Multivariate Diffusion Transformer with Decoupled Attention for High-Fidelity Mask-Text Collaborative Facial Generation*

Yushe Cao[1], Dianxi Shi[2] †, Xing Fu[3], Xuechao Zou[4], Haikuo Peng[5], Xueqi Li[2], Chun Yu[1] Junliang Xing[1]‡,

[1]Tsinghua University
[2]Intelligent Game and Decision Lab
[3]Ant Group
[4]Beijing Jiaotong University
[5]National University of Defense Technology
cao-ys23@mails.tsinghua.edu.cn, dxshi@nudt.edu.cn, fux008@gmail.com, xuechaozou@foxmail.com, penghk@nudt.edu.cn, lixueqilxqnudt@nudt.edu.cn, chunyu@tsinghua.edu.cn, jlxing@tsinghua.edu.cn

Abstract

While significant progress has been achieved in multimodal facial generation using semantic masks and textual descriptions, conventional feature fusion approaches often fail to enable effective cross-modal interactions, thereby leading to suboptimal generation outcomes. To address this challenge, we introduce MDiTFace—a customized diffusion transformer framework that employs a unified tokenization strategy to process semantic mask and text inputs, eliminating discrepancies between heterogeneous modality representations. The framework facilitates comprehensive multimodal feature interaction through stacked, newly designed multivariate transformer blocks that process all conditions synchronously. Additionally, we design a novel decoupled attention mechanism by dissociating implicit dependencies between mask tokens and temporal embeddings. This mechanism segregates internal computations into dynamic and static pathways, enabling caching and reuse of features computed in static pathways after initial calculation, thereby reducing additional computational overhead introduced by mask condition by over 94% while maintaining performance. Extensive experiments demonstrate that MDiTFace significantly outperforms other competing methods in terms of both facial fidelity and conditional consistency.

1 Introduction

High-fidelity facial synthesis (Ning et al. 2023; Chen et al. 2024b; Melnik et al. 2024; Wang et al. 2025a) holds transformative potential across digital identity systems, adversarial security, and immersive media applications (Terhörst et al. 2023; Rehaan, Kaur, and Kingra 2024; Diao et al. 2025). While recent advances in generative modeling, such as GANs and diffusion models, have enabled photorealistic facial generation, deploying these technologies in practice remains hindered by in fine-grained controllability, posing challenges in precisely generating facial images that align with user expectations.

Modern conditional generation frameworks have transitioned from unimodal to multimodal paradigms to tackle these challenges. Initially, text-driven methods constructed end-to-end pipelines to transform semantic descriptions into facial attributes (Kim, Kwon, and Ye 2022; Zhu and Mu 2023). However, they encountered difficulties in achieving precise spatial localization (e.g., pose, geometry) due to the inherent ambiguity of language. Subsequently, visual modalities such as semantic masks (Chen et al. 2022; Ergasti et al. 2024) were incorporated to enforce geometric constraints, but this often came at the cost of reduced semantic versatility in attributes like age, gender, or complexion. This dichotomy has spurred research into multimodal collaborative facial synergies (Meng et al. 2025; Sowmya and Meeradevi 2024; Kim et al. 2024) that leverage the expressive capacity of natural language for high-level attributes and the precision of visual cues for low-level spatial control.

Existing state-of-the-art multimodal facial synthesis methods generally adhere to two primary architectural paradigms: (1) Generative Adversarial Network (GAN)-based frameworks (Du et al. 2023; Kim et al. 2024; Meng et al. 2025), such as StyleGAN variants, which project heterogeneous modality conditions into a latent space via multiple encoding networks for subsequent fusion; and (2) diffusion models (Nair, Bandara, and Patel 2023; Huang et al. 2023; Peng et al. 2024; Kim et al. 2024), which integrate unimodal expert diffusion models through dynamic weighting during the inference process. However, these approaches typically process multiple modalities in isolation or employ simplistic feature fusion strategies, resulting in suboptimal multimodal feature interaction—particularly in scenarios that demand collaborative optimization between visual masks and textual descriptions.

*This work was supported in part by the CAAI-Ant Group Research Fund, and in part by the Natural Science Foundation of China under Grant No. 62222606.

†Co-Corresponding Author.
‡Co-Corresponding Author.

In this work, we propose a customized diffusion transformer framework termed MDiTFace, which aims to tackle the challenge of inadequate cross-modal feature interaction in high-fidelity mask-text collaborative facial synthesis. The framework employs a unified tokenization approach to process multimodal conditions from mask and text inputs and projects them into a shared latent space via modality-specific embedders, effectively bridging the gap between heterogeneous modality representations. We design a dedicated multivariate transformer block capable of simultaneously processing multi-conditional token sequences derived from both text and mask modalities. The internal attention computation within this block significantly enhances cross-modal feature interaction capabilities. Additionally, by decoupling the implicit dependency between mask tokens and temporal embeddings, we introduce a novel decoupled attention mechanism. It explicitly splits the internal computational flow into dynamic and static pathways, enabling relevant features in the static pathway to be cached and reused across denoising steps after initial computation. This approach reduces the additional computational overhead introduced by mask condition by over 94% while maintaining model performance.

In summary, our core innovations and contributions are as follows:

- We propose a customized diffusion transformer framework, MDiTFace, which resolves the issue of insufficient multimodal feature interaction in high-fidelity mask-text collaborative facial synthesis, significantly improving the quality of synthesized facial images.
- We design a dedicated multivariate transformer block that can simultaneously process multimodal token sequences from both text and mask inputs, with its internal attention computation substantially facilitating bidirectional and flexible interaction between multimodal features.
- We introduce a novel decoupled attention mechanism that explicitly divides the internal computational flow into dynamic and static pathways. By reusing computational features in the static pathway, it reduces the computational overhead associated with mask tokens by over 94% while preserving model performance..

Comprehensive qualitative and quantitative evaluations demonstrate that MDiTFace achieves state-of-the-art performance in both facial fidelity and multimodal conditional consistency.

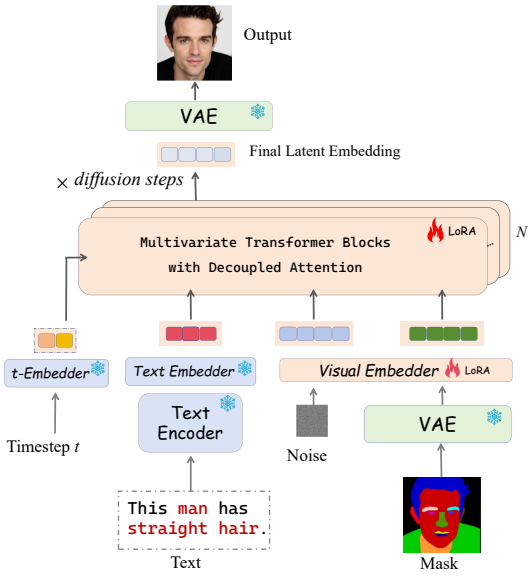


Figure 1: Overall framework of our MDiTFace method.

2 Related Works

2.1 Diffusion Models for Image Generation

Recent advancements in generative modeling have witnessed a paradigm shift from GAN-based architectures (You et al. 2022; Du et al. 2023; Che Azemin et al. 2024) to diffusion models (Po et al. 2024; He et al. 2025; Chang et al. 2025; Huang et al. 2025), driven by their superior training stability and output diversity. Early diffusion models like DDPM (Ho, Jain, and Abbeel 2020) and DDIM (Song, Meng, and Ermon 2020) established foundational sampling pipelines, while subsequent works focused on computational efficiency (Rombach et al. 2022; Xue et al. 2024; Chadebec et al. 2025) and conditional control (Podell et al. 2023; Yoon et al. 2023; Peng et al. 2024). Latent diffusion models (LDM) (Rombach et al. 2022) reduced computational costs by operating in compressed latent spaces, enabling high-resolution synthesis. ControlNet (Zhang, Rao, and Agrawala 2023), ControlNeXt (Peng et al. 2024), and T2I-Adapter (Mou et al. 2024) further enhanced controllability by introducing external conditioning networks, though their multimodal integration is still confined to the addition of independent feature maps.

The integration of vision transformers has driven recent breakthroughs, where diffusion transformers (DiTs) (Peebles and Xie 2023; Wang et al. 2025b) demonstrate superior performance over U-Net architectures (Ronneberger, Fischer, and Brox 2015) through global attention mechanisms, with representative works including FLUX.1 (Labs 2024), OminiControl (Tan et al. 2024, 2025), and OmniGen (Xiao et al. 2025). However, these methods often incur significantly higher computational costs.

2.2 Facial Image Synthesis with Multimodal Conditioning

Facial synthesis techniques (Kim et al. 2023; Melnik et al. 2024; Song et al. 2025) have evolved from unimodal control paradigms to multimodal control paradigms, with recent research efforts focusing on the collaborative integration of various modalities, such as text and semantic masks (Du et al. 2023; Meng et al. 2025). Methods like TediGAN (Xia et al. 2021) and MM2Latent (Meng et al. 2025) leverage the disentangled properties of the $\mathcal{W}$ latent space in StyleGANs to achieve multimodal-driven facial synthesis by fusing latent vectors from multiple modalities. In contrast, diffusion-based approaches, such as GCDP (Park et al. 2023), have

aimed to enhance semantic alignment in synthesized faces; however, they are limited by the restrictive assumptions of joint Gaussian distributions. Additionally, methods like UaC (Nair, Bandara, and Patel 2023) and CoDiffusion (Huang et al. 2023) dynamically integrate multiple unimodal expert models during inference, which results in a more than two-fold increase in computational cost. Meanwhile, DDG (Kim et al. 2024) combines diffusion models with GANs, utilizing ControlNet (Zhang, Rao, and Agrawala 2023; Peng et al. 2024) for semantic mask and text multimodal feature fusion. Still, the expressiveness of the StyleGAN latent space representation constrains the quality of synthesized faces.

3 Method

3.1 Preliminaries

Diffusion transformer (DiT) (Peebles and Xie 2023) successfully breaks through the limitations of local perception by adopting a pure transformer architecture to replace the traditional U-Net backbone network. This achieves an integration of global dependency modeling and multimodal information fusion, thereby significantly enhancing the model's capability to model semantic consistency in complex scenarios. At time step t, the model utilizes stacked tansformer blocks to synchronously process noisy image tokens $\mathbf{X}_t \in \mathbb{R}^{N\times d}$ and text condition tokens $\mathbf{C}_T \in \mathbb{R}^{L\times d}$. Each transformer block incorporates a multi-head attention computation via Equation 1, ensuring in-depth interaction and fusion between visual and semantic information.

$$\text{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are linearly projected from the concatenated tokens $[\mathbf{C}_T;\mathbf{X}_t]$, and d_h denotes the head dimension.

Rotary Positional Embedding. To overcome the translation invariance limitation of traditional absolute position encoding, DiT introduces Rotary Position Encoding (RoPE) (Heo et al. 2024) to explicitly model spatial relationships. Each visual token at position (i,j) is modulated via:

$$\mathbf{X}_{i,j} \leftarrow \mathbf{X}_{i,j}\cdot\mathbf{R}(i,j), \tag{2}$$

where $\mathbf{R}(i,j)$ denotes frequency-domain rotation. Text tokens are assigned (0,0) to maintain semantic coherence.

3.2 Model Design

Figure 1 illustrates the overall framework of the proposed MDiTFace method, which aims to synthesize high-fidelity facial images under the collaborative constraints of textual descriptions and semantic masks. Built upon the FLUX.1 text-to-image generation model (Labs 2024; Labs et al. 2025), MDiTFace employs a unified tokenization strategy to process heterogeneous mask-text conditional inputs, projecting them through modality-specific embedders into a shared latent space where the noise image tokens $\mathbf{X}_t$ resides, thereby eliminating representation disparities across heterogeneous modalities. To achieve synchronous processing and achieve a thorough fusion of multiple conditional token sequences, which include mask tokens C_M and text tokens C_T, we have conducted a comprehensive redesign and subsequent implementation of the multivariate transformer block. These blocks incorporate an innovatively designed decoupled attention mechanism that enables flexible interaction among multimodal tokens while partitioning internal computations into dynamic and static pathways. By eliminating the implicit dependency between mask tokens C_M and temporal embeddings, the relevant features of the static pathway can be cached after the initial computation and reused across denoising steps, thereby significantly reducing redundant computational overhead.

Unified Tokenization. To achieve collaborative facial synthesis combining semantic masks and textual descriptions, the key challenge lies in efficiently integrating semantic mask conditions into the foundational model. An intuitive approach is to directly concatenate the VAE-encoded mask condition $\mathbf{C}_M$ with the noisy image $\mathbf{X}_t$ along the channel dimension, as shown in Equation 3:

$$\mathbf{X}_t \leftarrow \text{Concat}(\mathbf{X}_t, \mathbf{C}_M). \tag{3}$$

However, experimental results indicate that this simple fusion strategy leads to poor mask-condition consistency in synthesized facial images (see Table 3 for details). To address this issue, our proposed MDiTFace method employs a unified tokenization strategy to process multimodal input conditions: The mask is encoded by reusing the VAE encoder, preserving spatial position information while maintaining architectural simplicity. The text is encoded using the pre-trained T5 encoder (Raffel et al. 2020) to capture semantic richness. Subsequently, the encoded features from both modalities are projected into a latent space shared with the noisy image tokens $\mathbf{X}_t$ via modality-specific embedders, bridging representation gaps between heterogeneous modalities. This transformation is formalized in Equation 4:

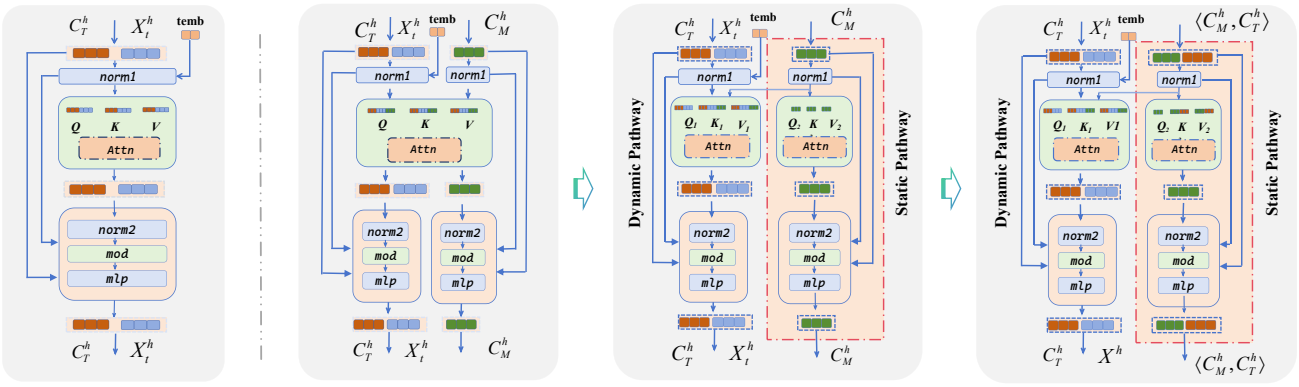
$$\begin{aligned}\mathbf{C}_M &= \text{VisualEmbedder}(\mathbf{VAE}(\text{Mask})) \in \mathbb{R}^{N\times d},\\ \mathbf{C}_T &= \text{TextEmbedder}(\mathbf{T5}(\text{Text})) \in \mathbb{R}^{L\times d}.\end{aligned} \tag{4}$$

To strengthen spatial alignment, we apply the same RoPE to the mask tokens $\mathbf{C}_M$ as to the noisy image tokens $\mathbf{X}_t$, ensuring precise correspondence between them in the spatial dimension. The operation is defined in Equation 5:

$$\mathbf{C}_{M_{i,j}} \leftarrow \mathbf{C}_{M_{i,j}}\cdot\mathbf{R}(i,j). \tag{5}$$

Finally, a joint token sequence $[\mathbf{C}_T;\mathbf{X}_t;\mathbf{C}_M]$ is constructed by concatenating text tokens $\mathbf{C}_T$, noisy image tokens $\mathbf{X}_t$, and mask tokens $\mathbf{C}_M$ along the sequence dimension. This unified tokenization provides a foundational representation for further interaction of multimodal features.

Multivariate Transformer Block. In the vanilla transformer block, its inherent dual-stream token processing paradigm is primarily designed for text-driven image synthesis tasks (as shown in Figure 2)(a), making it difficult to directly accommodate the synchronous processing requirements of multimodal conditional tokens (mask tokens $\mathbf{C}_M$ and text tokens $\mathbf{C}_T$). To address this limitation, we have specifically reconstructed the internal architecture of the transformer block, upgrading its intrinsic dual-stream attention mechanism to a tri-stream attention architecture (as depicted in Figure 2(b)) to support the collaborative processing of multiple conditional tokens.



(a) Vanilla (b) Holistic Tri-stream Attention (c) Hard Decoupled Attention (d) Improved Decoupled Attention

Figure 2: Internal attention design of the multivariate transformer block. (a) The vanilla dual-stream attention in FLUX.1, which exclusively supports text-modal conditioning. (b) Extended holistic tri-stream attention supporting mask-text multimodal conditions at significantly increased computational cost; (c) Hard-decoupled attention with dynamic and static pathways,efficiency improves, but at the cost of performance degradation; (d) Improved decoupled attention restoring mask-to-text perceptual pathways for balanced efficiency and model performance.

Specifically, this module performs linear projections on the inputs of the three modalities separately using independent parameter matrices to generate queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$). The calculation process is as follows:

$$\begin{aligned}\mathbf{Q} &= [\mathbf{W}_q^T\mathbf{C}_T; \mathbf{W}_q^X\mathbf{X}_t; \mathbf{W}_q^M\mathbf{C}_M],\\ \mathbf{K} &= [\mathbf{W}_k^T\mathbf{C}_T; \mathbf{W}_k^X\mathbf{X}_t; \mathbf{W}_k^M\mathbf{C}_M],\\ \mathbf{V} &= [\mathbf{W}_v^T\mathbf{C}_T; \mathbf{W}_v^X\mathbf{X}_t; \mathbf{W}_v^M\mathbf{C}_M].\end{aligned} \tag{6}$$

Through tri-stream attention computation, noisy image tokens $\mathbf{X}_t$, text tokens $\mathbf{C}_T$, and mask tokens $\mathbf{C}_M$ can achieve sufficient dynamic interaction and feature fusion, effectively enhancing the semantic correlation among multimodal features. However, the introduction of multiple conditional tokens inevitably increases the computational complexity during the inference phase, particularly when handling high-resolution facial synthesis tasks, significantly constraining the generation efficiency.

Decoupled Attention Mechanism. To effectively mitigate the increase in computational overhead, we innovatively propose a decoupled attention mechanism. Its initial design is illustrated in Figure 2(c), where the attention computation is decoupled into a dynamic pathway and a static pathway through structural design. Specifically, the dynamic pathway incorporates the perception of mask features while retaining the ability to be modulated by timestep embedding. The static pathway focuses solely on the self-attention computation of mask tokens $\mathbf{C}_M$ and decouples its dependency on the diffusion timesteps. It only needs to be computed once at the initial stage of diffusion, and the cached features can be directly reused in subsequent denoising steps, thereby significantly reducing computational complexity. However, experiments (as shown in Table 4) reveal that this hard-decoupled design disrupts the effective perception of text tokens $\mathbf{C}_T$ by mask tokens $\mathbf{C}_M$, leading to a significant decline in mask-condition consistency of the synthesized facial images.

To address this deficiency, we optimize and improve the structure of the decoupled attention mechanism, as depicted in Figure 2(d). We use the concatenated sequence of multimodal condition tokens $\langle \mathbf{C}_M, \mathbf{C}_T \rangle$ as the input for static pathway computation, thereby restoring the perception from mask tokens $\mathbf{C}_M$ to text tokens $\mathbf{C}_T$ and ensuring the integrity and effectiveness of feature interaction. The final computational definitions of the dynamic pathway and static pathway are as follows:

Dynamic pathway (Modulated by timesteps):

$$\begin{aligned}\mathbf{Q_1} &= [\mathbf{W}_q^T\mathbf{C}_T; \mathbf{W}_q^X\mathbf{X}_t],\\ \mathbf{K_1} &= [\mathbf{W}_k^T\mathbf{C}_T; \mathbf{W}_k^X\mathbf{X}_t; \mathbf{W}_k^T\mathbf{C}_M],\\ \mathbf{V_1} &= [\mathbf{W}_v^T\mathbf{C}_T; \mathbf{W}_v^X\mathbf{X}_t; \mathbf{W}_v^T\mathbf{C}_M].\end{aligned} \tag{7}$$

$$\text{Attn}\left([\mathbf{C}_T; \mathbf{X}_t]\right) = \text{Softmax}\left(\frac{\mathbf{Q_1}\mathbf{K_1}^\top}{\sqrt{d_h}}\right)\mathbf{V}_1, \tag{8}$$

Static pathway (Cache and reuse after computation):

$$\begin{aligned}\mathbf{Q_2} &= [\mathbf{W}_q^T\mathbf{C}_M; \mathbf{W}_q^T\mathbf{C}_T],\\ \mathbf{K_2} &= [\mathbf{W}_k^T\mathbf{C}_M; \mathbf{W}_k^T\mathbf{C}_T],\\ \mathbf{V_2} &= [\mathbf{W}_v^T\mathbf{C}_M; \mathbf{W}_v^T\mathbf{C}_T].\end{aligned} \tag{9}$$

$$\text{Attn}([\mathbf{C}_M; \mathbf{C}_T]) = \text{Softmax}\left(\frac{\mathbf{Q_2}\mathbf{K}_2^\top}{\sqrt{d_h}}\right)\mathbf{V}_2. \tag{10}$$

3.3 Computational Analysis

We conduct a computational complexity analysis of the attention calculations across different architectures depicted in Figure 2. Taking $\mathbf{X}_t, \mathbf{C}_M \in \mathbb{R}^{N\times d}$ and $\mathbf{C}_T \in \mathbb{R}^{L\times d}$ as examples, where N denotes the number of image tokens and

L represents the number of text tokens. Suppose a total of T denoising timesteps are executed during inference. The computational complexity of the architecture in Figure 2(a) is $\mathcal{O}(T \cdot (N+L)^2)$. Due to the introduction of mask tokens $\mathbf{C}_M$, the computational complexity of the architecture in Figure 2(b) increases significantly to $\mathcal{O}(T \cdot (2N+L)^2)$, which becomes infeasible when N is large. In contrast, our optimized decoupled attention mechanism in Figure 2(d) benefits from the fact that computations on the static pathway need to be performed only once, reducing its computational complexity to $\mathcal{O}\big(T \cdot (N+L) \cdot (2N+L) + (N+L)^2\big)$. Given that in high-resolution facial synthesis scenarios where $N \gg L$, this design optimization can reduce the overall computational load by nearly 50%.

3.4 Training Strategy

During the training phase, we adopt a lightweight adaptation architecture and optimize the model with the aid of flow matching loss, while introducing a stochastic condition dropout mechanism to enhance the model's performance and generalization capability.

Training Objective. We employ a lightweight adaptation architecture, which only requires low-rank adaptation (LoRA) (Hu et al. 2021) fine-tuning on the visual embedder and some linear layers within the multivariate transformer blocks. This approach results in training fewer than 0.1% additional parameters, thereby significantly reducing the computational cost of training. The model is optimized using the flow matching loss, with the specific formula as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\epsilon,X_t,C_T,C_M}\left[\|v_\theta(X_t, C_T, C_M, t) - \mu_t(X_t)\|_2^2\right], \tag{11}$$

where, θ denotes the trainable parameters, $v_\theta(*)$ signifies the predicted velocity field and $\mu_t(*)$ represents the target velocity field.

Stochastic Condition Dropout. To further augment the generalization ability of our model, enabling it to accommodate both multi-condition collaborative facial synthesis (integrating both mask and text inputs) and single-condition facial synthesis (utilizing either text or mask input), we incorporate a stochastic condition-dropping mechanism during the training phase. Specifically, for either the text condition or the mask condition, we randomly drop them with a predefined probability p (set to $p = 0.1$ in our experiments). The detailed implementation can be found in Equation 12.

$$C_{T/M} = \begin{cases} \phi & \text{with probability } p, \\ C_{T/M} & \text{with probability } 1-p. \end{cases} \tag{12}$$

4 Experiments

4.1 Experimental Setup

In this section, we provide a comprehensive introduction to the specific experimental settings, encompassing datasets, evaluation metrics, and implementation details.

Datasets. MM-CelebA (Lee et al. 2020) is a large-scale, multimodal facial synthesis benchmark dataset widely used for evaluation. It comprises 30,000 highly diverse, high-resolution real-world facial images. Each image is accompanied by 10 distinct textual descriptions and a manually annotated semantic mask. The semantic masks cover 19 semantic categories, including major facial components such as hair, skin, eyes, and nose, as well as accessory categories like glasses and clothing. For model training and testing, the MM-CelebA dataset is split into training and test sets in a 9:1 ratio. In addition, to further evaluate the generalization capability of our model, we extended the FFHQ-Text dataset (Zhou 2021) into a mask-text multimodal version, denoted as MM-FFHQ, where the semantic masks were obtained using a pre-trained facial parser (Zheng et al. 2022).

Metrics. We employ a range of widely recognized metrics to evaluate the performance of our model, spanning image quality, conditional consistency, and human preference. For image quality,we employ the no-reference TOPIQ metric (Chen et al. 2024a) to evaluate the overall quality of synthesized facial images. Additionally, we utilize the CLIP Maximum Mean Discrepancy (CMMD) to quantify the distributional divergence between synthesized and real facial images in the CLIP feature space (Jayasumana et al. 2024), and the Learned Perceptual Image Patch Similarity (LPIPS) to assess perceptual distortion between synthesized and original images (Zhang et al. 2018). For conditional consistency, we measure semantic alignment between text and images using the CLIP score (Radford et al. 2021), evaluate pixel-level alignment between segmentation masks and images via Mask Accuracy, and quantify structural similarity between synthesized and real facial images using the DINO Structure Distance (DSD) (Tumanyan et al. 2022). Additionally, we employ a pre-trained aesthetic scoring model (Xu et al. 2023) to compute the Image Reward Score (IRS), which serves as an indirect proxy for the human-centric visual appeal of the generated images.

Implementation Details. All experiments were conducted on a server equipped with 8 NVIDIA A100 GPUs, each with 80 GB of memory. We employed parameter-efficient LoRA fine-tuning with a rank of 8, and selected Prodigy (Mishchenko and Defazio 2024) as the optimizer. The base learning rate was set to 1.0, and the weight decay coefficient was configured as 0.01. During training, we used a batch size of 16 and optimized the model for 5,000 steps. For inference, we employed the Flow-Matching Euler Discrete scheduler, with a random seed of 42 and 28 sampling steps.

4.2 Quantitative Evaluation

Table 1 presents the quantitative comparison on the MM-CelebA benchmark dataset, demonstrating that the proposed MDiTFace outperforms all competing methods across all metrics, thereby establishing a new performance benchmark. Specifically, in terms of image fidelity measured by TOPIQ, LPIPS, and CMMD, MDiTFace achieves improvements of 2.59%, 3.27% and 34.33%, respectively, compared to the second-best method. For conditional consistency evaluated using Mask (%), CLIP.T (%), and DSD metrics, MDiTFace shows enhancements of 0.87%, 2.26% and 8.81%, respectively, over the runner-up. This con-

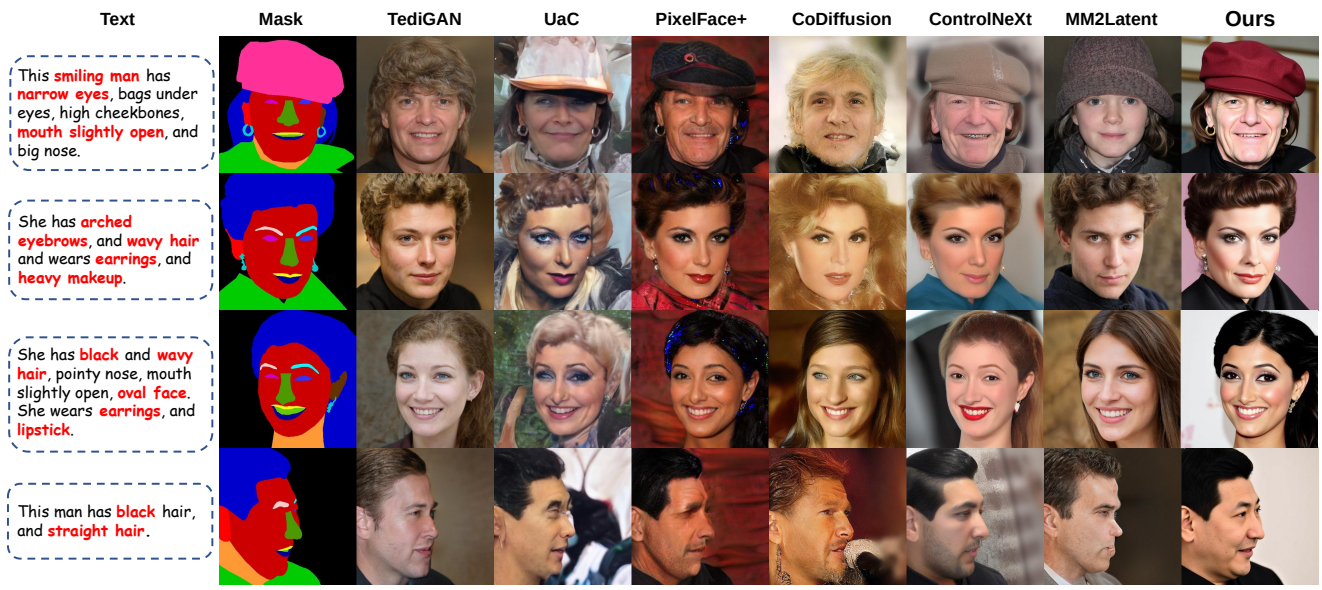


Figure 3: Qualitative comparison with state-of-the-art methods of mask-text collaborative facial generation.

Category	Methods	Paradigm	TOPIQ↑	LPIPS↓	CMMD↓	Mask(%)↑	CLIP.T(%)↑	DSD↓	IRS↑
Comparisons	TediGAN (2021)	GAN	0.7130	0.4885	1.562	87.89	23.90	2.46	-0.1446
	UaC(2023)	Diffusion	0.6027	0.5761	1.982	84.71	25.52	3.41	-0.4001
	PixelFace+ (2023)	GAN	0.5720	0.5640	1.273	93.82	26.16	2.61	0.4608
	CoDiffusion (2023)	Diffusion	0.7381	0.5845	0.734	83.75	24.51	3.22	-0.1265
	ControlNeXt (2024)	Diffusion	0.8037	0.5469	1.042	91.65	25.88	2.66	0.3675
	MM2Latent (2025)	GAN	0.8252	0.5124	1.178	85.05	24.61	3.05	0.2754
Ours	**MDiTFace**	Diffusion	**0.8466**	**0.4725**	**0.482**	**94.64**	**26.75**	**2.38**	**0.6993**

Table 1: Quantitative comparison with state-of-the-art methods on the MM-CelebA test set. (Bold indicates the best, and underline indicates the second-best.)

firms that MDiTFace generates facial images with not only higher fidelity but also better semantic alignment with multimodal input conditions. Additionally, MDiTFace attains the highest Image Reward Score of 0.6993, proving its superior performance in human-centric visual appeal. Table 2 displays the quantitative evaluation on the MM-FFHQ dataset. MDiTFace continues to maintain leading performance across multiple metrics, demonstrating its robust zero-shot generalization capability. More quantitative experiments are provided in the Supplementary Material.

4.3 Qualitative Comparison

Figure 3 displays the facial images synthesized by the proposed MDiTFace and competing methods under the same multimodal conditions. It is evident that the facial images synthesized by MDiTFace exhibit higher fidelity. Meanwhile, in terms of fine-grained attributes such as earrings, hair color, and hat shape, MDiTFace can better align with the input conditions, as shown in rows 1-3. This is consistent with the results of quantitative evaluation experiments. In contrast, existing methods suffer from various deficiencies. For instance, methods like UaC (Nair, Bandara, and Patel 2023), PixelFace+ (Du et al. 2023), and CoDiffusion (Huang et al. 2023) generate facial images with noticeably inadequate fidelity, resulting in unsatisfactory image quality. Other methods, such as TediGAN (Xia et al. 2021), ControlNeXt (Peng et al. 2024), and MM2Latent (Meng et al. 2025), encounter challenging issues of attribute loss or drift, leading to discrepancies between the synthesized facial images and the expected input conditions. Additionally, when handling the task of facial synthesis under extreme poses, as depicted in row 4, MDiTFace still maintains significant robustness compared to existing methods. More examples can be found in the Supplementary Material.

4.4 User Study

We conducted a user study with 30 participants to estimate choice preferences from a human perspective. We ran-

Category	Methods	Paradigm	TOPIQ↑	LPIPS↓	CMMD↓	Mask(%)↑	CLIP.T(%)↑	DSD↓	IRS↑
Comparisons	TediGAN (2021)	GAN	0.6992	0.6420	1.091	86.81	25.03	3.32	-0.4847
	UaC (2023)	Diffusion	0.6106	0.6752	1.796	86.17	26.97	6.38	-0.8851
	PixelFace+ (2023)	GAN	0.5609	0.6705	1.917	95.02	26.60	3.08	-0.2616
	CoDiffusion (2023)	Diffusion	0.5035	0.6767	2.178	81.78	23.03	3.85	-1.2101
	ControlNeXt (2024)	Diffusion	0.7594	0.6165	1.073	86.52	27.80	3.84	-0.4063
	MM2Latent (2025)	GAN	0.8198	0.6371	**0.437**	84.24	26.37	3.52	-0.1197
Ours	**MDiTFace**	Diffusion	**0.8503**	**0.4969**	1.135	**96.13**	**28.05**	**2.79**	**0.0842**

Table 2: Quantitative comparison with state-of-the-art methods on the MM-FFHQ dataset.

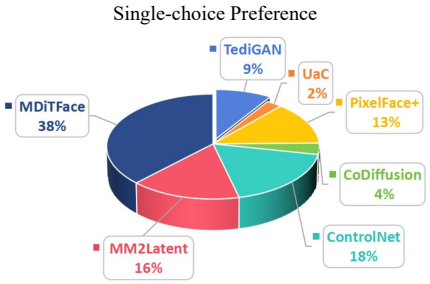


Figure 4: User study. Our method secured the highest proportion of user support.

Process	Mask(%)↑	CLIP.T(%)↑	IRS↑	CMMD↓
Concat	88.73	26.08	0.3699	**0.476**
Ours	**94.64**	**26.75**	**0.6993**	0. 482

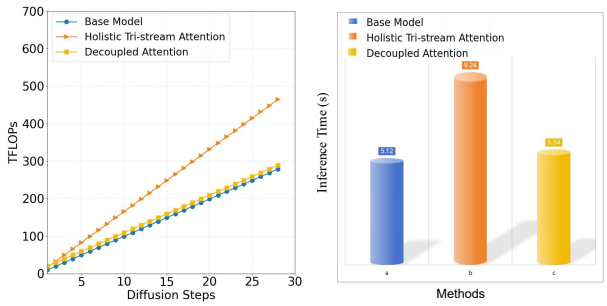
Table 3: Effectiveness comparison of different mask condition processing approaches.

domly selected 50 sets of test data and shuffled the order of facial images synthesized by all the methods. Participants were required to choose one facial image that best met the multi-constraint conditions of masks and text prompts. Figure 4 presents the statistical results. Our method received the highest proportion (38%) of user support, far outperforming the second-place method (ControlNeXt, 18%) and the third-place method (MM2Latent, 16%). These findings further validate the effectiveness of MDiTFace in mask-text collaborative facial synthesis.

4.5 Ablation Studies

We conducted ablation studies to comparatively analyze different mask-handling approaches, validate the effectiveness of our decoupled attention mechanism, and assess the impact of hyperparameter settings in our experiments.

Handling of the mask condition. Table 3 presents a quantitative comparison of different mask condition processing approaches. The method denoted by *Concat* is as described in Equation 3, which simply performs channel-



(a) Cumulative Floating-Point Operations Count (b) Comparison of inference time

Figure 5: Additional computational overhead introduced by mask condition.

wise concatenation of the noisy image latent and the mask. The experimental results demonstrate that this processing approach leads to suboptimal performance in terms of mask condition consistency, achieving a score of only 88.73 on the Mask(%) metric. This represents a 5.91% gap compared to the unified tokenization method we employ. These findings provide compelling evidence for the superiority of using unified tokenization to handle multimodal conditions in mask-text collaborative face generation tasks.

Effectiveness of Decoupled Attention. Table 4 presents a systematic quantitative comparison across the three multivariate transformer architectures depicted in Figure 2(b)-(d). The holistic tri-stream attention mechanism (Figure 2(b)) incurs an additional computational overhead of 185.79 TFLOPs due to its simultaneous processing of mask conditions. The hard-decoupled variant (Figure 2(c)) effectively reduces computational load by eliminating implicit coupling between static pathways and temporal embeddings, but at the expense of weakened mask-text interaction capabilities, as evidenced by a 3.5% absolute decline in the Mask(%) metric (from 94.72 to 91.22). In contrast, our improved decoupled attention mechanism (Figure 2(d)) reconstructs perceptual pathways from mask tokens to text tokens by introducing joint multimodal condition tokens ⟨CM, CT⟩ into static pathways. This approach maintains statistical equivalence in model performance while dramatically reducing mask-induced computational overhead from 185.79

Type	Mask(%)↑	CLIP.T(%)↑	IRS↑	CMMD↓	ΔTFLOPs
(b)	94.72	26.86	0.6967	0.524	185.79
(c)	91.22	26.80	0.7039	0.602	6.64
(d)	94.64	26.75	0.6993	0.482	9.95

Table 4: Comprehensive quantitative analysis on the effectiveness of the proposed decoupled attention mechanism.

r	Mask(%)↑	CLIP.T(%)↑	CMMD↓
4	94.78	26.52	0.477
8	94.64	26.75	0.482
16	94.64	26.70	0.470
32	94.73	26.63	0.494

Table 5: Ablation study on LoRA rank setting.

p	Mask(%)↑	CLIP.T(%)↑	DSD↓
0.1	94.64	26.75	2.38
0.3	94.21	26.68	2.40
0.5	94.21	26.71	2.41
0.7	93.61	26.90	2.54
0.9	84.00	27.02	3.11

Table 6: Ablation study on hyperparameter p of the stochastic condition dropout probability.

TFLOPs to 9.95 TFLOPs (a 94.7% reduction). Figure 5 further validates the significant computational efficiency gains through cumulative floating-point operation statistics and inference time comparisons.

The setting of hyperparameters r and p. We conducted a comparative analysis of model performance under different settings of the LoRA rank (r), with the relevant results presented in Table 5. The experiments revealed that increasing the value of r did not significantly enhance model performance, which, to a certain extent, confirms the robustness of our proposed method. All quantitative scores reported in this paper were obtained when the LoRA rank was set to 8. Additionally, we introduced a stochastic condition dropout strategy during the training phase, aiming to enable the model to simultaneously adapt to facial generation tasks driven by unimodal conditions. We also conducted a comparative analysis of the impact of the hyperparameter p setting on model performance, with the relevant experimental results shown in Table 6. It can be observed that, under a fixed number of fine-tuning steps, as the value of the hyperparameter p increases, the model's learning efficiency for newly introduced mask conditions significantly decreases. This trend is particularly evident in the Mask(%) and DSD metrics when p exceeds 0.5.

4.6 Multimodal compatibility

Although the proposed MDiTFace method is primarily tailored for the mask-text collaborative facial generation task, its intrinsic architecture demonstrates robust generalization capabilities, allowing it to seamlessly accommodate combined inputs from diverse multimodal conditions. For example, multimodal pairings such as depth-text and sketch-text can all be efficiently processed by MDiTFace. As depicted in Figure 6, MDiTFace exhibits exceptional performance in the sketch-text collaborative facial generation task. In this scenario, the sketch provides coarse contour and structural information of the face, while the text serves to refine facial attributes, including skin tone, expressions, and fine-grained feature details. MDiTFace precisely captures key information from both the sketch and text, organically fusing them to generate high-quality, highly realistic facial images that closely align with the input descriptions. This fully underscores its powerful compatibility in handling multimodal condition combinations.

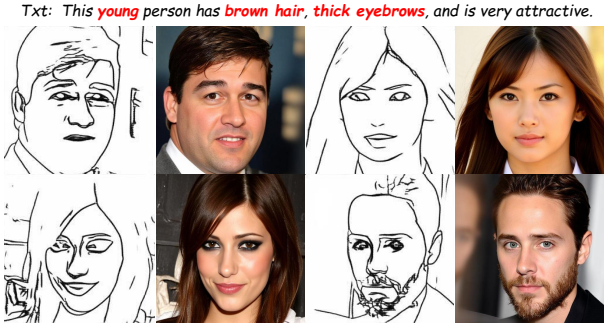


Figure 6: Sketch-text jointly driven facial synthesis.

5 Conclusion

This study proposes MDiTFace, a high-fidelity mask-text collaborative facial synthesis framework built upon a customized diffusion transformer. The framework employs unified tokenization to process multimodal conditions, effectively bridging the representational gaps between heterogeneous modalities. The restructured multivariate transformer blocks is capable of efficiently and synchronously handling both mask and text tokens. Its internally designed decoupled attention mechanism not only facilitates flexible interactions among multimodal features but also explicitly partitions the computational flow into dynamic and static pathways by decoupling implicit dependencies between mask tokens and temporal embeddings. Relevant features in the static pathway only need to be computed once and then cached for reuse across denoising steps. This mechanism reduces redundant computational overhead introduced by mask conditions by over 94% while maintaining model performance, thereby significantly enhancing the efficiency of facial synthesis.

A Supplementary Material

This supplementary material offers a thorough, multi-dimensional analysis of the proposed MDiTFace method, emphasizing its key features and strengths. To verify its generalization capacity, we conduct additional quantitative experiments and provide more visual comparisons with cutting-edge approaches, demonstrating its high fidelity in facial image generation and strict compliance with given constraints. We also conduct experiments to evaluate MDiTFace's performance under multi-modal constraints, with a particular focus on its robustness and diversity in handling complex inputs. For unimodal-driven (text or mask) scenarios, both quantitative and qualitative results confirm its outstanding generalization ability in facial image synthesis. Finally, we discuss the limitations of the proposed MDiTFace method and propose directions for future research.

A.1 Additional Quantitative Experiments

Table 7 presents the quantitative evaluation results of our proposed MDiTFace method for multimodal facial synthesis, with experiments carried out on the MM-FairFace dataset. Built on the FairFace dataset (Karkkainen and Joo 2021) (released by MIT in 2019 to reduce race-, gender-, and age-related biases in face recognition), MM-FairFace inherits its race, gender, and age annotations for each facial image. We constructed MM-FairFace by using a pre-trained facial parser (Zheng et al. 2022) to generate semantic masks for each image, following a methodology similar to that for the MM-FFHQ dataset.

The experimental results clearly show that MDiTFace outperforms most competing methods across the majority of evaluation metrics, demonstrating its strong generalization ability. For instance, on the Mask(%) metric, it achieves a score of 95.95, a 4.74% improvement over the second-best method. Similarly, on the LPIPS and CMMD metrics, it surpasses the second-best method by 11.83% and 9.23%, respectively. These results indicate that MDiTFace not only better understands and aligns multimodal input conditions but also generates facial images with higher fidelity, closely resembling real facial images.

A.2 Qualitative Comparisons with SOTA Methods

Figures 7, 8, and 9 display visual comparisons of typical facial images synthesized by our proposed MDiTFace method and competing state-of-the-art approaches on the MM-CelebA (Lee et al. 2020), MM-FFHQ (Zhou 2021), and MM-FairFace (Karkkainen and Joo 2021) datasets, respectively. Visually, our method yields facial images with superior naturalness and realism, accompanied by a marked reduction in artifacts. Moreover, compared to other methods, MDiTFace shows distinct advantages in semantic mask alignment (reflecting spatial layout) and fine-grained attribute matching (e.g., hair color, accessories, facial makeup).

Notably, on the MM-FairFace dataset (Karkkainen and Joo 2021), many competing methods exhibit facial synthesis collapse due to their failure to interpret racial annotations, resulting in poor generalization. In contrast, MDiTFace shows superior robustness by rationally deducing skin tone from textual racial cues, highlighting its strong generalization in textual semantic comprehension.

A.3 Robustness under Multimodal Combination

To validate the robust performance of the proposed MDiTFace method in facial synthesis under multi-modal condition constraints, we randomly selected m mask conditions and n text conditions, and constructed $m \times n$ condition combinations. Figure 10 displays the facial images generated based on these combinations. When observed row by row, under the same mask condition, as the text conditions vary in aspects such as age, gender, and makeup, MDiTFace can well maintain the alignment characteristics of these semantics. Even in unconventional cases, such as women with extremely short hair or men with extremely long hair, the synthesized facial images remain natural and realistic. When observed column by column, under the same text condition, the pose and contour of the synthesized face change with the variation of mask conditions, which fully demonstrates the method's strong spatial position alignment capability. The above-mentioned observations indicate that MDiTFace exhibits remarkable robustness under condition combinations.

A.4 Diversity in Facial Generation

As shown in Fig. 11, we conduct an in-depth analysis of the facial image generation diversity of the proposed MDiTFace under the collaborative mask-text constraints. The experimental results clearly demonstrate that, while adhering to the spatial layout constraints set by the mask conditions and the semantic attribute constraints specified by the text conditions, our MDiTFace can still maintain a high level of generation diversity across other attribute dimensions of facial images. Specifically, these variable attributes cover multiple aspects, such as variations in facial skin tone and hair color, diverse styles of accessories like earrings and glasses, and flexible background transformations. This notable characteristic holds immense application potential in practical scenarios such as data augmentation and interactive creation.

A.5 Support for Unimodal Driving

During training, we employed a stochastic condition dropout strategy, enabling MDiTFace to theoretically learn facial synthesis using either text or mask conditions alone. This hypothesis was rigorously validated through both quantitative and qualitative experiments.

Quantitatively, as depicted in Tables 8 and 9, we conducted a comparison between MDiTFace and state-of-the-art specialized facial generation models under text-only and mask-only conditions. Although these competing models were specifically designed for particular modalities, MDiTFace achieved comparable performance and outperformed them significantly in terms of TIPIQ, CMMD, and Mask(%) metrics. This highlights its remarkable generalization capability in unimodal-driven synthesis. Qualitatively, Figure 12 showcases facial images generated exclusively from masks

Category	Methods	Paradigm	TOPIQ↑	LPIPS↓	CMMD↓	Mask(%)↑	CLIP.T(%)↑	DSD↓	IRS↑
Comparisons	TediGAN (2021)	GAN	0.6707	0.7257	2.367	71.55	22.44	5.50	-0.2374
	UaC (2023)	Diffusion	0.3647	0.6983	3.636	74.33	22.38	4.66	-1.2595
	PixelFace+ (2023)	GAN	0.3214	0.6987	3.304	91.20	23.91	4.65	-1.2848
	CoDiffusion (2023)	Diffusion	0.3825	0.7556	2.221	53.44	21.83	5.85	1.4839
	ControlNeXt (2024)	Diffusion	0.5483	0.7576	1.138	74.14	**27.05**	5.12	-0.5563
	MM2Latent (2025)	GAN	**0.7552**	0.7179	1.390	67.43	24.05	5.21	-0.0827
Ours	**MDiTFace**	Diffusion	0.7182	**0.6157**	**1.033**	**95.94**	26.40	**4.22**	**0.2424**

Table 7: Quantitative comparison with state-of-the-art methods on the MM-FairFace dataset.

Methods	TOPIQ↑	CMMD↓	CLIP.T(%)↑	IRS↑
Clip2Latent (2022)	0.7861	1.410	27.56	**1.0129**
GCDP (2023)	0.6118	1.456	25.94	0.4563
E3-FaceNet (2024)	0.7319	2.749	**27.94**	0.8150
Ours MDiTFace	**0.8086**	**0.609**	27.25	0.8569

Table 8: Performance comparison with text-only facial generation methods.

Methods	TOPIQ↑	CMMD↓	Mask(%)↑	DSD↓
INADE (2021)	0.6156	1.871	93.60	2.57
E2Style (2022)	0.8007	1.129	90.68	2.36
SemFlow+ (2024)	0.7098	1.767	94.35	**2.30**
Ours MDiTFace	**0.8648**	**0.478**	**94.74**	2.39

Table 9: Performance comparison with mask-only facial generation methods.

or text descriptions. These visual results further demonstrate MDiTFace's proficiency in achieving diverse facial synthesis under single-condition constraints.

A.6 Limitations and Prospects

Although this study has made significant strides in the field of multimodal collaborative facial synthesis by utilizing mask-text inputs, there are still several promising directions for further improvement.

Modality Combination Exploration. Currently, our experimental validation is confined to the fusion of mask and text modalities. Although the proposed MDiTFace architecture is inherently general and theoretically amenable to diverse modality combinations, we have not yet extended our investigations to incorporate other potential pairings, such as sketch-text and keypoint-text. Expanding the scope of modality combinations could unlock new capabilities and applications for our model.

Computational Efficiency Optimization. To improve computational efficiency, we have introduced a structured decoupled attention mechanism. This mechanism segregates internal computations into dynamic and static pathways, contingent on whether timestep embeddings modulate the process. By caching computational features within the static pathway and reusing them across denoising steps, we have effectively reduced redundant computations related to the mask condition by over 94%. Nevertheless, the iterative denoising nature inherent to diffusion models results in MDiTFace's inference time cost remaining substantially higher than that of comparable generative adversarial network-based methods. To mitigate this limitation, future work could explore the integration of techniques such as accelerated sampling or sparse attention to further enhance the generation efficiency of MDiTFace and alleviate the current constraints on its inference speed.

Figure 7: Qualitative comparison with state-of-the-art methods on the MM-CelebA dataset.

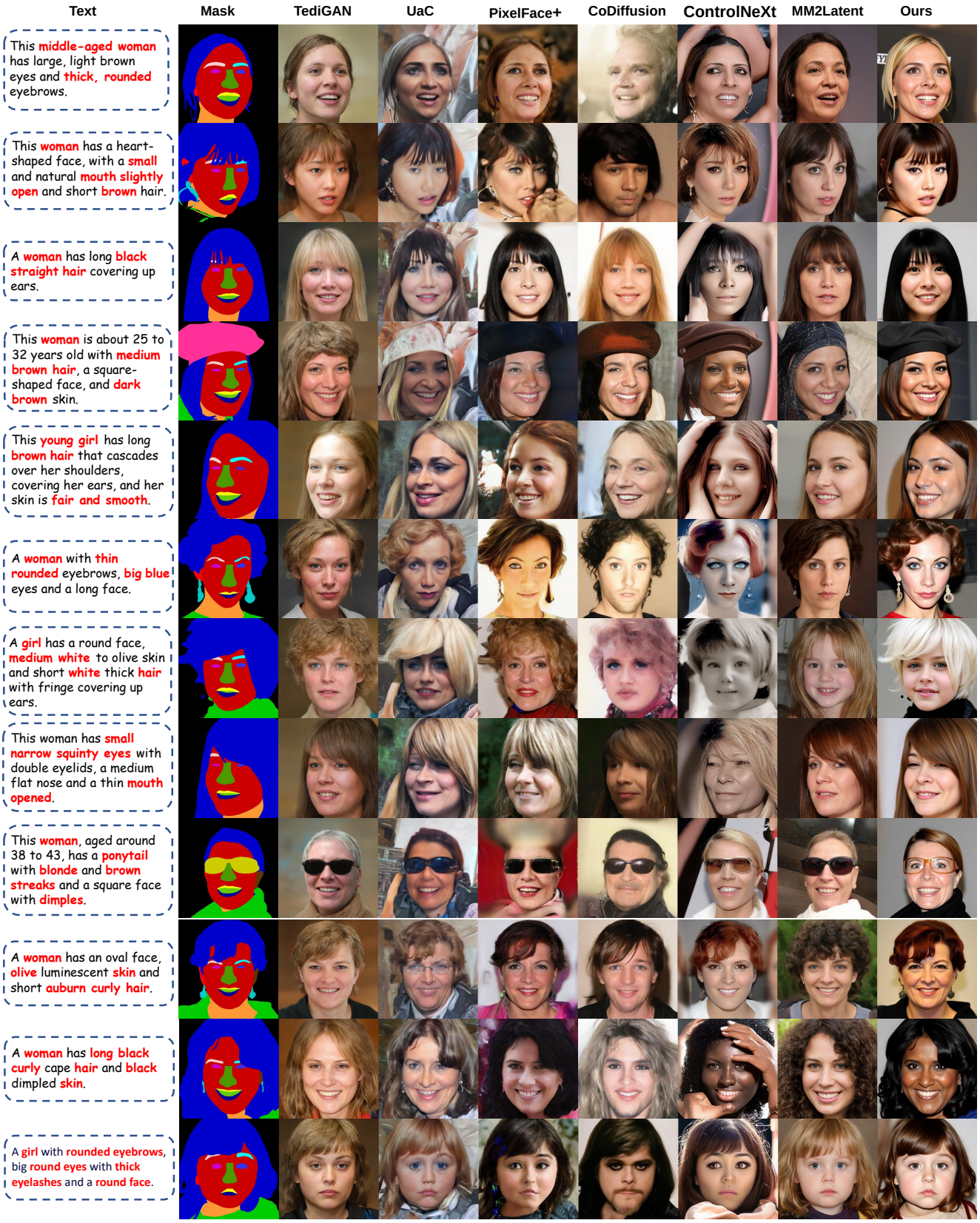


Figure 8: Qualitative comparison with state-of-the-art methods on the MM-FFHQ dataset.

Figure 9: Qualitative comparison with state-of-the-art methods on the MM-FairFace dataset.

Figure 10: Robustness validation of MDiTFace: collaborative facial synthesis examples with arbitrary mask-text combinations.

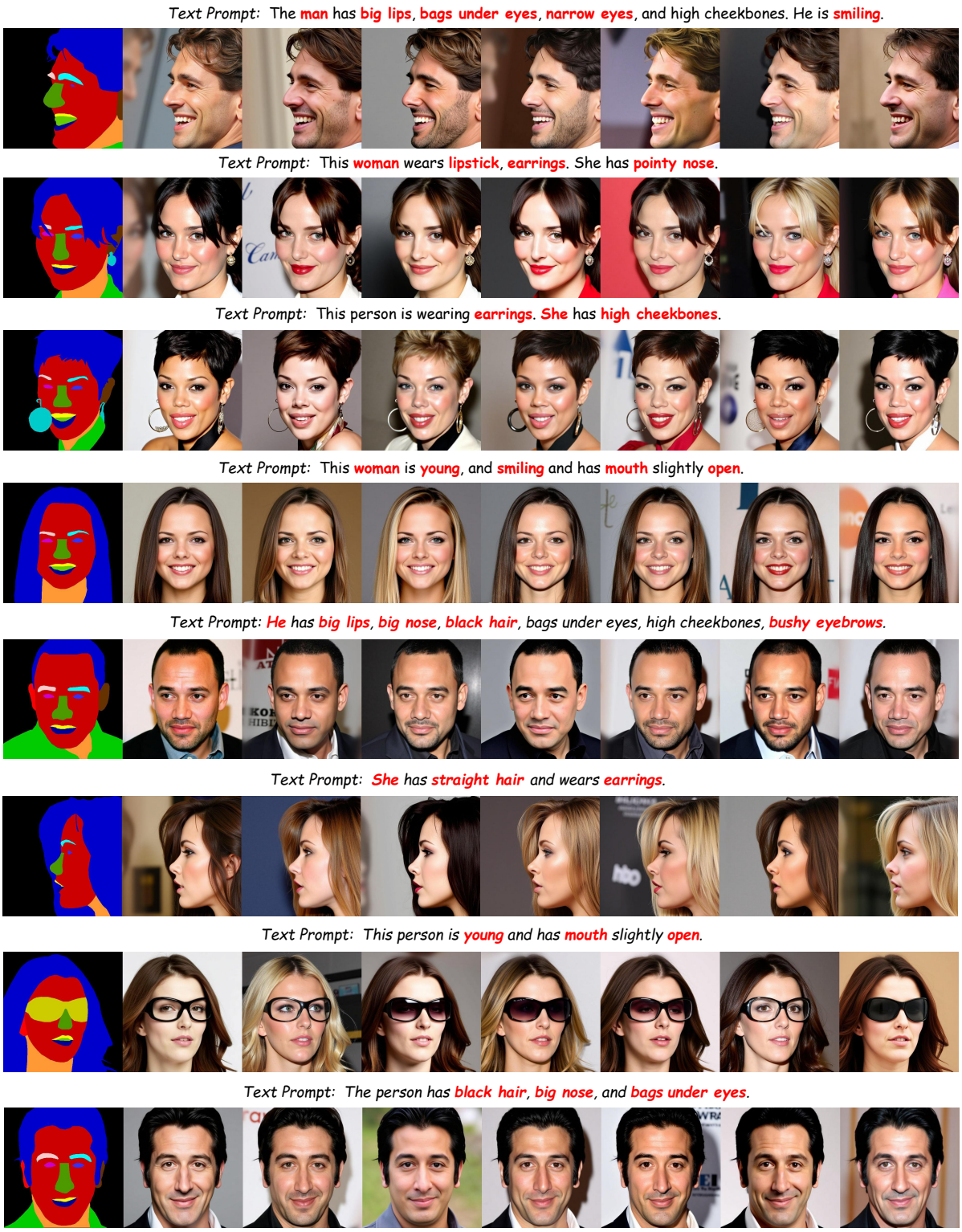
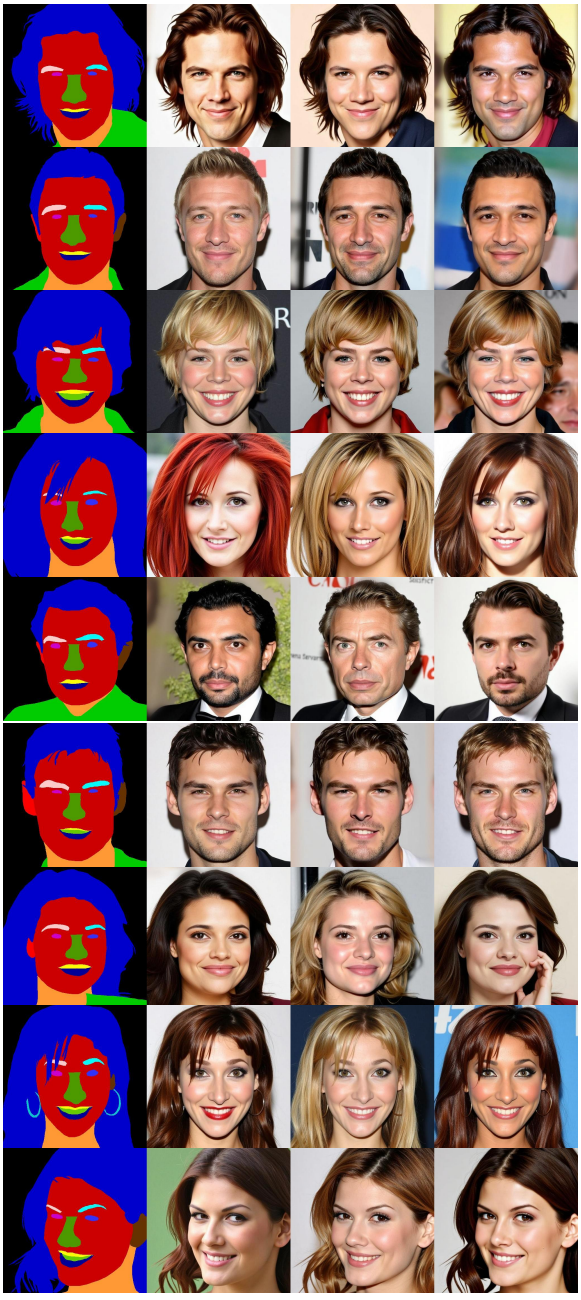


Figure 11: Investigating the generative diversity of MDiTFace under mask-text collaborative constraints.



(a) Mask-Only-Driven

(b) Text-Only-Driven

Figure 12: Examples of unimodal-driven facial synthesis supported by MDiTFace.

References

Chadebec, C.; Tasar, O.; Benaroche, E.; and Aubin, B. 2025. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *AAAI Conference on Artificial Intelligence*, 15686–15695.

Chang, Z.; Koulieris, G. A.; Chang, H. J.; and Shum, H. P. 2025. On the design fundamentals of diffusion models: A survey. *Pattern Recognition*, 111934.

Che Azemin, M. Z.; Mohd Tamrin, M. I.; Hilmi, M. R.; and Mohd Kamal, K. 2024. Assessing the efficacy of StyleGAN 3 in generating realistic medical images with limited data availability. In *International Conference on Software and Computer Applications*, 192–197.

Chen, A.; Liu, R.; Xie, L.; Chen, Z.; Su, H.; and Yu, J. 2022. Sofgan: A portrait image generator with dynamic styling. *ACM Transactions on Graphics*, 41(1): 1–26.

Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024a. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*.

Chen, Z.; Fang, S.; Liu, W.; He, Q.; Huang, M.; and Mao, Z. 2024b. DreamIdentity: Enhanced editability for efficient face-identity preserved image generation. In *AAAI Conference on Artificial Intelligence*, 1281–1289.

Diao, X.; Cheng, M.; Barrios, W.; and Jin, S. 2025. Ft2tf: First-person statement text-to-talking face generation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4821–4830.

Du, X.; Peng, J.; Zhou, Y.; Zhang, J.; Chen, S.; Jiang, G.; Sun, X.; and Ji, R. 2023. Pixelface+: Towards controllable face generation and manipulation with text descriptions and segmentation masks. In *ACM International Conference on Multimedia*, 4666–4677.

Ergasti, A.; Ferrari, C.; Fontanini, T.; Bertozzi, M.; and Prati, A. 2024. Controllable face synthesis with semantic latent diffusion models. In *International Conference on Pattern Recognition*, 337–352.

He, C.; Shen, Y.; Fang, C.; Xiao, F.; Tang, L.; Zhang, Y.; Zuo, W.; Guo, Z.; and Li, X. 2025. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Heo, B.; Park, S.; Han, D.; and Yun, S. 2024. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, 289–305.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Huang, J.; Liu, Y.; Yan, M.; Lv, J.; Liu, J.; Xiong, W.; Zhang, H.; Cao, L.; and Chen, S. 2025. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, Z.; Chan, K. C.; Jiang, Y.; and Liu, Z. 2023. Collaborative diffusion for multi-modal face generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6080–6090.

Jayasumana, S.; Ramalingam, S.; Veit, A.; Glasner, D.; Chakrabarti, A.; and Kumar, S. 2024. Rethinking fid: Towards a better evaluation metric for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9307–9315.

Karkkainen, K.; and Joo, J. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2426–2435.

Kim, J.; Oh, C.; Do, H.; Kim, S.; and Sohn, K. 2024. Diffusion-driven gan inversion for multi-modal face image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10403–10412.

Kim, M.; Liu, F.; Jain, A.; and Liu, X. 2023. Dcface: Synthetic face generation with dual condition diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12715–12725.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5549–5558.

Melnik, A.; Miasayedzenkau, M.; Makaravets, D.; Pirshtuk, D.; Akbulut, E.; Holzmann, D.; Renusch, T.; Reichert, G.; and Ritter, H. 2024. Face generation and editing with stylegan: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3557–3576.

Meng, D.; Tzelepis, C.; Patras, I.; and Tzimiropoulos, G. 2025. MM2Latent: Text-to-facial image generation and editing in GANs with multimodal assistance. In *European Conference on Computer Vision*, 88–106.

Mishchenko, K.; and Defazio, A. 2024. Prodigy: an expeditiously adaptive parameter-free learner. In *International Conference on Machine Learning*, 35779–35804.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, 4296–4304.

Nair, N. G.; Bandara, W. G. C.; and Patel, V. M. 2023. Unite and conquer: Plug & play multi-modal synthesis using diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6070–6079.

Ning, X.; Nan, F.; Xu, S.; Yu, L.; and Zhang, L. 2023. Multi-view frontal face image generation: A survey. *Concurrency and Computation: Practice and Experience*, 35(18): e6147.

Park, M.; Yun, J.; Choi, S.; and Choo, J. 2023. Learning to generate semantic layouts for higher text-image correspondence in text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision*, 7591–7600.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Peng, B.; Wang, J.; Zhang, Y.; Li, W.; Yang, M.-C.; and Jia, J. 2024. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*.

Pinkney, J. N.; and Li, C. 2022. clip2latent: Text driven sampling of a pre-trained StyleGAN using denoising diffusion and CLIP. *British Machine Vision Conference*.

Po, R.; Yifan, W.; Golyanik, V.; Aberman, K.; Barron, J. T.; Bermano, A.; Chan, E.; Dekel, T.; Holynski, A.; Kanazawa, A.; et al. 2024. State of the art on diffusion models for visual computing. In *Computer Graphics Forum*, e15063.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 1–67.

Rehaan, M.; Kaur, N.; and Kingra, S. 2024. Face manipulated deepfake generation and recognition approaches: A survey. *Smart Science*, 53–73.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, 234–241. Springer.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, W.; Ye, Z.; Sun, M.; Hou, X.; Li, S.; and Hao, A. 2025. AttriDiffuser: Adversarially enhanced diffusion model for text-to-facial attribute image synthesis. *Pattern Recognition*, 111447.

Sowmya, B.; and Meeradevi, S. S. 2024. Generative adversarial networks with attentional multimodal for human face synthesis. *Indonesian Journal of Electrical Engineering and Computer Science*, 33(2): 1205–1215.

Tan, Z.; Chai, M.; Chen, D.; Liao, J.; Chu, Q.; Liu, B.; Hua, G.; and Yu, N. 2021. Diverse semantic image synthesis via probability distribution modeling. In *conference on computer vision and pattern recognition*, 7962–7971.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*.

Tan, Z.; Xue, Q.; Yang, X.; Liu, S.; and Wang, X. 2025. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*.

Terhörst, P.; Ihlefeld, M.; Huber, M.; Damer, N.; Kirchbuchner, F.; Raja, K.; and Kuijper, A. 2023. Qmagface: Simple and accurate quality-aware face recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 3484–3494.

Tumanyan, N.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2022. Splicing ViT Features for Semantic Appearance Transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10748–10757.

Wang, C.; Li, X.; Qi, L.; Ding, H.; Tong, Y.; and Yang, M.-H. 2024. Semflow: Binding semantic segmentation and image synthesis via rectified flow. *Advances in Neural Information Processing Systems*, 138981–139001.

Wang, J.; Yu, Y.; Chen, J.; Dai, Q.; and Jiang, Y.-G. 2025a. FaceA-Net: Facial Attribute-Driven ID Preserving Image Generation Network. In *AAAI Conference on Artificial Intelligence*, 7736–7743.

Wang, Z.; Xia, X.; Chen, R.; Yu, D.; Wang, C.; Gong, M.; and Liu, T. 2025b. Lavin-dit: Large vision diffusion transformer. In *Computer Vision and Pattern Recognition Conference*, 20060–20070.

Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Zhang, W.; Yuan, L.; Hua, G.; and Yu, N. 2022. E2Style: Improve the efficiency and effectiveness of StyleGAN inversion. *IEEE Transactions on Image Processing*, 3267–3280.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2256–2265.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; and Liu, Z. 2025. Omnigen: Unified image generation. In *Computer Vision and Pattern Recognition Conference*, 13294–13304.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*.

Xue, S.; Liu, Z.; Chen, F.; Zhang, S.; Hu, T.; Xie, E.; and Li, Z. 2024. Accelerating diffusion sampling with optimized time steps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8292–8301.

Yoon, E. B.; Park, K.; Kim, S.; and Lim, S. 2023. Score-based generative models with Lévy processes. *Advances in Neural Information Processing Systems*, 40694–40707.

You, A.; Kim, J. K.; Ryu, I. H.; and Yoo, T. K. 2022. Application of generative adversarial networks (GAN) for ophthalmology image domains: a survey. *Eye and Vision*, 6.

Zhang, J.; Zhou, Y.; Zheng, Q.; Du, X.; Luo, G.; Peng, J.; Sun, X.; and Ji, R. 2024. Fast text-to-3D-aware face generation and manipulation via direct cross-modal mapping and geometric regularization. In *International Conference on Machine Learning*, 60605–60625.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zheng, Y.; Yang, H.; Zhang, T.; Bao, J.; Chen, D.; Huang, Y.; Yuan, L.; Chen, D.; Zeng, M.; and Wen, F. 2022. General facial representation learning in a visual-linguistic manner. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18697–18709.

Zhou, Y. 2021. Generative adversarial network for text-to-face synthesis and manipulation. In *ACM International Conference on Multimedia*, 2940–2944.

Zhu, J.; and Mu, L. 2023. GrainedCLIP and DiffusionGrainedCLIP: Text-guided advanced models for fine-grained attribute face image processing. *IEEE Access*, 11: 99030–99045.