

# Collaborative Diffusion for Multi-Modal Face Generation and Editing –Supplementary File–

Ziqi Huang Kelvin C.K. Chan Yuming Jiang Ziwei Liu<sup>✉</sup>  
S-Lab, Nanyang Technological University  
{ziqi002, chan0899, yuming002, ziwei.liu}@ntu.edu.sg

In this *supplementary file*, we elaborate on the implementation details of the *Collaborative Diffusion* framework in Section A. We then provide further explanations on experimental details in Section B. More qualitative results and visualizations are provided in Section C. Finally, we discuss the potential societal impacts in Section D.

## A. Implementation Details

In this section, we describe the implementation details of our *Collaborative Diffusion* framework.

### A.1. Multi-Modal Collaborative Synthesis

We adopt LDM [11] as our uni-modal diffusion models for its good balance between quality and speed. LDM [11] applies diffusion models in the latent space of autoencoders to reduce the computation overhead of training and sampling. Our framework supports both  $256 \times 256$  and  $512 \times 512$  resolution, and we will use the  $256 \times 256$  version in subsequent discussions for simplicity.

We train a Variational Autoencoder (VAE) [8], where the encoder compresses  $256 \times 256 \times 3$  resolution images into the  $64 \times 64 \times 3$  latent space, and the decoder reconstructs the  $256 \times 256 \times 3$  images from the  $64 \times 64 \times 3$  latent codes. The VAE is trained on the CelebA-HQ [6] Dataset by minimizing the following objective:

$$L_{VAE} = 1.0 \cdot L_{rec} + 1.0 \cdot L_{vgg} + 10^{-6} \cdot L_{kl}, \quad (1)$$

where  $L_{rec}$  is the  $L_1$  distance between the reconstructed image and the input image,  $L_{vgg}$  is the perceptual loss [5] using VGG-16 [13], and  $L_{kl}$  is the Kullback–Leibler divergence term which regularizes the VAE latent space towards the Gaussian distribution. The KL term is largely scaled down by a factor of  $10^{-6}$  for two reasons: 1) KL regularization was required in the original VAE for directly sampling latent codes from the Gaussian prior. In this work, we simply use VAE as an image compression tool instead of a generative model, so that we do not need strong regularization of VAE latent space. Diffusion models will take care of sampling meaningful latent codes from the weakly regularized VAE latent space. 2) Weaker KL regularization allows relatively stronger focus on image reconstruction, and thus potentially less distortion during VAE’s compression-reconstruction process. All our *dynamic diffusers* and uni-modal diffusion models are applied in the  $64 \times 64 \times 3$  latent space of the pre-trained VAE. The *reverse process* of diffusion models gradually denoises the Gaussian  $\mathbf{x}_T \in \mathbb{R}^{64 \times 64 \times 3}$  to  $\mathbf{x}_0 \in \mathbb{R}^{64 \times 64 \times 3}$  which will then be decoded to a synthesized image of size  $256 \times 256 \times 3$  using VAE’s decoder. In subsequent discussions, we will term  $\mathbf{x}_T$  as the “latent code”, and  $\mathbf{x}_0$  as the “image” to avoid confusion between diffusion models’ latent space and VAE’s latent space.

The text conditions are converted to a sequence of 77 tokens using BERT-tokenizer [2], and are then embedded using 32 transformer encoder layers to obtain the  $77 \times 640$  text condition embedding. The segmentation masks are downsampled to  $32 \times 32$  resolution, and each pixel is expanded to a  $1 \times 19$  one-hot vector to encode the 19 classes of facial components. The uni-modal diffusion models are trained with learning rate of  $2 \times 10^{-6}$  and batch size of 32 on CelebA-HQ [6]’s  $256 \times 256$  images and corresponding condition annotations.

The *dynamic diffuser*  $\mathbf{D}_{\theta_m}$  takes the noisy image  $\mathbf{x}_t$ , timestep  $t$ , and the condition  $c_m$  as input, and predicts the *influence function*  $\mathbf{I}_{m,t}$ . Since the input noisy image  $\mathbf{x}_t \in \mathbb{R}^{64 \times 64 \times 3}$  and the output *influence function*  $\mathbf{I}_{m,t} \in \mathbb{R}^{64 \times 64 \times 1}$  has the same spatial resolution, we implement *dynamic diffuser* as a UNet [12].

---

<sup>✉</sup>Corresponding author.

The timestep  $t$  is injected to the *dynamic diffuser* using Adaptive Layer Normalization (AdaLN) [11]:

$$h_{out} = (1 + s(t))\text{LayerNorm}(h_{in}) + b(t), \quad (2)$$

where  $s(\cdot)$  and  $b(\cdot)$  are linear layers that project the timestep  $t$  to the scale and bias respectively, and  $h_{in}$  and  $h_{out}$  are the intermediate activations before and after timestep injection.

The condition  $c_m$  is fed into the *dynamic diffuser* via cross-attention [16] with the intermediate activations  $h$ :

$$h_{out} = \text{CrossAttention}(h_{in}, c_m) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \quad (3)$$

$$Q = W_Q \cdot h_{in}, \quad K = W_K \cdot c_m, \quad V = W_V \cdot c_m, \quad (4)$$

where  $W_Q$ ,  $W_K$ , and  $W_V$  are learnable projection matrices. We provide an overview of the hyperparameters of *dynamic diffusers* in Table A1.

Table A1. Hyperparameters of Dynamic Diffusers.

	Dynamic Diffuser for Text Branch	Dynamic Diffuser for Mask Branch
VAE latent space shape	64×64×3	64×64×3
Number of parameters	13.1M	13.1M
Diffusion steps	1000	1000
Channels	32	32
Attention resolutions	8,4,2	8,4,2
Batch size	8 samples × 4 GPUs	8 samples × 4 GPUs
Number of iterations	187k	187k
Learning rate	2×10 <sup>-6</sup>	2×10 <sup>-6</sup>

Our *dynamic diffuser* has a much smaller model size than the conditional diffusion model, as shown in Table A2.

Table A2. Comparison of Model Size. A *dynamic diffuser* is much smaller than a uni-modal conditional diffusion model.

Model Name	Number of Parameters
Mask-Driven Pre-trained Diffusion Model	403.6M
Text-Driven Pre-trained Diffusion Model	403.6M
Dynamic Diffuser for Mask Branch	13.1M
Dynamic Diffuser for Text Branch	13.1M

## A.2. Collaborative Editing

In this work, all face editing results, including user study and the qualitative results, are applied on 256×256 *real images* in the validation split of the CelebA-HQ Dataset.

We use Imagic [7] to demonstrate that our *Collaborative Diffusion* framework can be extended from synthesis to editing. Imagic is a text-based image editing method using diffusion models, and involves three steps to complete an edit. Given the input image  $\mathbf{x}_{input}$  and target text  $c_{text,target}$ , Imagic first optimizes the text condition so that the diffusion model  $\epsilon_{\theta_{text}}$  can reconstruct the input image:

$$c_{text,opt} = \operatorname{argmin}_{c_{text}} \mathbb{E}_{\epsilon,t} \|\epsilon - \epsilon_{\theta_{text}}(\mathbf{x}_t, t, c_{text})\|^2, \quad (5)$$

where  $c_{text}$  is initialized as  $c_{text,target}$  before optimization, and  $\mathbf{x}_t$  is constructed using the *diffusion process* via  $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{input} + \sqrt{1 - \alpha_t}\epsilon$ . To further improve the fidelity of the input image, the diffusion model  $\epsilon_{\theta_{text}}$  is then fine-tuned with the optimized condition  $c_{text,opt}$  being fixed:

$$\theta_{text,opt} = \operatorname{argmin}_{\theta_{text}} \mathbb{E}_{\epsilon,t} \|\epsilon - \epsilon_{\theta_{text}}(\mathbf{x}_t, t, c_{text,opt})\|^2. \quad (6)$$

Finally, Imagic interpolates between  $c_{text,target}$  and  $c_{text,opt}$  to obtain the interpolated condition  $c_{text,int}$ :

$$c_{text,int} = \alpha \cdot c_{text,target} + (1 - \alpha) \cdot c_{text,opt}. \quad (7)$$

The edited image is synthesized using the interpolated text condition  $c_{text,int}$  and the fine-tuned diffusion model  $\epsilon_{\theta_{text}}$ .

We generalize Imagic to achieve mask-driven editing by optimizing the mask condition embedding  $c_{mask,target}$  and fine-tuning the pre-trained mask-driven model  $\epsilon_{\theta_{mask}}$ . We then use *Collaborative Diffusion* to integrate any text-driven edit and mask-driven edit on the same input image into a collaborative edit.

## B. Further Explanations on Experimental Details

### B.1. Dataset

The CelebA-HQ Dataset [6] consists of 30,000 high-resolution images. We use the multi-modal annotations for these images in the CelebAMask-HQ [9] Dataset and the CelebA-Dialog Dataset [4]. The 30,000 images are split into the training set (27,000 images) and validation set (3,000 images). The training of uni-modal diffusion models and the *dynamic diffusers* are conducted on the training set, and all the results reported and shown in this work are using multi-modal conditions from the validation set.

The segmentation masks in the CelebAMask-HQ Dataset has 19 classes including facial components and accessories: ‘background’, ‘skin’, ‘nose’, ‘left eye’, ‘right eye’, ‘left eyebrow’, ‘right eyebrow’, ‘left ear’, ‘right ear’, ‘mouth’, ‘upper lip’, ‘lower lip’, ‘hair’, ‘hat’, ‘eyeglass’, ‘earring’, ‘necklace’, ‘neck’, and ‘cloth’.

The texts in the CelebA-Dialog Dataset provide fine-grained natural language descriptions of the five attributes: ‘Bangs’, ‘Eyeglasses’, ‘Beard’, ‘Smiling’, ‘Age’. To avoid conflict between segmentation masks and texts, we trimmed the descriptions regarding ‘Bangs’, ‘Eyeglasses’ and ‘Smiling’ from the natural language descriptions as they are described by segmentation masks as well.

### B.2. Implementation Details on Comparison Methods

**TediGAN** [17, 18]. *TediGAN* is a StyleGAN-based method for text-driven face generation and manipulation. It can be extended to support other modality’s guidance by projecting the conditions into StyleGAN’s  $\mathcal{W}+$  latent space, and performing style mixing to achieve multi-modal control. We use TediGAN [17, 18]’s official implementation for text-driven and mask-driven generation and editing. For multi-modal driven generation, we mix the style codes of text and mask using TediGAN’s style mixing control mechanism. For editing, the style codes are initialized using the inverted  $\mathcal{W}+$  codes of the input image, and the remaining steps are the same as generation.

**Composable** [10]. Both *Composable* and *Ours* use the same set of pre-trained uni-modal conditional diffusion models described in Section A. To accelerate the time-consuming sampling process of diffusion models while maintaining fair comparisons, we use DDIM [14] with 50 steps in all experiments (*i.e.*, quantitative, qualitative, and user study) involving *Composable* or *Ours*.

## C. More Qualitative Results

We show various qualitative results in Figure A1-A5, which are located at the end of this Supplementary File.

### C.1. Generation and Editing

We provide more face generation results in Figure A1 and Figure A2, and face editing results in Figure A3.

### C.2. Visualization of Influence Functions

In Figure A4 and Figure A5, we visualize the *influence functions* to show their spatial-temporal variation. Given the mask condition in Figure A4(a), Figure A4(b) displays the *influence functions* of the mask-driven collaborator at each DDIM sampling step  $t = 980, 960, \dots, 20, 0$ , from the left to right, top to down. The text branches’ *influence functions* are displayed similarly. In Figure A4(f), we show the intermediate diffusion results  $\mathbf{x}_t$  for  $t = 980, 960, \dots, 20, 0$  by decoding them to the image space using the VAE decoder. The final synthesized image is displayed in Figure A4(e). Figure A5 displays the intermediate results using a different set of multi-modal conditions, and is arranged in the same way as Figure A4.

## D. Potential Societal Impacts

*Collaborative Diffusion* can achieve high-quality real image editing driven by different modalities. However, such capabilities could be applied to maliciously manipulate real human faces. Therefore, we advise users to use *Collaborative Diffusion* only for proper recreational purposes.

The rapid progress in generative models unleashes creativity, but inevitably introduces various societal concerns. First, it becomes easier to create false imagery or maliciously manipulate the data, which could lead to the spread of misinformation. Second, training data might be revealed during the sampling process without explicit consent from data owner [15]. Third, generative models potentially suffer from the biases present in the training data [3]. For *Collaborative Diffusion*, we conducted training on CelebA-HQ [6]’s faces of various celebrities, which could potentially deviate from the looks of the general population. We hope to see more research to alleviate the risks and biases of generative models, and we advise all to apply generative models with discretion.

## References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1
- [3] Patrick Esser, Robin Rombach, and Björn Ommer. A note on data biases in generative models. In *NeurIPS Workshop*, 2020. 4
- [4] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-Edit: Fine-grained facial editing via dialog. In *ICCV*, 2021. 3
- [5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1
- [6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 1, 3, 4
- [7] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2
- [8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1
- [9] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 3
- [10] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022. 3
- [11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1
- [12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1
- [13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1
- [14] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3
- [15] Patrick Tinsley, Adam Czajka, and Patrick Flynn. This face does not exist... but it might be yours! identity leakage in generative models. In *WACV*, 2021. 4
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2
- [17] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. TediGAN: Text-guided diverse face image generation and manipulation. In *CVPR*, 2021. 3
- [18] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards open-world text-guided face image generation and manipulation. *arXiv preprint arXiv:2104.08910*, 2021. 3

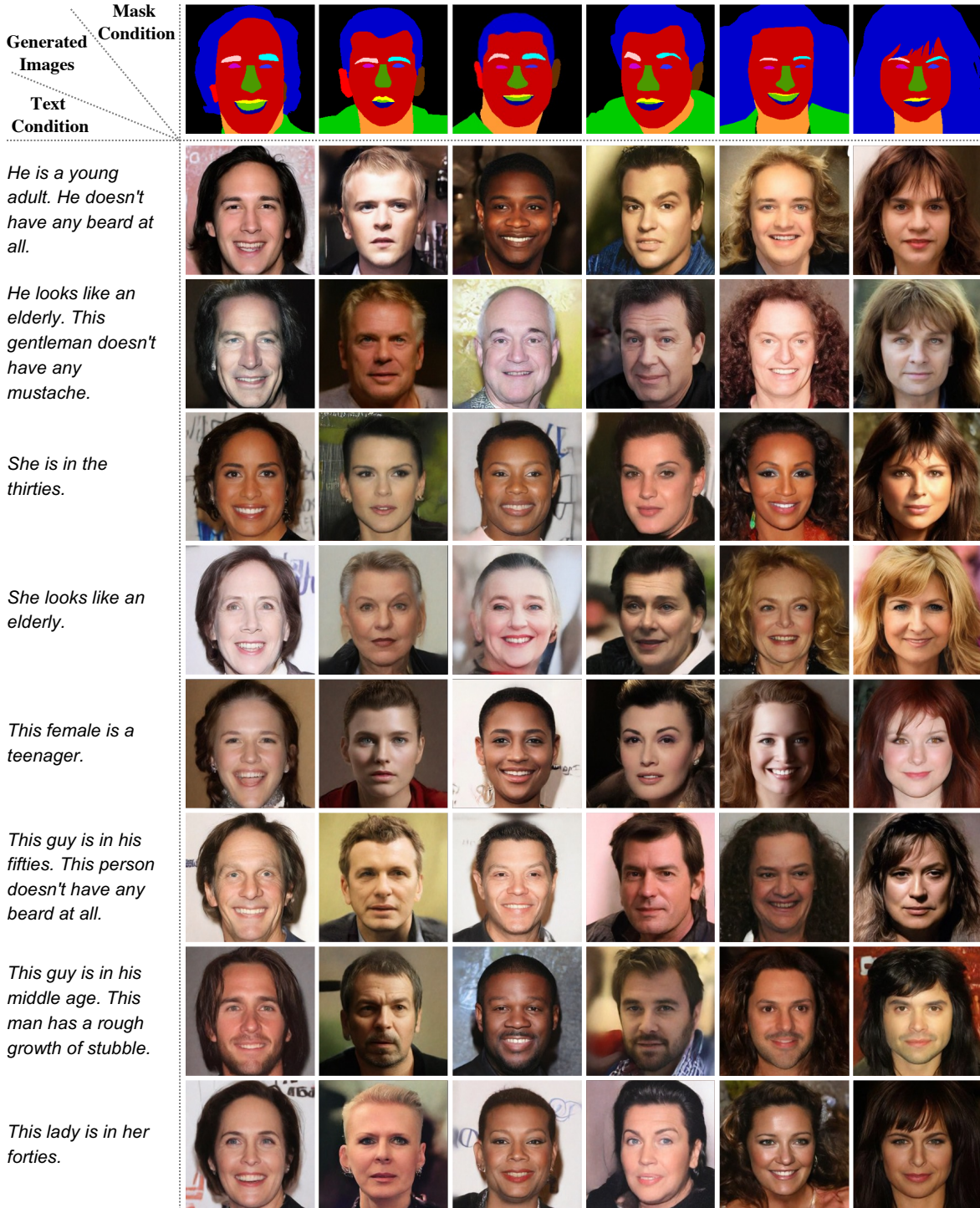


Figure A1. **More Face Generation Results (A)**. Our method generates realistic images under different combinations of multi-modal conditions, even for relatively rare combinations in the training distribution, such as a man with long hair.

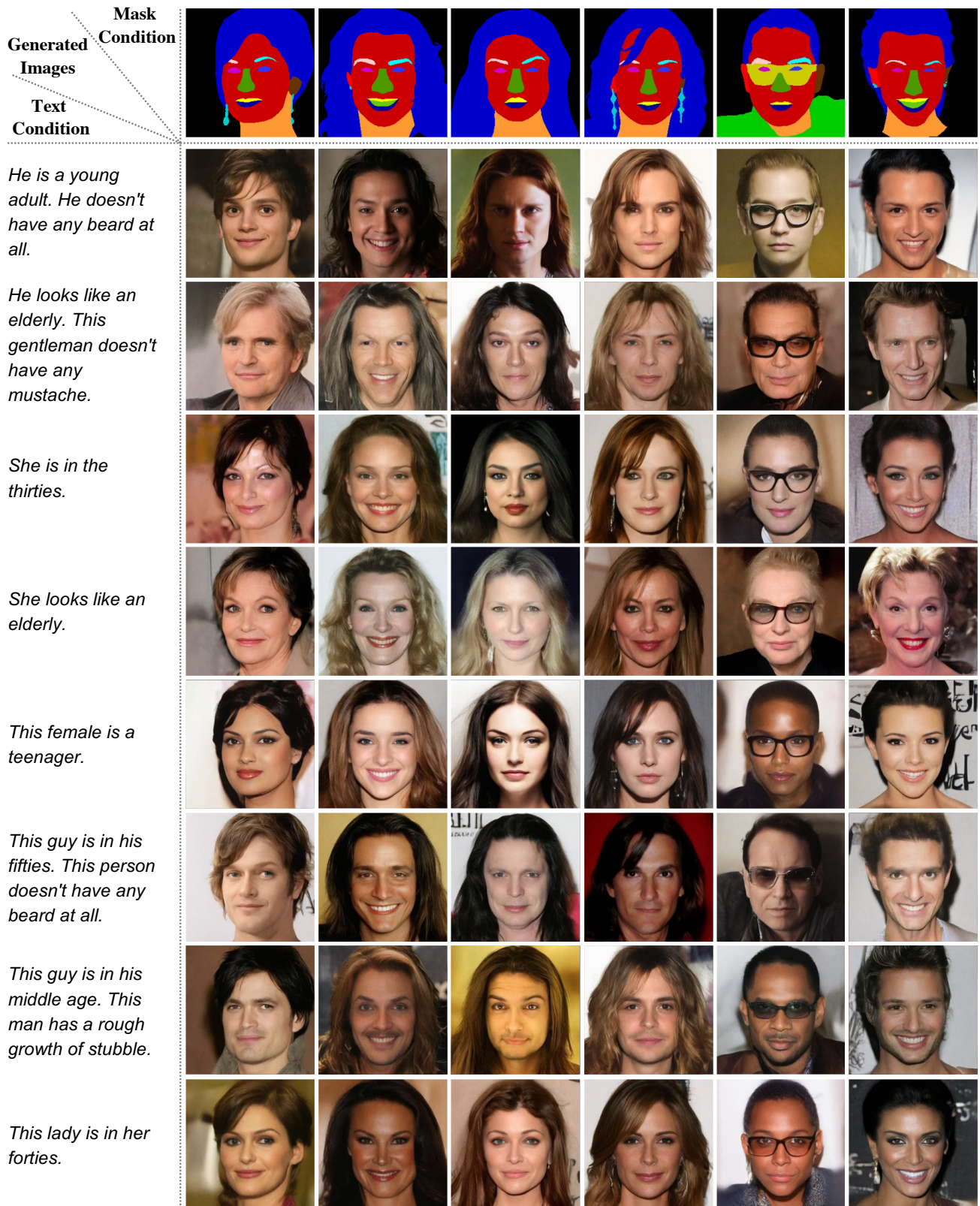


Figure A2. **More Face Generation Results (B)**. Our method generates realistic images under different combinations of multi-modal conditions, even for relatively rare combinations in the training distribution, such as a man with long hair.



















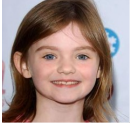






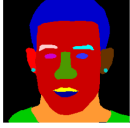







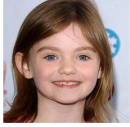

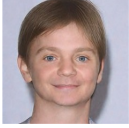



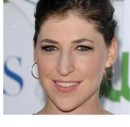



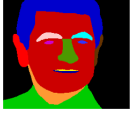




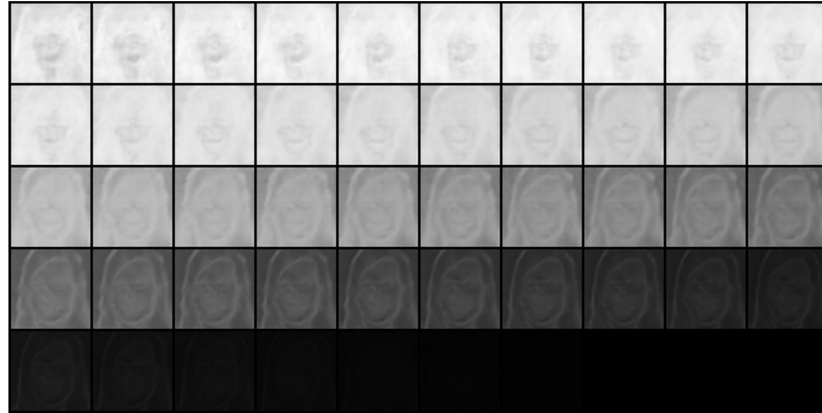
Input Image	Target Mask	Target Text	Edited Image	Input Image	Target Mask	Target Text	Edited Image
		<i>This woman looks like an elderly.</i>				<i>This female is in the middle age.</i>	
		<i>He is a young adult. He doesn't have any beard at all.</i>				<i>He is a young adult. He doesn't have any beard at all.</i>	
		<i>This woman looks like an elderly.</i>				<i>This man doesn't have any mustache at all. This guy is in his forties.</i>	
		<i>This female is in the middle age.</i>				<i>This female is in the middle age.</i>	
		<i>He is a teen. The face is covered with short pointed beard.</i>				<i>He is a young adult. He doesn't have any beard at all.</i>	
		<i>This female is in the middle age.</i>				<i>He is a young adult. He doesn't have any beard at all.</i>	
		<i>He is a teen. The face is covered with short pointed beard.</i>				<i>He is a young adult. He doesn't have any beard at all.</i>	
		<i>He looks like an elderly. This gentleman doesn't have any mustache.</i>				<i>She is a teenager.</i>	

Figure A3. **Face Editing Results.** Given the input real image and target conditions, we display the edited image using our method.



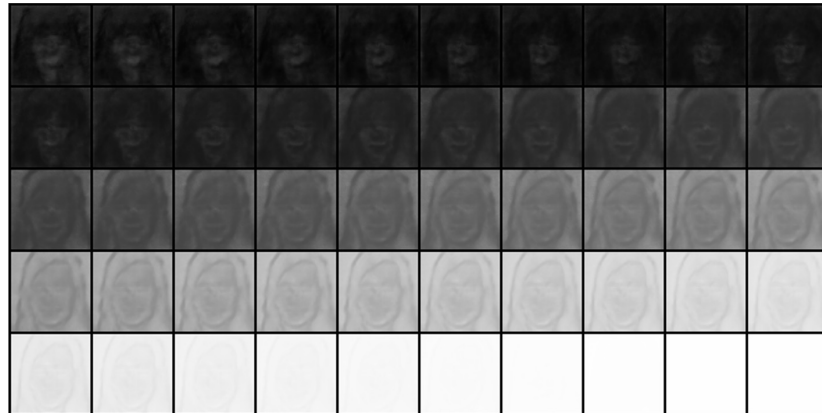
(a)  $\mathbf{c}_{mask}$



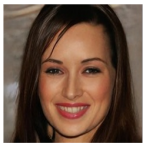
(b)  $\mathbf{I}_{mask,t}$  for  $t = 980, 960, \dots, 20, 0$

This person is  
in her forties.

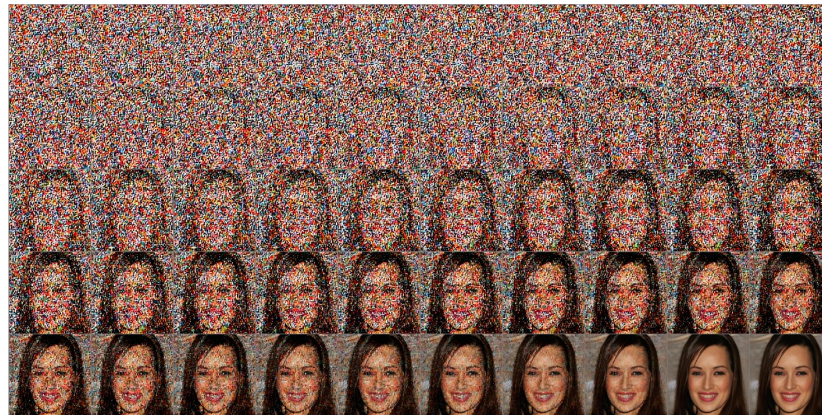
(c)  $\mathbf{c}_{text}$



(d)  $\mathbf{I}_{text,t}$  for  $t = 980, 960, \dots, 20, 0$



(e)  $\mathbf{x}_0$



(f)  $\mathbf{x}_t$  for  $t = 980, 960, \dots, 20, 0$

Figure A4. **Visualization of Influence Functions (A).** The *influence function* varies spatially at different face regions, and temporally at different diffusion timesteps. The spatial-temporal adaptivity of *influence functions* facilitates effective collaboration.





(a)  $\mathcal{C}_{mask}$



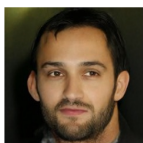
(b)  $I_{mask,t}$  for  $t = 980, 960, \dots, 20, 0$

This man has  
beard of medium  
length. He is in  
his thirties.

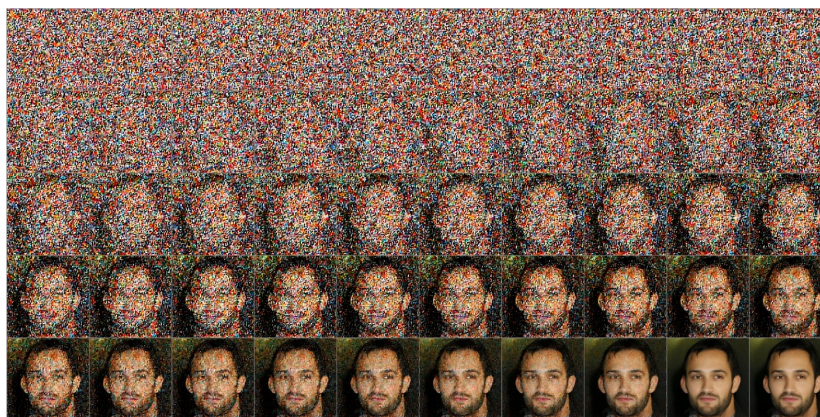
(c)  $\mathcal{C}_{text}$



(d)  $I_{text,t}$  for  $t = 980, 960, \dots, 20, 0$



(e)  $\mathbf{x}_0$



(f)  $\mathbf{x}_t$  for  $t = 980, 960, \dots, 20, 0$

Figure A5. **Visualization of Influence Functions (B).** The *influence function* varies spatially at different face regions, and temporally at different diffusion timesteps. The spatial-temporal adaptivity of *influence functions* facilitates effective collaboration.