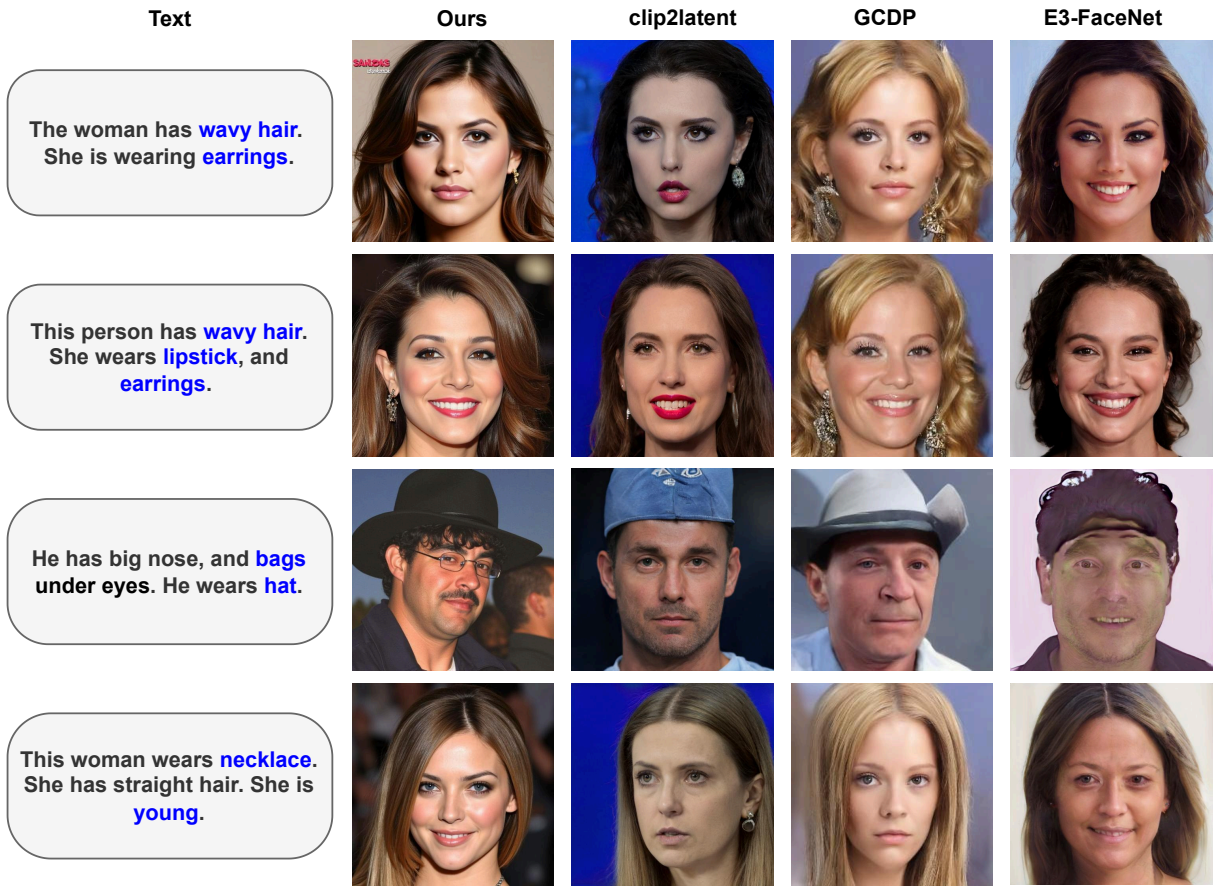

Figure 3: Comparative results of text-to-face generation on the MM-CelebA-HQ dataset.

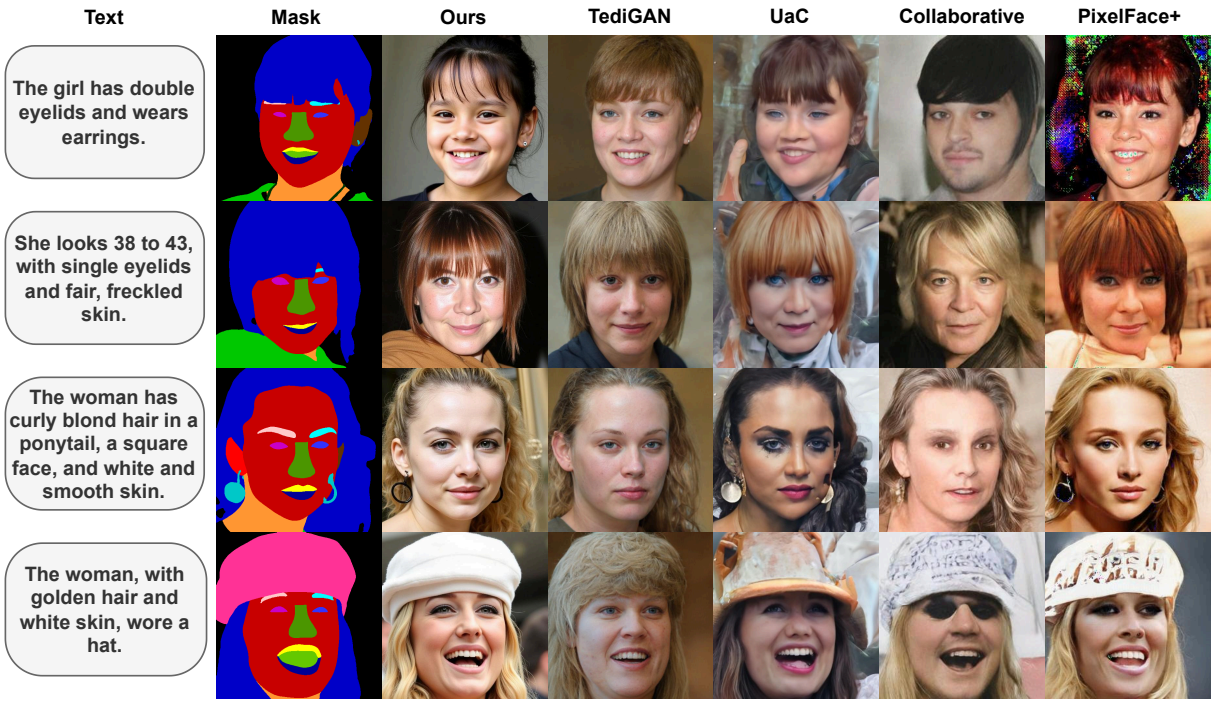

Figure 4: Comparative results of zero-shot generalization on the MM-FFHQ-Female dataset.

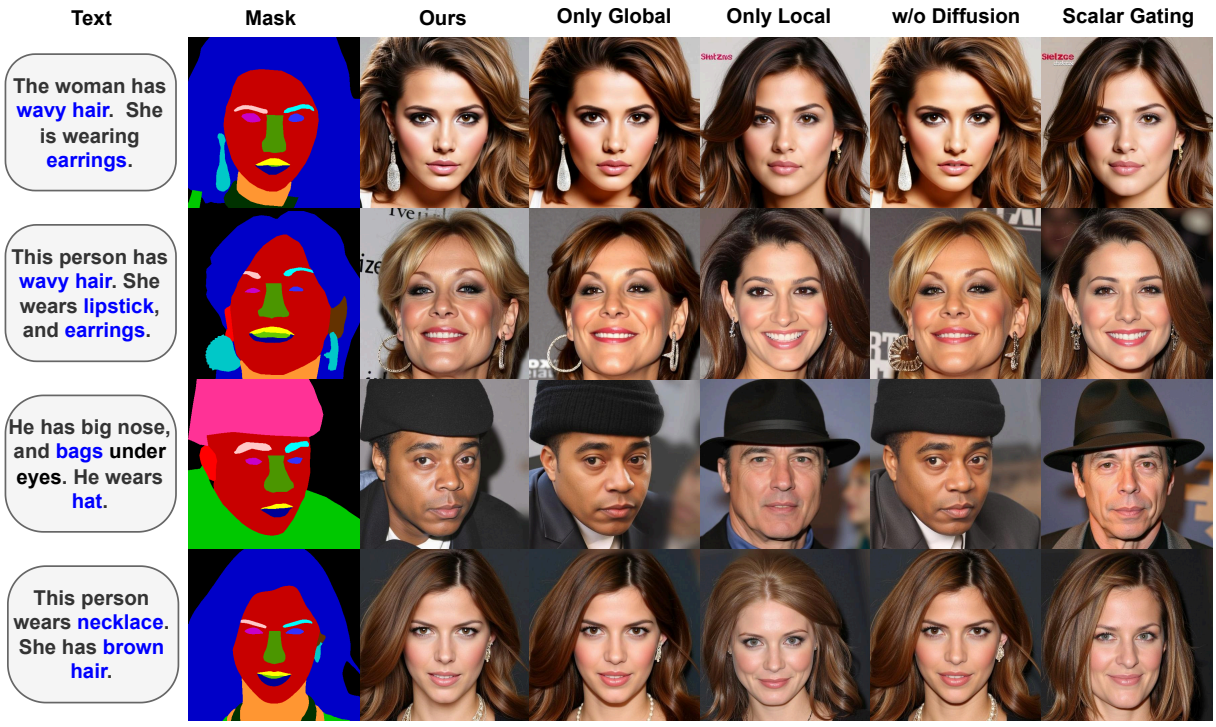

Figure 5: Comparative results of ablation studies on the MM-CelebA-HQ dataset.