Collaborative Diffusion for Multi-Modal Face Generation and Editing

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

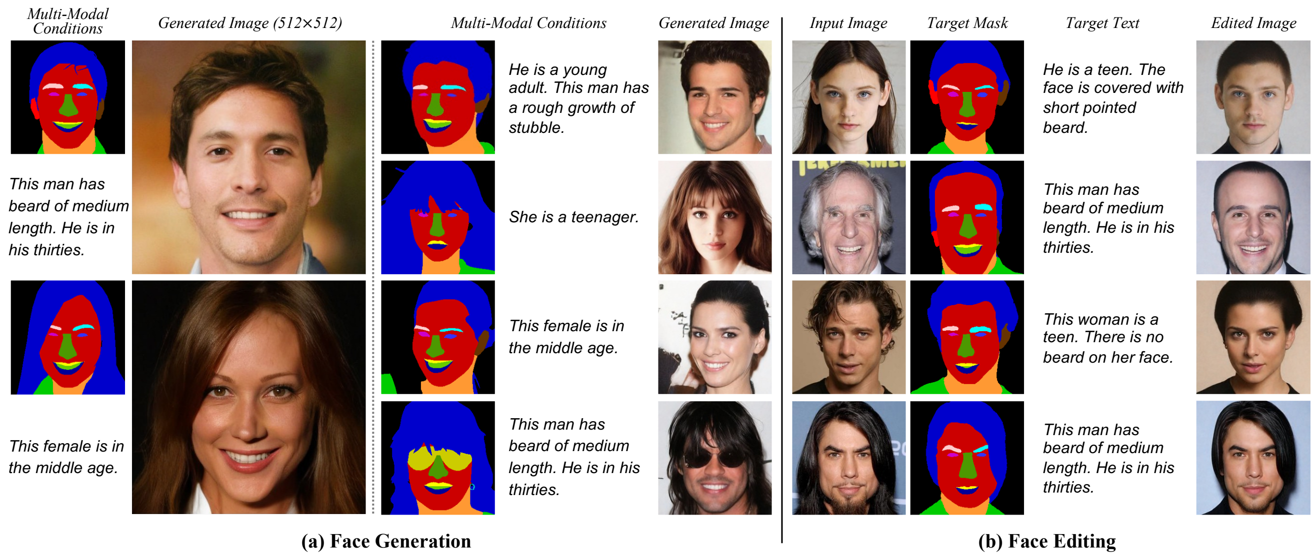

Figure 1. We propose ***Collaborative Diffusion***, where users can use multiple modalities to control face generation and editing. **(a) Face Generation**. Given multi-modal controls, our framework synthesizes high-quality images consistent with the input conditions. **(b) Face Editing**. Collaborative Diffusion also supports multi-modal editing of real images with promising identity preservation capability.

Abstract

Diffusion models arise as a powerful generative tool recently. Despite the great progress, existing diffusion models mainly focus on uni-modal control, i.e.*, the diffusion process is driven by only one modality of condition. To further unleash the users' creativity, it is desirable for the model to be controllable by multiple modalities simultaneously,* e.g. *generating and editing faces by describing the age (text-driven) while drawing the face shape (mask-driven).*

*In this work, we present **Collaborative Diffusion**, where pre-trained uni-modal diffusion models collaborate to achieve multi-modal face generation and editing without re-training. Our key insight is that diffusion models driven by different modalities are inherently complementary regarding the latent denoising steps, where bilateral connections can be established upon. Specifically, we propose dynamic diffuser, a meta-network that adaptively hallucinates multi-modal denoising steps by predicting the spatial-temporal influence functions for each pre-trained uni-modal model. Collaborative Diffusion not only collaborates generation capabilities from uni-modal diffusion models, but also integrates multiple uni-modal manipulations to perform multi-modal editing. Extensive qualitative and quantitative experiments demonstrate the superiority of our framework in both image quality and condition consistency.*

1. Introduction

Recent years have witnessed substantial progress in image synthesis and editing with the surge of diffusion models [9, 20, 56, 58]. In addition to the remarkable synthesis quality, one appealing property of diffusion models is the flexibility of conditioning on various modalities, such as texts [3, 16, 29, 30, 39, 48, 51], segmentation masks [48, 65, 66], and sketches [7, 65]. However, existing explorations are largely confined to the use of a single modality at a time. The exploitation of multiple conditions remains under-explored. As a generative tool, its controllability is still limited.

To unleash users' creativity, it is desirable that the model

✉Corresponding author.
Project page: https://ziqihuangg.github.io/projects/collaborative-diffusion.html
Code: https://github.com/ziqihuangg/Collaborative-Diffusion

is simultaneously controllable by multiple modalities. While it is trivial to extend the current supervised framework with multiple modalities, training a large-scale model from scratch is computationally expensive, especially when extensive hyper-parameter tuning and delicate architecture designs are needed. More importantly, each trained model could only accept a fixed combination of modalities, and hence re-training is necessary when a subset of modalities are absent, or when additional modalities become available. The above demonstrates the necessity of a unified framework that effectively exploits pre-trained models and integrate them for multi-modal synthesis and editing.

In this paper, we propose ***Collaborative Diffusion***, a framework that synergizes pre-trained uni-modal diffusion models for multi-modal face generation and editing without the need of re-training. Motivated by the fact that different modalities are complementary to each other (*e.g.*, *text* for age and *mask* for hair shape), we explore the possibility of establishing lateral connections between models driven by different modalities. We propose *dynamic diffuser* to adaptively predict the spatial-temporal *influence function* for each pre-trained model. The *dynamic diffuser* dynamically determines spatial-varying and temporal-varying influences of each model, suppressing contributions from irrelevant modalities while enhancing contributions from admissible modalities. In addition to multi-modal synthesis, the simplicity and flexibility of our framework enable extension to multi-modal face editing with minimal modifications. In particular, the *dynamic diffuser* is first trained for collaborative synthesis. It is then fixed and combined with existing face editing approaches [29, 50] for multi-modal editing. It is worth-mentioning that users can select the best editing approaches based on their needs without the need of altering the *dynamic diffusers*.

We demonstrate both qualitatively and quantitatively that our method achieves superior image quality and condition consistency in both synthesis and editing tasks. Our contributions can be summarized as follows:

- We introduce ***Collaborative Diffusion***, which exploits pre-trained uni-modal diffusion models for multi-modal controls without re-training. Our approach is the first attempt towards flexible integration of uni-modal diffusion models into a single collaborative framework.
- Tailored for the iterative property of diffusion models, we propose *dynamic diffuser*, which predicts the spatial-varying and temporal-varying *influence functions* to selectively enhance or suppress the contributions of the given modalities at each iterative step.
- We demonstrate the flexibility of our framework by extending it to face editing driven by multiple modalities. Both quantitative and qualitative results demonstrate the superiority of *Collaborative Diffusion* in multi-modal face generation and editing.

2. Related Work

Diffusion Models. Diffusion models [20, 56, 58] have recently become a mainstream approach for image synthesis [9, 11, 38] apart from Generative Adversarial Networks (GANs) [14], and success has also been found in various domains including video generation [17, 19, 55, 63], image restoration [21, 52], semantic segmentation [1, 4, 15], and natural language processing [2]. In the diffusion-based framework, models are trained with score-matching objectives [22, 64] at various noise levels, and sampling is done via iterative denoising. Existing works focus on improving the performance and efficiency of diffusion models through enhanced architecture designs [16, 48] and sampling schemes [57]. In contrast, this work focuses on exploiting existing models, and providing a succinct framework for multi-modal synthesis and editing without large-scale re-training of models.

Face Generation. Existing face generation approaches can be divided into three main directions. Following the GAN paradigm, the *StyleGAN* series [26–28] boost the quality of facial synthesis, and provide an interpretable latent space for steerable style controls and manipulations. The *vector-quantized* approaches [12, 61] learn a discrete codebook by mapping the input images into a low-dimensional discrete feature space. The learned codebook is then sampled, either sequentially [12, 61] or parallelly [5, 6, 16], for synthesis. In contrast to the previous two approaches, *diffusion models* are trained with a stationary objective, without the need of optimizing complex losses (*e.g.*, adversarial loss) or balancing multiple objectives (*e.g.*, codebook loss versus reconstruction loss). With training simplicity as a merit, diffusion-based approaches have become increasingly popular in recent years. Our framework falls in the diffusion-based paradigm. In particular, we leverage pre-trained diffusion models for multi-modal generation and editing.

Conditional Face Generation and Editing. Conditional generation [10, 11, 13, 31, 34, 37, 40, 43–46, 51, 65, 67–71] and editing [8, 32, 41, 53, 54, 67, 68] is an active line of research focusing on conditioning generative models on different modalities, such as texts [24, 41, 67, 68], segmentation masks [32, 33, 40, 47], and audios [59]. For example, StyleCLIP [41], DiffusionCLIP [30], and many others [35, 60] have demonstrated remarkable performance in text-guided face generation and editing. However, most existing models do not support simultaneous conditioning on multiple modalities (*e.g.*, text and mask at the same time), and supporting additional modalities often requires time-consuming model re-training and extensive hyper-parameter tuning, which are not preferable in general. In this work, we propose *Collaborative Diffusion* to exploit pre-trained uni-modal diffusion models [48] (*e.g.*, text-driven and mask-driven models) to achieve multi-modal conditioning without model re-training.

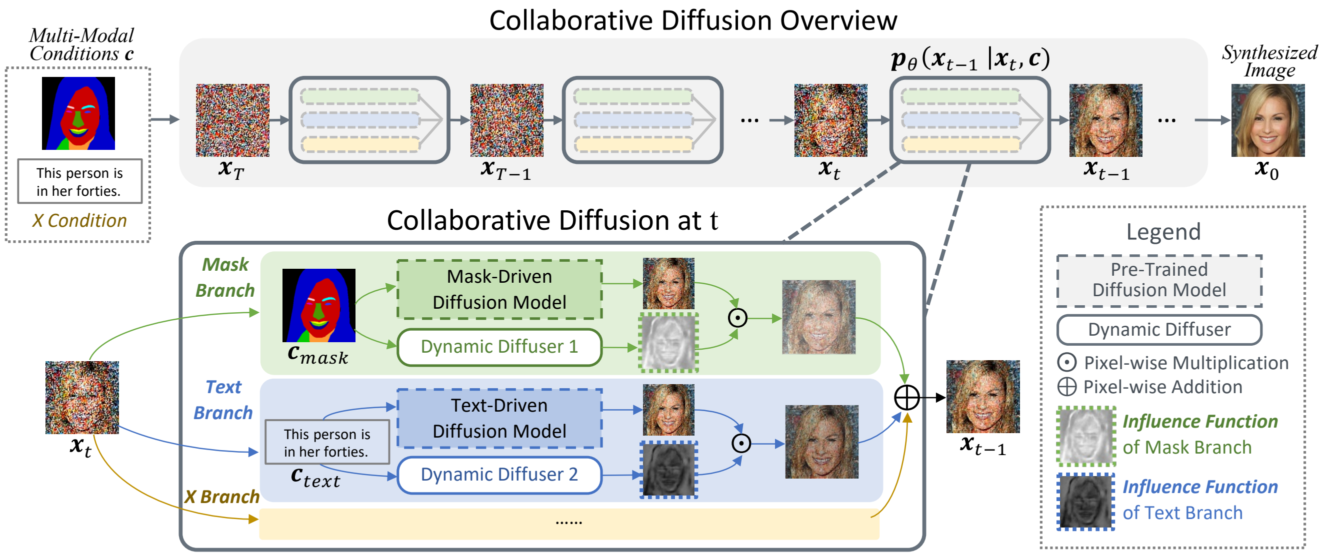

Figure 2. **Overview of Collaborative Diffusion**. We use pre-trained uni-modal diffusion models to perform multi-modal guided face generation and editing. At each step of the reverse process (*i.e.*, from timestep t to $t-1$), the *dynamic diffuser* predicts the spatial-varying and temporal-varying *influence function* to selectively enhance or suppress the contributions of the given modality.

3. Collaborative Diffusion

We propose ***Collaborative Diffusion***, which exploits multiple pre-trained uni-modal diffusion models (Section 3.1) for multi-modal generation and editing. The key of our framework is the *dynamic diffuser*, which adaptively predicts the *influence functions* to enhance or suppress the contributions of the pre-trained models based on the spatial-temporal influences of the modalities. Our framework is compatible with most existing approaches for both multi-modal guided synthesis (Section 3.2) and multi-modal editing (Section 3.3).

3.1. Uni-Modal Conditional Diffusion Models

Diffusion models are a class of generative models that model the data distribution in the form of $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T}$. The *diffusion process* (*a.k.a. forward process*) gradually adds Gaussian noise to the data and eventually corrupts the data $\mathbf{x}_0$ into an approximately pure Gaussian noise $\mathbf{x}_T$ using a variance schedule $\beta_1, \ldots, \beta_T$:

$$
\begin{aligned}
q(\mathbf{x}_{1:T}|\mathbf{x}_0) &:= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \\
q(\mathbf{x}_t|\mathbf{x}_{t-1}) &:= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}).
\end{aligned} \tag{1}
$$

Reversing the *forward process* allows sampling new data $\mathbf{x}_0$ by starting from $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$. The *reverse process* is defined as a Markov chain where each step is a learned Gaussian transition $(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$:

$$
\begin{aligned}
p_\theta(\mathbf{x}_{0:T}) &:= p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \\
p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &:= \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)).
\end{aligned} \tag{2}
$$

Training diffusion models relies on minimizing the variational bound on $p(x)$'s negative log-likelihood. The commonly used optimization objective L_{DM} [20] reparameterizes the learnable Gaussian transition as $\boldsymbol{\epsilon}_\theta(\cdot)$, and temporally reweights the variational bound to trade for better sample quality:

$$
L_{\mathrm{DM}}(\theta) := \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2 \right], \tag{3}
$$

where $\mathbf{x}_t$ can be directly approximated by $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, with $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ and $\alpha_t := 1 - \beta_t$.

To sample data $\mathbf{x}_0$ from a trained diffusion model $\boldsymbol{\epsilon}_\theta(\cdot)$, we iteratively denoise $\mathbf{x}_t$ from $t = T$ to $t = 1$ with noise $\mathbf{z}$:

$$
\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z} \tag{4}
$$

The unconditional diffusion models can be extended to model conditional distributions $p_\theta(\mathbf{x}_0|c)$, where $\mathbf{x}_0$ is the image corresponding to the condition c such as class labels, segmentation masks, and text descriptions. The conditional diffusion model receives an additional input $\tau(c)$ and is trained by minimizing $\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \tau(c))\|^2$, where $\tau(\cdot)$ is an encoder that projects the condition c to an embedding $\tau(c)$. For brevity, we will use c to represent $\tau(c)$ in our subsequent discussions.

Algorithm 1 Dynamic Diffuser Training

repeat
  $\mathbf{x}_0, c_1, c_2, ..., c_M \sim q(\mathbf{x}_0, c_1, c_2, ..., c_M)$
  $t \sim \text{Uniform}(\{1, \ldots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  for $m = 1, ..., M$ **do**
    $\boldsymbol{\epsilon}_{pred,m,t} = \boldsymbol{\epsilon}_{\theta_m}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, c_m)$
    $\mathbf{I}_{m,t} = \mathbf{D}_{\phi_m}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, c_m)$
  end for
  $\hat{\mathbf{I}}_{m,t,p} = \frac{\exp(\mathbf{I}_{m,t,p})}{\sum_{j=1}^{M}\exp(\mathbf{I}_{j,t,p})}$, softmax at each pixel p
  $\boldsymbol{\epsilon}_{pred,t} = \sum_{m=1}^{M} \hat{\mathbf{I}}_{m,t} \odot \boldsymbol{\epsilon}_{pred,m,t}$
  Take gradient descent step on
    $\nabla_\phi \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{pred,t} \right\|^2$ where $\phi = \{\phi_m | m = 1, ..., M\}$
until converged

Algorithm 2 Collaborative Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  for $m = 1, ..., M$ **do**
    $\boldsymbol{\epsilon}_{pred,m,t} = \boldsymbol{\epsilon}_{\theta_m}(\mathbf{x}_t, t, c_m)$
    $\mathbf{I}_{m,t} = \mathbf{D}_{\phi_m}(\mathbf{x}_t, t, c_m)$
  end for
  $\hat{\mathbf{I}}_{m,t,p} = \frac{\exp(\mathbf{I}_{m,t,p})}{\sum_{j=1}^{M}\exp(\mathbf{I}_{j,t,p})}$, softmax at each pixel p
  $\boldsymbol{\epsilon}_{pred,t} = \sum_{m=1}^{M} \hat{\mathbf{I}}_{m,t} \odot \boldsymbol{\epsilon}_{pred,m,t}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{pred,t}\right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

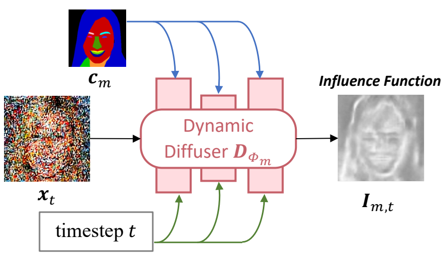

Figure 3. **Dynamic Diffuser.** It predicts the *influence function* that represents the desired level of contribution from each pre-trained diffusion model. The *influence function* varies at different timestep t and different spatial locations, and is dependent on the state of diffusion $\mathbf{x}_t$ and the condition input c_m.

3.2. Multi-Modal Collaborative Synthesis

In our ***Collaborative Diffusion***, multiple uni-modal diffusion models collaborate at each step of the denoising process for multi-modal guided synthesis. The core of our framework is the *dynamic diffuser*, which determines the extent of contribution from each collaborator by predicting the spatial-temporal *influence functions*.

Problem Formulation. Given M pre-trained uni-modal conditional diffusion models $\{\epsilon_{\theta_m}\}$ (each models the distribution $p(\mathbf{x}_0|c_m)$), where the modality index $m = 1, ..., M$, our objective is to sample from $p(\mathbf{x}_0|\mathbf{c})$, where $\mathbf{c} = \{c_1, c_2, \cdots, c_M\}$, without altering pre-trained models.

Dynamic Diffuser. In diffusion models, each step of the *reverse process* requires predicting the noise $\boldsymbol{\epsilon}$. With multiple diffusion models collaborating, we need to carefully determine *when*, *where*, and *how much* each diffusion model contributes to the prediction $\boldsymbol{\epsilon}$. At each diffusion timestep $t = T, ..., 1$, the influence $\mathbf{I}_{m,t}$ from each pre-trained diffusion model $\boldsymbol{\epsilon}_{\theta_m}$ is adaptively determined by a *dynamic diffuser* $\mathbf{D}_{\phi_m}$:

$$\mathbf{I}_{m,t} = \mathbf{D}_{\phi_m}(\mathbf{x}_t, t, c_m), \tag{5}$$

where $m = 1, ..., M$ is the index of the modalities, $\mathbf{I}_{m,t} \in \mathbb{R}^{h \times w}$, $\mathbf{x}_t$ is the noisy image at time t, c_m is the condition of the m^{th} modality, and $\mathbf{D}_{\phi_m}$ is the *dynamic diffuser* implemented by a UNet [49]. To regularize the overall influence strength, we perform softmax across all modalities' $\mathbf{I}_{m,t}$ at each pixel p to obtain the final *influence function* $\hat{\mathbf{I}}_{m,t}$:

$$\hat{\mathbf{I}}_{m,t,p} = \frac{\exp(\mathbf{I}_{m,t,p})}{\sum_{j=1}^{M}\exp(\mathbf{I}_{j,t,p})}. \tag{6}$$

Multi-Modal Collaboration. We use the M learned *influence functions* $\hat{\mathbf{I}}_{m,t}$ to control the contribution from each pre-trained diffusion model at each denoising step:

$$\boldsymbol{\epsilon}_{pred,t} = \sum_{m=1}^{M} \hat{\mathbf{I}}_{m,t} \odot \boldsymbol{\epsilon}_{\theta_m}(x_t, t, c_m) \tag{7}$$

where $\boldsymbol{\epsilon}_{\theta_m}$ is the m^{th} collaborator, and $\odot$ denotes pixel-wise multiplication. The complete training procedure of *dynamic diffusers* and the image sampling strategy in our *Collaborative Diffusion* framework are detailed in Algorithm 1 and Algorithm 2, respectively.

Relation with Composable Diffusion. Composable Diffusion [36] combines two instances of the same text-to-image diffusion model to achieve compositional visual generation via intermediate results addition. Our framework is related to Composable Diffusion in terms of composing diffusion models for image synthesis, but is substantially different in terms of the task nature and methodology. Composable Diffusion aims to decompose the text condition into elementary segments to factorize the conditional synthesis problem, while we aim to integrate uni-modal collaborators to achieve multi-modal controls. Furthermore, different from Composable Diffusion, which can only compose instances of the

Algorithm 3 Collaborative Editing

Require:
input image $\mathbf{x}_{input}$, target conditions $c_{m,target}$,
diffusion models $\boldsymbol{\epsilon}_{\theta_m}$, dynamic diffusers $\mathbf{D}_{\phi_m}$, $(m = 1, \ldots, M)$,
interpolation scale α

for $m = 1, ..., M$ **do** ▷ Uni-Modal Editing
  $c_m = c_{m,target}$
  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_{input} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$
  $c_{m,opt} = \text{argmin}_{c_m} \mathbb{E}_{\boldsymbol{\epsilon},t} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta_m}(\mathbf{x}_t, t, c_m)\|^2$
  $\theta_{m,opt} = \text{argmin}_{\theta_m} \mathbb{E}_{\boldsymbol{\epsilon},t} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta_m}(\mathbf{x}_t, t, c_{m,opt})\|^2$
  $c_{m,int} = \alpha \cdot c_{m,target} + (1-\alpha) \cdot c_{m,opt}$
end for
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Collaborate the Uni-Modal Edits
for $t = T, \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  for $m = 0, ..., M$ **do**
    $\boldsymbol{\epsilon}_{pred,m,t} = \boldsymbol{\epsilon}_{\theta_{m,opt}}(\mathbf{x}_t, t, c_{m,int})$
    $\mathbf{I}_{m,t} = \mathbf{D}_{\phi_m}(\mathbf{x}_t, t, c_{m,int})$
  end for
  $\hat{\mathbf{I}}_{m,t,p} = \frac{\exp(\mathbf{I}_{m,t,p})}{\sum_{j=1}^{M} \exp(\mathbf{I}_{j,t,p})}$, softmax at each pixel p
  $\boldsymbol{\epsilon}_{pred,t} = \sum_{m=1}^{M} \hat{\mathbf{I}}_{m,t} \odot \boldsymbol{\epsilon}_{pred,m,t}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{pred,t}\right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

same text-based synthesis model (*i.e.*, the weights of the diffusion models being composed are the same), our framework possesses the flexibility in integrating models with different weights, architectures, and modalities through the learned *dynamic diffusers*.

3.3. Multi-Modal Collaborative Editing

In addition to face synthesis, our framework is also capable of combining multiple facial manipulations, each of which is guided by a different modality, into a collaborative edit. Our collaborative framework in theory can integrate any existing uni-modal diffusion-based editing approach for collaborative editing.

In this work, we demonstrate such possibility by extending Imagic [29] to the multi-modal paradigm. We first follow Imagic to fine-tune the embeddings and models to better capture the input face identity during editing. The trained *dynamic diffusers* discussed in Section 3.2 are then used to combine the fine-tuned models. Algorithm 3 displays the complete procedure of collaborative editing. Note that fine-tuning the pre-trained models is for identity preservation proposed in Imagic, which is independent to our framework. The extension of our framework to editing requires no further training of *dynamic diffusers*.

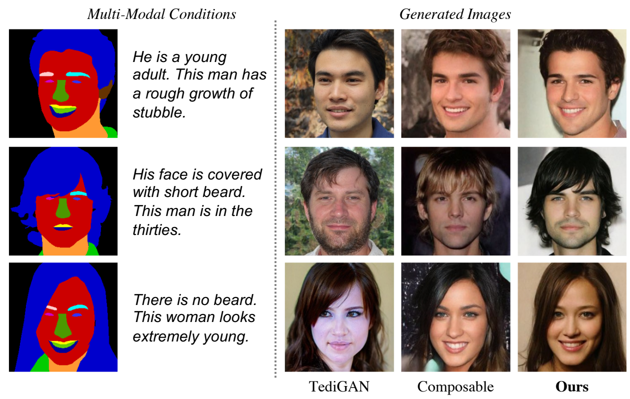

Figure 4. **Qualitative Comparison of Face Generation**. In the second example, TediGAN and Composable fails to follow mask, while ours generates results highly consistent to both conditions.

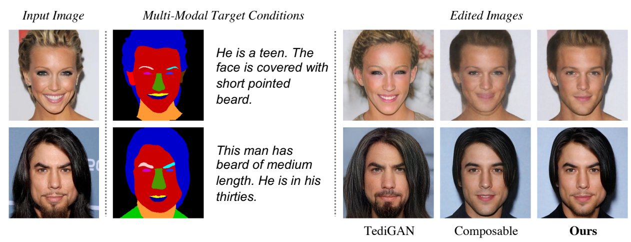

Figure 5. **Qualitative Comparison of Face Editing**. While TediGAN's hair shape is inconsistent with mask, and Composable fails to generate beard according to text, our framework produces results highly consistent with both conditions while maintaining identity.

4. Experiments

4.1. Experimental Setup

Datasets. CelebAMask-HQ [32] consists of manually annotated segmentation masks for each of the 30,000 images in the CelebA-HQ dataset [25]. Each mask has up to 19 classes, including the main facial components such as hair, skin, eyes, and nose, and accessories such as eyeglasses and cloth. CelebA-Dialog [23] provides rich and fine-grained natural language descriptions for the images in CelebA-HQ. In this work, we train the mask-driven diffusion models on CelebAMask-HQ, and the text-driven models on CelebA-Dialog. For *dynamic diffuser*, we simply combine the mask and text for each image without further processing.

Implementation Details. We adopt LDM [48] for our uni-modal diffusion models since it achieves a good balance between quality and speed. Our *dynamic diffuser* is a time-conditional UNet [49], with conditional embeddings injected via cross-attention [62]. The *dynamic diffuser* is $\times 30$ smaller than the pre-trained diffusion model. The detailed architecture and settings are provided in the supplementary material.

4.2. Comparison Methods

TediGAN [67, 68] is a StyleGAN-based face synthesis and manipulation method. It projects both text and mask condi-

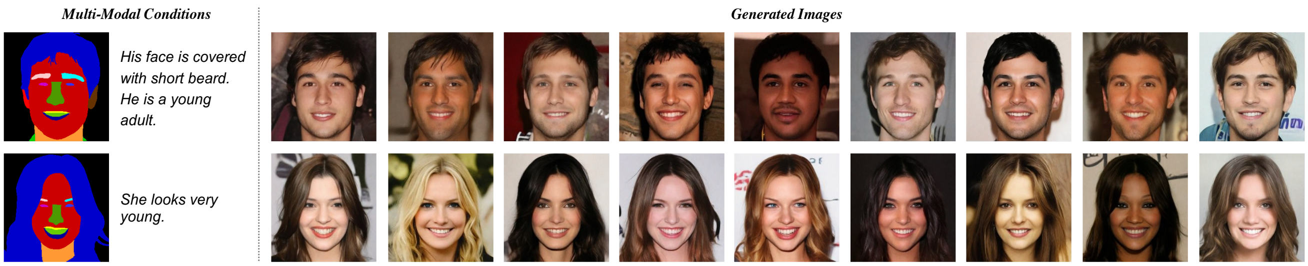

Figure 6. **Diversity of Face Generation**. Given input conditions, our method demonstrates promising synthesis diversity in facial attributes that are not constrained by a specific combination of multi-modal conditions, such as hair color and skin tone.

tions into StyleGAN's $\mathcal{W}+$ latent space, and performs style mixing to achieve multi-modal control.

Composable Diffusion [36] parallely implements two instances of the same text-to-image diffusion model to perform compositional scene generation. We generalize it by replacing the two text-to-image model instances by the text-driven model and the mask-driven model trained using LDM [48]. The two models are composed by averaging the intermediate diffusion results at each step of the *reverse process*.

4.3. Evaluation Metrics

FID. We use Frechet Inception Distance (FID) [18] to measure the quality of images synthesized by different methods. FID computes the feature representation's distance between generated images and real images. A lower FID implies better sample quality.

CLIP Score. CLIP [42] is a vision-language model trained on large-scale datasets. It uses an image encoder and a text encoder to project images and texts to a common feature space, respectively. The CLIP score is computed as the cosine similarity between the normalized image and text embeddings. A higher score usually indicates higher consistency between the output image and the text caption.

Mask Accuracy. For each output image, we predict the segmentation mask using the face parsing network provided by CelebAMask-HQ [32]. Mask accuracy is the pixel-wise accuracy against the ground-truth segmentation. A higher average accuracy indicates better consistency between the output image and the segmentation mask.

User Study. We conduct user study with 25 human evaluators to measure the effectiveness of the methods perceptually. For *face generation*, we randomly sample multi-modal conditions in the validation split of CelebA-HQ Dataset, and then synthesize output images given the conditions. Evaluators are provided with the input conditions and the output images, and they are asked to choose the best image based on 1) image photo-realism, 2) image-text consistency, 3) image-mask consistency. User study is conducted similarly for *face editing*, except that 1) evaluators are also provided with the input image for editing, and 2) they are also asked to assess the level of identity preservation between the edited image and input image.

Table 1. **Quantitative Results of Face Generation**. Compared with TediGAN and Composable Diffusion, our method synthesizes images with better quality (lower FID), and higher consistency with the text and mask conditions.

Method	FID ↓	Text (%) ↑	Mask (%) ↑
TediGAN [67, 68]	157.81	24.27	72.19
Composable [36]	124.62	23.94	76.11
Ours	**111.36**	**24.51**	**80.25**

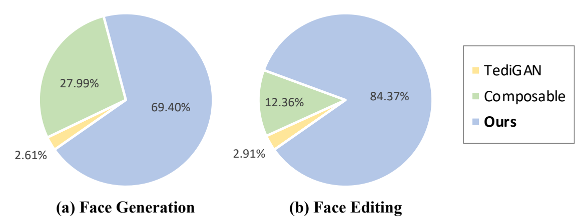

Figure 7. **User Study**. Among the three methods, the majority of users vote for our results as the best for both generation (69.40%) and editing (84.37%), in terms of image quality, condition consistency, and identity preservation.

4.4. Comparative Study

Qualitative Comparison. In Figure 8, we provide the generation results of our *Collaborative Diffusion* under different combinations of texts and masks. It is observed that our framework is capable of synthesizing realistic outputs consistent with the multi-modal inputs, even for relatively rare conditions, such as a man with long hair. We then compare our method with TediGAN [67, 68] and Composable Diffusion [36] in Figure 4. For example, in the second example, TediGAN and Composable Diffusion's results are inconsistent with the mask, while ours is able to maintain both mask and text consistency. We also show in Figure 6 that our method synthesizes images with high diversity without losing condition consistency.

In addition to image generation, it is also observed in

Figure 8. **Face Generation Results**. 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 9 that our framework is able to edit images based on multi-modal conditions while preserving identity. The comparison to existing works in Figure 5 also verifies our effectiveness. For example, in the second example, while TediGAN is unable to synthesize hair consistent to the mask and Composable Diffusion fails to generate beard according to the text, our framework is able to generate results highly consistent to both conditions while maintaining identity.

Quantitative Comparison. As shown in Table 1, our *Collaborate Diffusion* outperforms TediGAN and Composable Diffusion in all three objective metrics. In addition, as depicted in Figure 7, our framework achieves the best result in 69.40% and 84.37% of the time for face generation and editing, respectively. These results verify our effectiveness in both image quality and image-condition consistency.

4.5. Ablation Study

In this section, we visualize the *influence functions*, and show that it is necessary for the *influence function* to be both spatially and temporally varying to facilitate effective collaboration.

4.5.1 Spatial Variations of Influence Functions

As shown in Figure 10, the *influence functions* behaves differently at different spatial regions. For instance, the influence for the mask-driven model mainly lies on the contours of facial regions, such as the outline of hair, face, and eyes, as these regions are crucial in defining facial layout. In contrast, the influence for the text-driven model is stronger at skin regions including cheeks and chin. This is because the attributes related skin texture, such as age and beard length, are better described by text.

To verify its necessity, we remove the spatial variation of *influence functions*. From Table 2 we observe that removing the spatial variance results in deterioration in both output quality and condition consistency. This corroborates our hypothesis that it is important to assign different weights to different modalities at different spatial regions.

4.5.2 Temporal Variations of Influence Functions

In addition to the spatial variation, it is also observed in Figure 10 that the influence from the mask-driven model is stronger at earlier diffusion stages (*i.e.*, larger t), since early stages focus on initializing the facial layout using the mask-

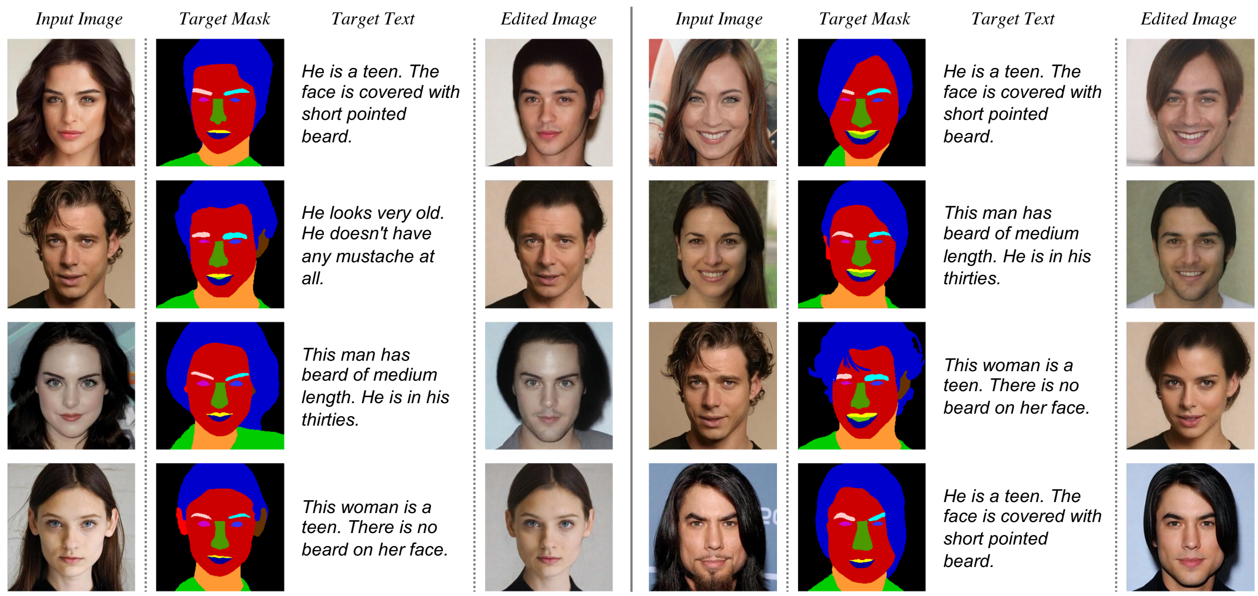

Figure 9. **Face Editing Results**. Given the input real image and target conditions, we display the edited image using our method.

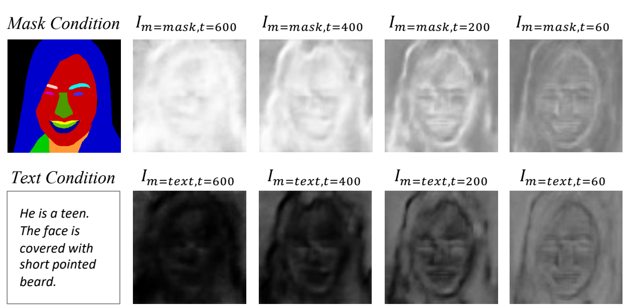

Figure 10. **Visualization of Influence Functions**. The *influence functions* vary spatially at different face regions, and temporally at different diffusion timesteps.

Table 2. **Ablation Study**. Temporal or spatial suppression in influence variation introduces performance drops, which shows the *necessity of influence functions' spatial-temporal adaptivity*.

Method	FID ↓	Text (%) ↑	Mask (%) ↑
Ours w/o Spatial	117.81	24.36	80.08
Ours w/o Temporal	117.34	24.48	77.07
Ours	**111.36**	**24.51**	**80.25**

driven model's predictions. At later stages, the influence from the text-driven model increases as the textural details (*e.g.*, skin wrinkles and beard length) are instantiated using information from the text.

We remove the temporal variation in the *influence functions* to demonstrate its importance. It is shown in Table 2 that both the quality and condition consistency drop without the temporal variation.

5. Conclusion

With generative AI gaining increasing attention, multi-modal conditioning becomes an indispensable direction to unleash creativity and enable comprehensive controls from human creators. In this work, we take the first step forward and propose ***Collaborative Diffusion***, where pre-trained uni-modal diffusion models collaboratively achieve multi-modal face generation and editing without being re-trained. With our *dynamic diffuser*, this framework could be used to extend arbitrary uni-modal approach to the multi-modal paradigm through predicting the relative influence of different modalities. We believe our idea of synergizing uni-modal models for multi-modal tasks would be a good inspiration for future works in various domains such as motion and 3D generation.

Limitations and Future Work. Since our framework focuses on exploiting pre-trained diffusion models, our performance is dependent on the capability of each model. An orthogonal direction is to train each collaborator on large-scale datasets for performance gain.

Potential Negative Societal Impacts. The facial manipulation capabilities of *Collaborative Diffusion* could be applied maliciously on real human faces. We advise users to apply it only for proper recreational purposes.

Acknowledgement. We would like to thank Liangyu Chen and Bo Li for proofreading. This study is supported by NTU NAP, MOE AcRF Tier 1 (2021-T1-001-088), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. SegDiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 2

[2] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021. 2

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 1

[4] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022. 2

[5] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *ECCV*, 2022. 2

[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022. 2

[7] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Adaptively-realistic image generation from stroke and sketch with diffusion model. In *WACV*, 2023. 1

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 1, 2

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. CogView: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021. 2

[11] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis. In *NeurIPS*, 2021. 2

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[13] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 2

[14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[15] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *NeurIPS*, 2022. 2

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 1, 2

[17] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022. 2

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 2

[22] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *JMLR*, 2005. 2

[23] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-Edit: Fine-grained facial editing via dialog. In *ICCV*, 2021. 5

[24] Yuming Jiang, Shuai Yang, Haonan Qju, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM TOG*, 2022. 2

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 5

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 2

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2

[29] 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. 1, 2, 5

[30] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 1, 2

[31] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *WACV*, 2021. 2

[32] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 2, 5, 6

[33] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. ManiGAN: Text-guided image manipulation. In *CVPR*, 2020. 2

[34] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, 2019. 2

[35] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. StyleT2I: Toward compositional and high-fidelity text-to-image synthesis. In *CVPR*, 2022. 2

[36] 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. 4, 6

[37] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *WACV*, 2023. 2

[38] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2

[39] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[40] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2

[41] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[45] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 2

[46] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *NeurIPS*, 2016. 2

[47] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *CVPR*, 2021. 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 5, 6

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4, 5

[50] 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. *arXiv preprint arXiv:2208.12242*, 2022. 2

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[52] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2022. 2

[53] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, 2020. 2

[54] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE TPAMI*, 2020. 2

[55] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[56] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 2

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[58] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1, 2

[59] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *IJCAI*, 2019. 2

[60] Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. AnyFace: Free-style text-to-face synthesis and manipulation. In *CVPR*, 2022. 2

[61] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[62] 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. 5

[63] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[64] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 2011. 2

[65] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 1, 2

[66] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 1

[67] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. TediGAN: Text-guided diverse face image generation and manipulation. In *CVPR*, 2021. 2, 5, 6

[68] 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. 2, 5, 6

[69] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 2

[70] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2

[71] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE TPAMI*, 2018. 2