

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

Anonymous Author(s)

Submission Id: 3972

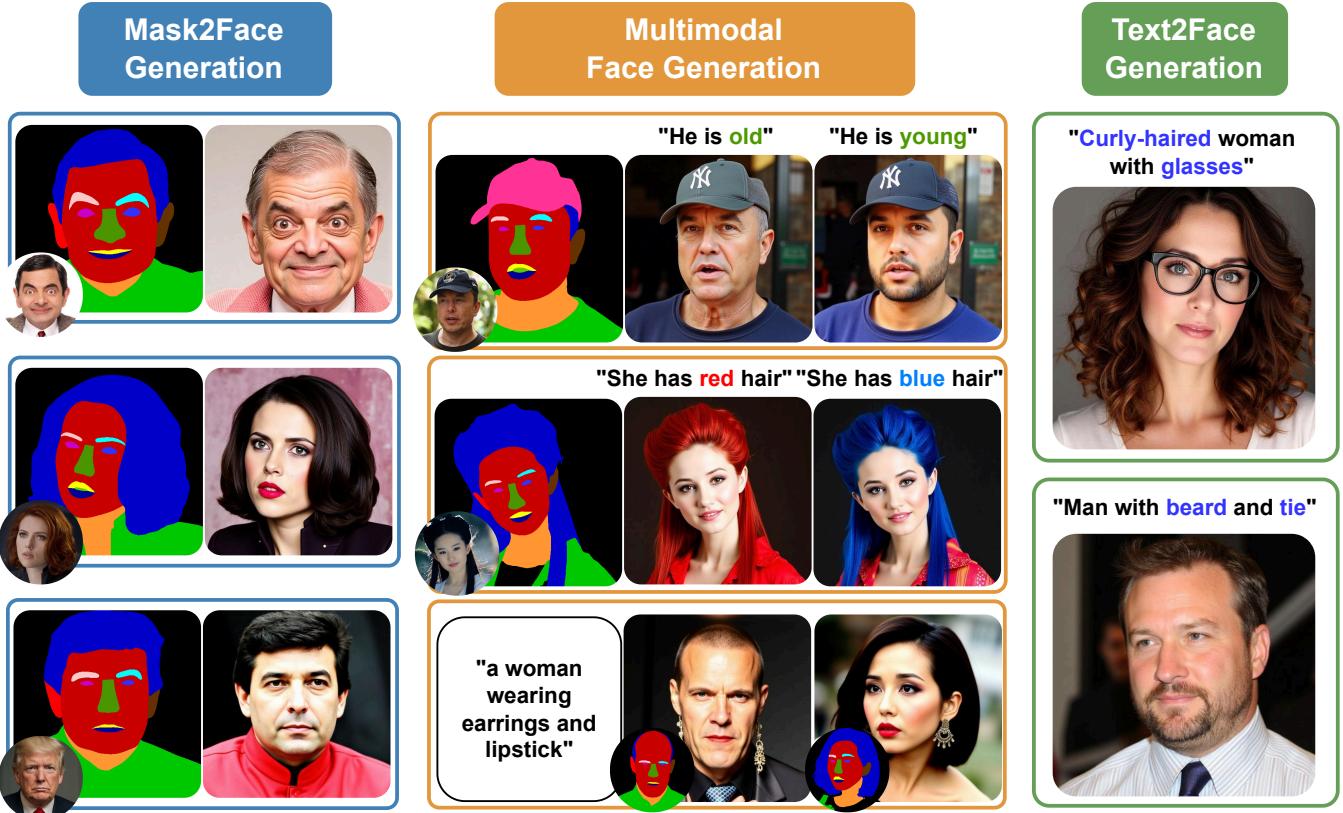


Figure 1: *Mixture of Global and Local Experts with Diffusion Transformer* (Face-MoGLE) is a unified and flexible framework for high-quality and controllable face generation. It supports text-to-face synthesis (left), mask-to-face synthesis (right), and multimodal face generation guided jointly by text and masks (middle). By harmonizing global context modeling with local detail refinement, Face-MoGLE produces highly photorealistic results with enhanced semantic consistency and visual fidelity.

Abstract

Controllable face generation poses critical challenges in generative modeling due to the intricate balance required between semantic controllability and photorealism. While existing approaches struggle with disentangling semantic controls from generation pipelines, we revisit the architectural potential of Diffusion Transformers (DiTs) through the lens of expert specialization. This paper introduces Face-MoGLE, a novel framework featuring: (1) Semantic-decoupled latent modeling through mask-conditioned space factorization, enabling precise attribute manipulation; (2) A mixture of global and local experts that captures holistic structure and region-level semantics for fine-grained controllability; (3) A dynamic gating network producing time-dependent coefficients that evolve with diffusion steps and spatial locations. Face-MoGLE provides a powerful and flexible solution for high-quality, controllable face generation,

with strong potential in generative modeling and security applications. Extensive experiments demonstrate its effectiveness in multimodal and monomodal face generation settings and its robust zero-shot generalization capability. Project page is available at <https://anonymous.4open.science/w/Face-MoGLE-BF75/>.

CCS Concepts

- Information systems → Multimedia content creation;
- Computing methodologies → Image processing.

Keywords

Mixture of Experts, Diffusion Transformer, Face Generation

117 1 Introduction

118 Face generation has become a central task in computer vision,
 119 with wide-ranging applications in digital content creation [21, 37],
 120 virtual reality [15], and human-computer interaction [40]. Beyond
 121 entertainment, this technology holds significant promise in security
 122 and public welfare. For example, it can assist criminal investigations
 123 by synthesizing suspect portraits from forensic sketches or textual
 124 descriptions [29], and support the search for missing persons by
 125 reconstructing plausible appearances from partial visual cues. Generating
 126 realistic and contextually appropriate faces offers profound
 127 potential for numerous practical uses.

128 Face generation involves the core challenges of achieving both
 129 high image fidelity and controllability over various facial attributes.
 130 Controllable face generation, in particular, focuses on generating
 131 facial images that can manipulate specific attributes, such as identity,
 132 expression, or appearance. Early efforts in controllable face
 133 generation were dominated by generative adversarial networks
 134 (GANs) [11, 20, 21], which enable high-resolution synthesis but often
 135 suffer from issues such as mode collapse, training instability, and
 136 limited adaptability to complex or multi-modal conditions. Flow-
 137 based methods [24] offer invertibility and likelihood estimation but
 138 fall short in sample quality. Recently, diffusion models [13, 44, 48]
 139 have become the de facto standard for high-fidelity image synthesis
 140 due to their strong generative performance and compatibility with
 141 conditional guidance, making them particularly effective for con-
 142 trollable face generation. Recently, several works [3, 4, 32, 59, 61, 62]
 143 have explored fine-tuning pre-trained diffusion models to support
 144 conditional inputs such as sketches or semantic maps, further en-
 145 hancing controllability.

146 Despite the success of diffusion models, most state-of-the-art
 147 models rely on U-Net backbones [44], which suffer from two in-
 148 herent limitations. First, the convolutional inductive bias of U-Nets
 149 restricts their ability to model long-range dependencies essential for
 150 holistic facial consistency, while their entangled feature representa-
 151 tions conflate structural and textural information, hindering precise
 152 attribute control. Second, existing methods [17, 33] tightly couple
 153 semantic masks with generation by directly concatenating masks
 154 and latent codes, which propagates mask errors to texture syn-
 155 thesis and limits fine-grained control over local regions. This cou-
 156pling also imposes impractical requirements for pixel-perfect masks
 157 during inference, compromising zero-shot generalization. These
 158 dual limitations motivate our architectural redesign to strengthen
 159 global modeling while decoupling semantic guidance from low-level
 160 synthesis. While pre-trained foundational diffusion transformers
 161 (DiT) [9, 26, 38] have recently demonstrated stronger generalization
 162 and higher fidelity compared to U-Net-based diffusion models, their
 163 application to face generation—especially under complex, multi-
 164 modal conditional settings—remains largely underexplored.

165 To address the limitations of existing controllable face genera-
 166 tion methods, we propose Face-MoGLE, a novel framework that
 167 decouples semantic masks into independent binary components,
 168 each corresponding to a distinct facial attribute such as hair, face
 169 contour, or nose. This semantic decoupling enables precise, region-
 170 specific control and lays the foundation for targeted refinement.
 171 To effectively model both global structure and local detail, we in-
 172 troduce a Mixture of Experts (MoE) design: global experts capture

175 holistic relationships across facial regions (e.g., ensuring spatial
 176 alignment between hair and face), while local experts focus on
 177 fine-grained features within individual regions (e.g., refining the
 178 texture of eyebrows or hair strands). These experts are seamlessly
 179 integrated into the Diffusion Transformer (DiT) backbone, which
 180 provides a powerful foundation for high-fidelity image synthesis
 181 and temporal-aware conditioning throughout the denoising pro-
 182 cess. Together, these components enable Face-MoGLE to achieve
 183 both semantically controllable and visually realistic face generation.
 184 In summary, our main contributions are as follows:

- We propose a unified and modular generation framework based on the Diffusion Transformer (DiT), which decouples semantic masks into binary components to enable structured and disentangled condition modeling.
- We design a Mixture of Global and Local Experts (MoGLE) architecture, where global experts capture holistic facial structures and local experts refine region-specific details for improved semantic alignment and visual fidelity.
- We introduce a diffusion-aware dynamic gating network that adaptively blends expert outputs with spatial and temporal awareness, enabling fine-grained control throughout the denoising process.

197 Experimental results showed that Face-MoGLE significantly out-
 198 performed state-of-the-art (SOTA) controllable face generation
 199 models across multiple benchmarks. Compared to strong diffu-
 200 sion baselines such as PixelFace+ [8] and DDGI [23], our model
 201 achieved better FID scores and higher condition consistency.

202 We extend the FFHQ-Text [64] dataset with high-quality se-
 203 mantic segmentation masks to enable precise region-level control
 204 for multimodal input tasks. Testing on this expanded dataset con-
 205 firms Face-MoGLE’s robust zero-shot generalization. Notably, Face-
 206 MoGLE produces images with high perceptual realism that can
 207 evade SOTA face forgery detectors, underscoring its promise in
 208 generative and security applications. As illustrated in Fig. 1, even
 209 without explicit single-modal training, it delivers strong results in
 210 Mask2Face and Text2Face tasks without retraining or architectural
 211 changes. Overall, by leveraging architectural innovations and ex-
 212 plicit condition disentanglement, Face-MoGLE offers a powerful
 213 and flexible solution for high-quality, controllable face generation.

2 Related Work

2.1 Diffusion Model

218 Diffusion models [6, 13, 35] have emerged as a powerful class of
 219 generative models, demonstrating superior performance to GANs
 220 in image synthesis, which had long been dominated by GAN-based
 221 approaches [22, 27, 52, 56]. Inspired by the physical diffusion pro-
 222 cess [47], these models learn to generate data by reversing a gradual
 223 noising process. In the forward pass, data is progressively corrupted
 224 with Gaussian noise over multiple timesteps, while the model is
 225 trained to recover the original sample through a denoising process.

226 The denoising diffusion probabilistic model proposed by Ho et
 227 al. [13] laid the foundation for this approach with impressive image
 228 synthesis performance. Later, Nichol and Dhariwal proposed the
 229 classifier-free guidance method [14], which enabled conditional
 230 generation without relying on external classifiers. Recent advances
 231 have transitioned from modeling in pixel space to latent

space. LDM [44] employs a U-Net [45] architecture for efficient denoising in a compressed latent space. More recent works replace the convolutional U-Net with vision transformers [7, 38], leveraging their global attention mechanisms and geometry-aware positional encodings to capture spatial dependencies better. These Diffusion Transformers (DiTs) [9, 26, 38, 58] have demonstrated strong scalability, with performance improvements that correlate with model capacity and training compute, establishing them as the new state-of-the-art in diffusion-based generation tasks.

2.2 Face Generation

Face generation has become a cornerstone task in computer vision, with increasing demands for controllability and high-fidelity synthesis. Current approaches can be broadly categorized into GAN-based, diffusion-based, and hybrid methods, and are typically applied to unimodal (e.g., mask or text) or multimodal settings.

In unimodal face generation, mask-to-face and text-to-face are two representative tasks. For mask-to-face generation, methods like MaskGAN [27] use semantic masks to enable interactive and diverse facial manipulation. INADE [50] introduces stochastic sampling on class distributions to enhance diversity, while E2Style [55] focuses on efficient and accurate StyleGAN inversion. SemFlow [54] further unifies image synthesis and segmentation using rectified flow, achieving reversible transformations. In text-to-face generation, clip2latent [39] and GCDP [36] employ CLIP-based guidance and semantic layout generation to improve text-image alignment. E3-FaceNet [60] additionally introduces 3D awareness and geometric regularization to improve realism and view consistency.

In multimodal face generation, the core challenge is aligning diverse conditions—such as text, masks, or sketches—while preserving image fidelity. TediGAN [56] and PixelFace+ [8] enable flexible content creation using both textual and visual inputs. MM2Latent [30] directly maps multimodal signals to the GAN latent space for efficient generation. Diffusion-based methods like Collaborative Diffusion [17] and UaC [33] support plug-and-play multimodal synthesis. DDGI [23] integrates GAN inversion with diffusion features to handle multi-condition face generation.

While prior works have made notable progress in specific settings, they often struggle to balance fine-grained semantic control with high-quality synthesis, especially under zero-shot or compositional generalization. In contrast, our proposed Face-MoGLE achieves strong results across unimodal and multimodal tasks.

2.3 Mixture of Experts

The Mixture of Experts (MoE) model is a neural architecture that enhances scalability and specialization by partitioning the input space among several expert networks. Each expert learns to model a subset of the data distribution, while a gating network dynamically assigns tokens to relevant experts based on input semantics. The concept was first introduced by Hinton et al. [18], who proposed a supervised learning framework in which each expert specializes in a distinct region of the input space. This early work provided a theoretical link between modular neural networks and competitive learning, and laid the foundation for modern sparse expert systems.

Subsequent advances have integrated MoEs into deep learning. Notably, Shazeer et al. [46] introduced the Sparsely-Gated

Mixture-of-Experts, demonstrating large-scale training efficiency. GShard [28] and Switch Transformers [10] refined these ideas, improving training stability and enabling trillion-parameter models. In the vision domain, V-MoE [43] brought sparse expert routing into Vision Transformers [7]. Inspired by these works, our Face-MoGLE leverages a mixture of global and local experts within a diffusion transformer. This enables high-fidelity, controllable face generation by dynamically selecting experts throughout the denoising process.

3 Method

3.1 Overall Pipeline

We build upon the DiT [38] architecture, utilizing FLUX [26] as our foundational model, to introduce *Mixture of Global and Local Experts (MoGLE)*—a simple yet powerful framework for fine-grained controllable face generation. This design accepts multimodal control conditions, aiming to harmonize global context modeling with local detail refinement for semantic masks within a unified model.

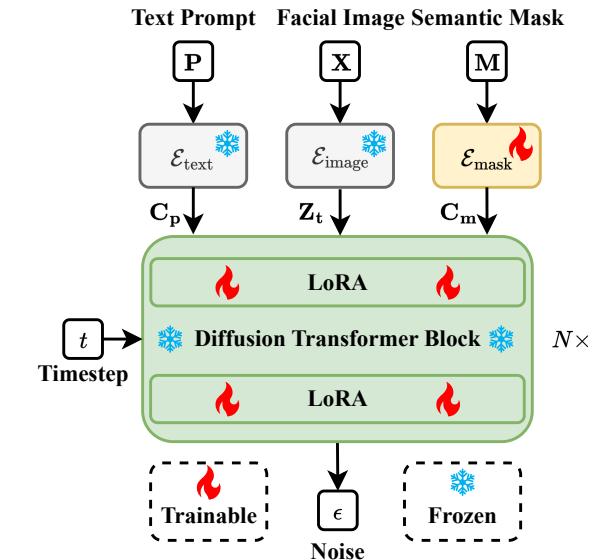


Figure 2: Training pipeline of the diffusion transformer.

Training. The training process is entirely conducted in the *latent space*, where both the forward diffusion and reverse denoising occur. This design avoids direct modeling in pixel space, which significantly improves computational efficiency [44]. As illustrated in Fig. 2, the model is conditioned on multimodal signals, including a text prompt and a semantic mask, and is trained to predict the noise added to the latent image tokens during the diffusion process.

Let the input text prompt be P , the facial image be $X \in \mathbb{R}^{H \times W \times 3}$, and the semantic mask be $M \in \mathbb{R}^{H \times W \times 3}$. These inputs are transformed into token sequences via the following encoders:

$$C_p = \mathcal{E}_{\text{text}}(P), \quad Z = \mathcal{E}_{\text{image}}(X), \quad C_m = \mathcal{E}_{\text{mask}}(M), \quad (1)$$

where $C_p \in \mathbb{R}^{L' \times d'}$, $Z \in \mathbb{R}^{L \times d}$, and $C_m \in \mathbb{R}^{L \times d}$ denote the token sequences output by the text, image, and mask encoders, respectively. L' and L are the token lengths, and d' , d are the embedding dimensions.

The text encoder $\mathcal{E}_{\text{text}}$ jointly leverages CLIP [41] and H5 [42], combining strong generalization with stylistic richness. The image encoder $\mathcal{E}_{\text{image}}$ is based on the encoder of a pretrained VAE [25, 53], which maps the image from pixel space to latent token representations. Both encoders are kept frozen during training.

In contrast, the mask encoder $\mathcal{E}_{\text{mask}}$ is a carefully designed component intended to enhance fine-grained and structured semantic control. We propose a *Mixture of Global and Local Experts* (MoGLE) architecture to obtain rich and flexible token-level representations from semantic masks. This design addresses the lack of spatial controllability in pretrained text-to-image generation DiTs [26, 44]. Details can be found in Section 3.2.

At each training step, a timestep $t \in \{1, \dots, T\}$ is sampled, and Gaussian noise is added to the image tokens:

$$Z_t = \sqrt{\bar{\alpha}_t} Z + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (2)$$

where $\bar{\alpha}_t$ denotes the cumulative product of the forward noise schedule. The resulting noisy tokens $Z_t \in \mathbb{R}^{L \times d}$ are fed into the denoising network. To improve robustness, we apply a drop probability of 0.1 independently to each condition (i.e., text prompt or semantic mask) during training, following previous works [51, 58, 59]. This allows the model to gracefully handle cases where one or both modalities are absent (i.e., set to \emptyset), thereby enabling flexible and controllable face generation.

Our denoising module is a diffusion transformer composed of N stacked blocks, each consisting of a frozen transformer backbone and trainable Low-Rank Adaptation (LoRA) [16, 51] modules for efficient fine-tuning:

$$\hat{\epsilon} = f_\theta(Z_t, t, C_p, C_m), \quad (3)$$

where f_θ denotes the denoising network, with only the LoRA modules updated during training. The model is optimized by minimizing the mean squared error between the predicted and true noise:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{t, Z, \epsilon} [\|\hat{\epsilon} - \epsilon\|_2^2]. \quad (4)$$

This objective guides the model to iteratively denoise latent image tokens while attending to textual and semantic signals.

Sampling. During inference, we adopt an improved sampling procedure [9, 26, 51] to iteratively denoise latent image tokens, initialized from pure Gaussian noise. The text prompt and semantic mask are encoded the same way as during training, using the frozen text encoder and the trained mask encoder, respectively. These condition tokens guide the denoising process toward generating semantically faithful and structurally aligned images. To support flexible generation, our model allows either condition to be dropped at test time by setting it to an empty input. This mirrors the condition drop strategy used during training and enables diverse use cases, such as semantic face synthesis without text, or text-to-face generation from text alone. After denoising completes, the generated latent tokens are decoded into the final image using the pretrained VAE [25, 53] decoder. The output image reflects high-level semantics and spatial structure derived from the conditioning inputs, enabling high-quality and controllable face generation.

3.2 Mixture of Global and Local Experts

Global and Local Experts. To obtain expressive and controllable semantic representations from the input mask, we design a Mixture of Global and Local Experts architecture, as shown in Fig. 3. The motivation is twofold: (1) to extract global structural priors from the full-face layout; and (2) to model region-specific semantics for enhanced controllability and fidelity in face generation.

Given a semantic mask $M \in \mathbb{R}^{H \times W \times 3}$, we first decouple it into n binary masks $\{M^{(i)}\}_{i=1}^n$, each representing a semantic region (e.g., face, hair, nose). All masks are passed through a shared frozen VAE encoder \mathcal{E}_{VAE} [25, 53], producing a sequence of latent tokens:

$$C_m^{(i)} = \mathcal{E}_{\text{VAE}}(M^{(i)}) \in \mathbb{R}^{L \times d}, \quad i = 0, 1, \dots, n \quad (5)$$

where $M^{(0)}$ denotes the full mask, used to derive global context. Each token sequence $C_m^{(i)}$ is processed by its corresponding expert:

$$C_m^{(i)\prime} = \text{Expert}_i(C_m^{(i)}) \in \mathbb{R}^{L \times d}. \quad i = 0, 1, \dots, n \quad (6)$$

The global expert captures high-level spatial priors, while local experts focus on fine-grained, region-specific semantics. This cooperative modeling enables the system to maintain structural consistency while improving the fidelity of generated faces.

Dynamic Gating Network. To dynamically integrate expert outputs across the diffusion process, we introduce a diffusion-aware gating network g_θ , illustrated in Fig. 4. This module takes the current noisy latent tokens Z_t , a learned timestep embedder $\mathcal{E}_{\text{time}}(t)$, and the global mask token $C_m^{(0)}$, and produces normalized weights:

$$\begin{aligned} [\omega_g^{(t)}, \omega_1^{(t)}, \dots, \omega_n^{(t)}] &= g_\theta(Z_t, \mathcal{E}_{\text{time}}(t), C_m^{(0)}) \\ \text{s.t. } \omega_g^{(t)} + \sum_{i=1}^n \omega_i^{(t)} &= 1, \quad \omega_g^{(t)}, \omega_i^{(t)} \in [0, 1] \end{aligned} \quad (7)$$

where $\omega_g^{(t)}$ and $\omega_i^{(t)}$ denote the spatial weight maps for the global expert and the i -th local expert at timestep t , respectively. Unlike static fusion, our gating mechanism produces spatially varying weights that evolve during the denoising process. The final semantic embedding is computed as:

$$C_m = \omega_g^{(t)} \cdot C_m^{(0)\prime} + \sum_{i=1}^n \omega_i^{(t)} \cdot C_m^{(i)\prime}. \quad (8)$$

To better understand the gating behavior, we visualize the spatial weight maps predicted for both global and selected local experts in Fig. 5. These maps reveal that different semantic regions are adaptively emphasized at stages of the diffusion process, validating the gating network's semantic awareness and spatial adaptivity.

4 Experiments

4.1 Datasets

In this study, we employ two benchmark datasets: MM-CelebA-HQ [56], an enriched version of CelebAMask-HQ [27] featuring high-resolution facial images annotated with attributes, utilized for both training and evaluation; and FFHQ-Text [64], a meticulously curated subset of FFHQ [22], comprising high-quality images of female faces accompanied by nuanced textual descriptions. We further refine this dataset into a multimodal variant, referred to as MM-FFHQ-Female, designed explicitly for zero-shot evaluation. In

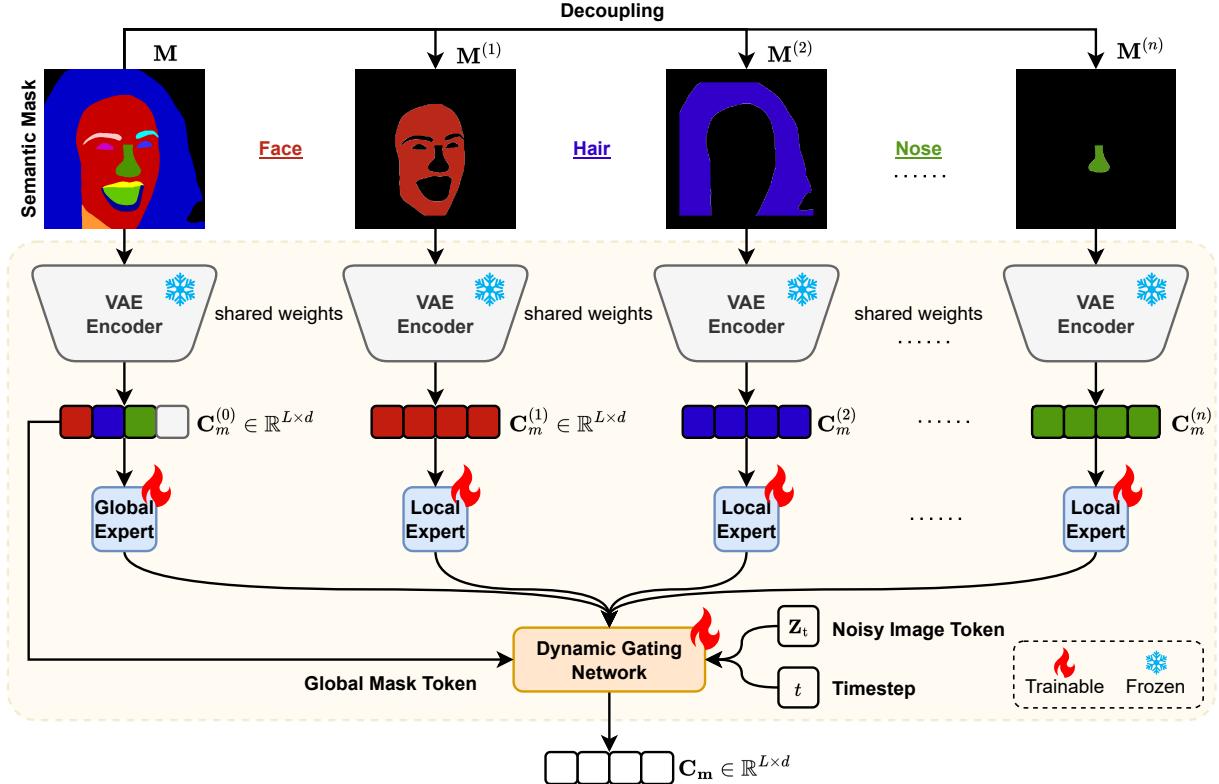


Figure 3: Architecture of the Mixture of Global and Local Experts (MoGLE) designed for semantic mask embedding.

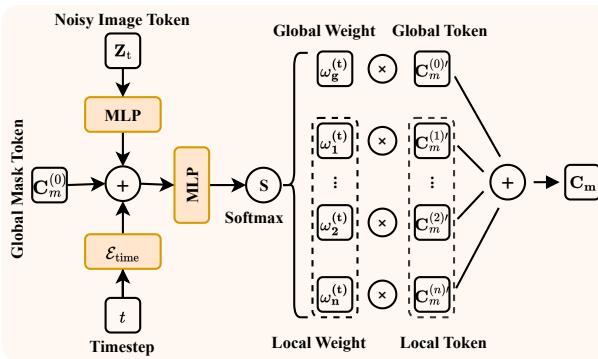


Figure 4: Structure of our dynamic gating network.

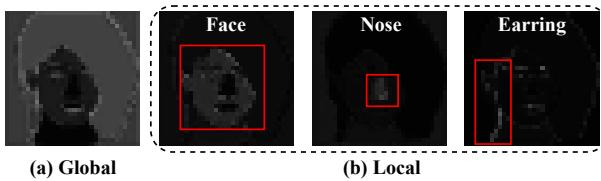


Figure 5: Visualization of global and partial local weight map.

accordance with [17], the initial 27,000 pairs from MM-Celeba-HQ are allocated for training, while the remaining 3,000 serve as the test set. To produce semantic masks for FFHQ-Text, we leverage two pretrained facial parsing models—FaRL [63] and SegFace [34]. For masks with overall accuracy (OA) below 0.8, we conduct manual annotation, whereas for those achieving $OA \geq 0.8$, we adopt a randomized sampling strategy, comprising 90% from FaRL and 10% from SegFace. More details of datasets are available in Appendix B.

4.2 Implementation Details

Our framework builds upon the open-source FLUX.1-dev [26]. During training, we set the batch size to 8 and utilize mixed-precision with the brain floating-point format to enhance computational efficiency and reduce memory usage. The optimizer employed is Prodigy [31], with a base learning rate of 1.0 and a weight decay of 0.01. A dropout probability 0.1 is applied to each controllable condition, excluding the global mask. The Low-Rank Adaptation (LoRA) [16, 51] is configured with a rank of 4 and a scaling factor of 4. A fixed random seed of 42 is used for all experiments to ensure reproducibility. Inference is performed with 28 sampling steps, and training is conducted for 4,000 steps, taking approximately 12 hours on a workstation equipped with 8×NVIDIA A100 80G GPUs.

4.3 Evaluation Metrics

Image Quality. The visual fidelity of generated images is a crucial indicator of model performance. We employ the following

Table 1: Comparison of face generation methods on the MM-CelebA-HQ dataset across different tasks.

Method	Venue	Paradigm	Image Generation Quality			Condition Alignment		IR \uparrow
			FID \downarrow	KID \downarrow	CMMMD \downarrow	Mask \downarrow	Text \uparrow	
Multimodal Face Generation								
TediGAN [56]	CVPR-21	GAN	83.35	74.75	1.562	2.46	23.90	-0.1446
UaC [33]	CVPR-23	Diffusion	75.35	63.78	1.982	3.41	25.52	-0.4001
Collaborative [17]	CVPR-23	Diffusion	24.48	13.50	0.734	3.22	24.51	-0.1265
PixelFace+ [8]	ACM MM-23	GAN	65.53	53.90	1.273	2.61	26.16	0.6403
DDGI [23]	CVPR-24	GAN & Diffusion	46.68	-	-	-	-	-
Face-MoGLE (Ours)	-	Diffusion	22.24	10.87	0.477	2.44	26.32	0.7014
Mask-to-Face Generation								
INADE [50]	CVPR-21	GAN	21.09	11.24	1.871	2.57	24.85	-0.0234
E2Style [55]	TIP-22	GAN	38.44	21.22	1.129	2.36	24.75	0.1649
SemFlow [54]	NeurIPS-24	Flow	56.65	41.48	1.767	2.30	25.65	-0.004
Face-MoGLE (Ours)	-	Diffusion	19.63	8.29	0.399	2.57	24.53	0.0398
Text-to-Face Generation								
clip2latent [39]	BMVC-22	GAN	63.89	38.55	1.410	5.05	27.56	1.0129
GCDP [36]	ICCV-23	Diffusion	72.67	43.81	1.456	4.68	25.94	0.4563
E3-FaceNet [60]	ICML-24	GAN	70.89	47.82	2.749	4.64	27.94	0.8150
Face-MoGLE (Ours)	-	Diffusion	34.81	21.85	0.636	4.94	26.91	0.9527

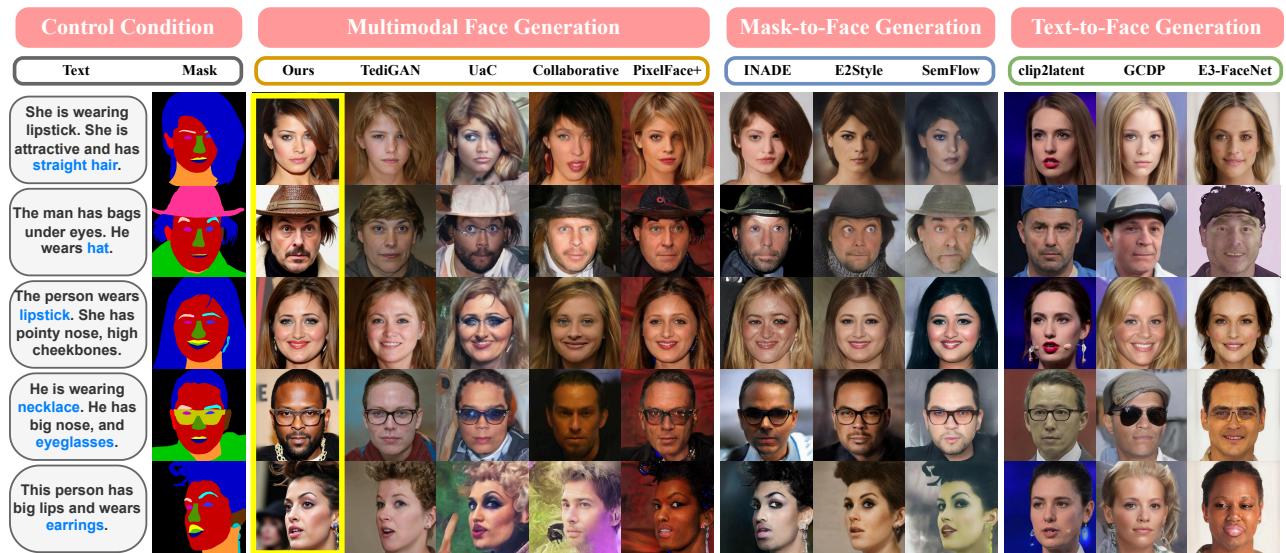
**Figure 6: Visualization results from different methods. The figure compares three generation paradigms: multimodal face synthesis (left) and two unimodal tasks – mask-to-face and text-to-face generation (middle and right). From top to bottom, each row highlights attribute-specific synthesis: straight hair, hat, lipstick, necklace, and earrings. Our method demonstrates superior alignment with the textual descriptions, semantic mask, and higher visual fidelity compared to other baseline models.**

Table 2: Comparison of multimodal face generation methods on the MM-FFHQ-Female dataset under zero-shot setting.

Method	Venue	Paradigm	Image Generation Quality			Condition Alignment		IR ↑
			FID ↓	KID ↓	CMMMD ↓	Mask ↓	Text ↑	
TediGAN [56]	CVPR-21	GAN	122.47	92.72	1.091	3.32	25.03	-0.4847
UaC [33]	CVPR-23	Diffusion	86.58	46.57	1.796	3.11	26.97	-0.8851
Collaborative [17]	CVPR-23	Diffusion	93.31	62.84	2.178	3.85	23.03	-1.2101
PixelFace+ [8]	ACM MM-23	GAN	<u>76.45</u>	<u>38.95</u>	1.917	<u>3.08</u>	26.60	<u>-0.2616</u>
Face-MoGLE (Ours)	-	Diffusion	62.93	31.27	<u>1.238</u>	2.77	28.06	0.1801

widely adopted metrics: *Fréchet Inception Distance* (*FID*) [12] quantifies the distributional discrepancy between generated and real images in the Inception feature space, with lower values indicating higher quality. *Kernel Inception Distance* (*KID*) [2], similar in spirit to *FID*, relies on kernel-based methods to provide an unbiased and more stable estimate. For readability, we multiply the *KID* by 1,000. *CLIP Maximum Mean Discrepancy* (*CMMMD*) [19] measures the alignment of conditional distributions, making it particularly suitable for conditional generation tasks.

Text Consistency. To evaluate how well the generated images align semantically with the input textual descriptions, we utilize the *CLIP Score* [41]. This metric leverages the CLIP vision-language model to compute similarity between text and image embeddings, with higher scores indicating more substantial semantic alignment. For better readability, we report the scores in percentage format.

Mask Consistency. We use *DINO Structure Distance* to measure how well the generated image aligns with source images under the guidance of the input semantic mask, which compares the self-similarity matrices of features from the DINO-ViT [5]. A smaller value indicates higher mask (structural) consistency.

Human Preference. To reflect subjective human perception, we incorporate the *Image Reward* (*IR*) score [57], derived from a learned aesthetic scoring model. This metric estimates the visual appeal and coherence of generated images and has been shown to correlate closely with human judgments.

Deepfake Detection. To further assess the realism of our synthesized facial images, we evaluate their ability to evade deepfake detection. It is measured using Area Under the Receiver Operating Characteristic Curve (*AUC*), Equal Error Rate (*EER*), and Average Precision (*AP*) [1, 49]. *AUC* and *EER* values close to 0.5 indicate increased similarity to authentic images and detector confusion, while lower *AP* reflects reduced confidence in fake identification.

4.4 Comparison with State-of-the-Art Methods

To evaluate the effectiveness of Face-MoGLE, we conduct comprehensive comparisons with representative state-of-the-art methods across three face generation tasks. As shown in Table 1, our model consistently performs well across key metrics of generated image quality, condition alignment, and human perceptual preference.

4.4.1 Multimodal Face Generation. Face-MoGLE achieves the best results in *FID* (22.24), *KID* (10.87), *CMMMD* (0.477), mask alignment

(2.44), and also ranks first in text alignment (26.32), indicating strong generation fidelity and semantic controllability. Since DDGI [23] has not released its code, we directly copy the results reported in its paper. Compared to the strongest diffusion-based baseline Collaborative [17], our method brings consistent improvements across all metrics, validating the effectiveness of the proposed framework.

4.4.2 Monomodal Face Generation.

Mask2Face Generation. For mask-to-face generation, Face-MoGLE again achieves the best results in *FID* (19.63), *KID* (8.29), and *CMMMD* (0.399). The mask alignment score (2.57) matches top-performing methods, demonstrating the model’s strength in preserving spatial structure while generating perceptually compelling faces.

Text2Face Generation. In the text-to-face setting, Face-MoGLE achieves superior performance in terms of generation fidelity and condition consistency. Specifically, it obtains the best results across all image quality metrics, with an *FID* of 34.81, *KID* of 21.85, and *CMMMD* of 0.636. Compared to other methods, Face-MoGLE is more effective at generating realistic and condition-consistent facial images from text descriptions, highlighting the effectiveness of our diffusion-based approach.

4.4.3 Visualization and Human Preference. We use the *IR* score [57] to assess the perceptual quality of the generated faces. As shown in Table 1, Face-MoGLE achieves the highest *IR* scores in multimodal (0.7014) and mask-to-face (0.0398) generation, and ranks second in text-to-face (0.9527), indicating that the generated images are both semantically aligned and preferred by human observers. Fig. 6 highlights these results, where Face-MoGLE effectively resolves cross-modal conflicts and generates more natural features in multimodal and mask-to-face tasks. Although slightly behind in text-to-face *IR*, it achieves better geometric accuracy and visual-textual consistency, striking a balance between structural fidelity and semantic alignment. More results are available in Appendix C.

4.5 Ablation Studies

4.5.1 Effect of Harmonizing Global and Local Experts. We conduct ablation studies on three configurations: *Global Expert* (holistic semantic mask), *Local Experts* (decoupled binary masks), and the *Combined Global + Local Experts*, as summarized in Table 3. The *Global Expert* alone yields a moderate *FID* of 30.36 but shows poor semantic mask alignment (Mask: 2.47). In contrast, *Local Experts*, while achieving the best text alignment (Text: 27.07) due to precise

Table 3: Effect of harmonizing global and local experts.

Expert Composition	FID ↓	KID ↓	Mask ↓	Text ↑
Only Global	30.36	18.16	2.47	26.30
Only Local	33.62	20.45	4.87	27.07
Global & Local	22.24	10.87	2.44	26.32

mapping between facial regions and textual semantics, suffer from the highest FID (33.62) due to the lack of holistic spatial context. Our unified framework, which dynamically integrates global and local features via a gating network, outperforms both, achieving a significantly improved FID of 22.24 and enhanced mask consistency (2.44). These results highlight the effectiveness of hierarchical expert fusion: global experts ensure topological coherence, while local experts provide fine-grained semantic control, jointly promoting visual fidelity and structural precision.

Table 4: Impact of various gating mechanisms.

Gating Mechanism	FID ↓	KID ↓	Mask ↓	Text ↑
w/o Diffusion	25.74	13.12	2.37	26.53
Scalar Gating	43.48	30.58	4.74	26.72
Matrix Gating	22.24	10.87	2.44	26.32

4.5.2 *Impact of Various Gating Mechanisms.* We evaluate three gating strategies: static weights, time-dependent scalar weights, and our spatiotemporal matrix weights. As Table 4 shows, matrix gating achieves the best FID (22.24) and KID (10.87), outperforming scalar gating by 48.9% and static weights by 13.6%. The severe degradation under scalar gating (FID:43.48) stems from its inability to handle spatial conflicts, whereas static weights (FID:25.74) lack temporal adaptability across diffusion stages. Our method resolves these through pixel-wise weight maps that evolve spatially and temporally, achieving optimal balance between semantic control (Mask:2.44) and photorealism. This conclusively demonstrates the necessity of spatiotemporal dynamics in expert fusion.

Table 5: Joint contribution of expert and gating.

Expert	Gating	FID ↓	KID ↓	Mask ↓	Text ↑
✗	✗	33.25	23.07	2.49	26.71
✓	✗	<u>26.55</u>	<u>14.87</u>	3.20	26.38
✗	✓	31.30	19.11	2.60	<u>26.64</u>
✓	✓	22.24	10.87	2.44	26.32

4.5.3 *Joint Contribution of Expert and Gating.* We ablate the individual and joint effects of global-local experts and dynamic gating. As shown in Table 5, using only experts (FID:26.55) or gating (FID:31.30) yields partial improvements over the baseline (FID:33.25), while their combined use achieves optimal FID (22.24). This synergy arises because experts decompose facial semantics into multiple binary components (mask error drops from 3.20 to 2.44), while

the gating network dynamically aligns these components across space and time. Although the baseline shows marginally higher text alignment (Text:26.71 vs. 26.32), its overly smoothed outputs lack semantic precision, whereas our full model balances photorealism and control. This validates that global-local experts and adaptive gating are mutually essential for high-fidelity and controllable generation.

4.5.4 *Zero-Shot Generalization Validation.* As shown in Table 2, Face-MoGLE achieves superior zero-shot generalization on the MM-FFHQ-Female dataset, outperforming existing methods on most metrics. Our framework attains state-of-the-art image quality (FID: 62.93, KID: 31.27), mask consistency (Mask: 2.77), text alignment (CLIP: 28.06), and human preference (Image Reward: 0.1801). While TediGAN shows better CMMD performance (1.091 vs. ours 1.238), this can be attributed to its StyleGAN backbone being pre-trained on the FFHQ dataset. Compared to the best diffusion-based baseline (UaC), our method reduces FID by 27.3% and KID by 32.9%, demonstrating the effectiveness of our mask-decoupling strategy and global-local MoE architecture. The dynamic gating network enables adaptive feature fusion during the diffusion process, contributing to robust performance on unseen semantic combinations. These results validate Face-MoGLE’s capability to synthesize high-fidelity faces under zero-shot conditions without task-specific fine-tuning.

4.5.5 *Evaluation with Deepfake Detection.* We evaluate the ability of our synthesized faces to evade deepfake detectors, as shown in Table 6. Specifically, we test against *NPR* [49], a general-purpose detector, and *Wavelet-CLIP* [1], which focuses on face-specific artifacts. Face-MoGLE achieves near-random AUC on *NPR* (0.50 vs. Collaborative’s 0.51) and significantly outperforms Collaborative on *Wavelet-CLIP* (0.46 vs. 0.75). In light of these results, we stress that this evaluation is conducted solely for defensive research purposes. Societal impacts and responsible AI are discussed in Appendix A.

Table 6: Performance comparison of two deepfake detection models against different face generation methods. Each cell is shown as *NPR* [49] / *Wavelet-CLIP* [1] detection results.

Method	AUC	EER	AP
TediGan [56]	0.73 / 0.81	0.35 / 0.26	0.71 / <u>0.76</u>
UaC [33]	0.45 / 0.96	0.53 / 0.11	0.43 / 0.96
Collaborative [17]	<u>0.51</u> / 0.75	<u>0.51</u> / 0.32	<u>0.50</u> / 0.79
PixelFace+ [8]	0.28 / 0.87	0.65 / 0.20	0.37 / 0.89
Face-MoGLE (Ours)	0.50 / 0.46	0.50 / 0.53	0.52 / 0.46

5 Conclusion

In this work, we introduced a mixture of global and local experts with a diffusion transformer for controllable face generation. Our method effectively decouples semantic mask information and dynamically selects experts to enhance image synthesis fidelity and condition alignment. Through extensive experiments, we demonstrated improvements in multimodal and monomodal generation and robust generalization. Future work may explore more efficient architectures and applications in real-world scenarios.

References

- [1] Lalith Bharadwaj Baru, Rohit Boddeda, Shilhora Akshay Patel, and Sai Mohan Gajapaka. 2025. Wavelet-Driven Generalizable Framework for Deepfake Face Forgery Detection. In *WACV*. 1661–1669.
- [2] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *ICLR*. 1–36.
- [3] Bocheng YuhangMa, wuliebucha, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. 2024. HiCo: Hierarchical Controllable Diffusion Model for Layout-to-image Generation. In *NeurIPS*.
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*. 18392–18402.
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*. 9650–9660.
- [6] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*, Vol. 34. 8780–8794.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. 1–21.
- [8] Xiaoxiong Du, Jun Peng, Yiyi Zhou, Jinlu Zhang, Siting Chen, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. 2023. PixelFace+: Towards Controllable Face Generation and Manipulation with Text Descriptions and Segmentation Masks. In *ACM MM*. 4666–4677.
- [9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*. 28 pages.
- [10] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *JMLR* 23, 120 (2022), 1–39.
- [11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, Vol. 33. 9841–9850.
- [12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*. 6629–6640.
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, Vol. 33. 6840–6851.
- [14] Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS*. 1–8.
- [15] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *TOG* 41, 4, Article 161 (2022), 19 pages.
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*. 1–13.
- [17] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. 2023. Collaborative diffusion for multi-modal face generation and editing. In *CVPR*. 6080–6090.
- [18] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.
- [19] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. 2024. Rethinking FID: Towards a Better Evaluation Metric for Image Generation . In *CVPR*. 9307–9315.
- [20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*.
- [21] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.
- [22] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*. 4401–4410.
- [23] Jihyun Kim, Changjae Oh, Hoseok Do, Soohyun Kim, and Kwanghoon Sohn. 2024. Diffusion-driven gan inversion for multi-modal face image generation. In *CVPR*. 10403–10412.
- [24] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, Vol. 31. 1–10.
- [25] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational {Bayes}. In *ICLR*. 1–14.
- [26] Black Forest Labs. 2024. FLUX. <https://github.com/black-forest-labs/flux>.
- [27] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*. 5549–5558.
- [28] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. {GS}hard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *ICLR*. 1–23.
- [29] Fang Liu, Xiaoming Deng, Yu-Kun Lai, Yong-Jin Liu, Cuixia Ma, and Hongan Wang. 2019. SketchGAN: Joint Sketch Completion and Recognition With Generative Adversarial Network. In *CVPR*.
- [30] Debin Meng, Christos Tzelepis, Ioannis Patras, and Georgios Tzimiropoulos. 2024. MM2Latent: Text-to-facial image generation and editing in GANs with multimodal assistance. In *ECCV*. 1–20.
- [31] Konstantin Mishchenko and Aaron Defazio. 2024. Prodigy: An Expeditiously Adaptive Parameter-Free Learner. In *ICML*. 35779–35804.
- [32] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, Vol. 38. 4296–4304.
- [33] Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. 2023. Unite and conquer: Plug & play multi-modal synthesis using diffusion models. In *CVPR*. 6070–6079.
- [34] Kartik Narayan, Vibashan VS, and Vishal M Patel. 2024. Segface: Face segmentation of long-tail classes. *arXiv preprint arXiv:2412.08647* (2024).
- [35] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *ICML*. 8162–8171.
- [36] Minho Park, Jooyeon Yun, Seunghwan Choi, and Jaegul Choo. 2023. Learning to generate semantic layouts for higher text-image correspondence in text-to-image synthesis. In *ICCV*. 7591–7600.
- [37] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*. 2085–2094.
- [38] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *ICCV*. 4195–4205.
- [39] Justin N. M. Pinkney and Chuan Li. 2022. clip2latent: Text driven sampling of a pre-trained StyleGAN using denoising diffusion and CLIP. In *BMVC*. 1–12.
- [40] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *ACM MM*. 484–492.
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, Vol. 139. 8748–8763.
- [42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR* 21, 140 (2020), 1–67.
- [43] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. In *NeurIPS*, Vol. 34. 8583–8595.
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*. 10684–10695.
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. 234–241.
- [46] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*. 1–19.
- [47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*. 2256–2265.
- [48] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *ICLR*. 1–20.
- [49] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*. 28130–28139.
- [50] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. 2021. Diverse semantic image synthesis via probability distribution modeling. In *CVPR*. 7962–7971.
- [51] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098* (3) (2024).
- [52] Hao Tang, Dan Xu, Yan Yan, Philip H.S. Torr, and Nicu Sebe. 2020. Local Class-Specific and Global Image-Level Generative Adversarial Networks for Semantic-Guided Scene Generation. In *CVPR*. 7870–7879.
- [53] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *NeurIPS*. 6309–6318.
- [54] Chaoyang Wang, Xiangtai Li, Lu Qi, Henghui Ding, Yunhai Tong, and Ming-Hsuan Yang. 2024. Semflow: Binding semantic segmentation and image synthesis via rectified flow. In *NeurIPS*, Vol. 37. 138981–139001.
- [55] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. 2022. E2Style: Improve the efficiency and effectiveness of StyleGAN inversion. *TIP* 31 (2022), 3267–3280.
- [56] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*. 2256–2265.
- [57] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *NeurIPS*, Vol. 36. 15903–15935.

- 1045 [58] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang,
1046 Yizhong Zhang, Xin Tong, and Baining Guo. 2024. Vasa-1: Lifelike audio-driven
1047 talking faces generated in real time. In *NeurIPS*, Vol. 37. 660–684.
- 1048 [59] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text
1049 compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).
- 1050 [60] Jinlu Zhang, Yiyi Zhou, Qiancheng Zheng, Xiaoxiong Du, Gen Luo, Jun Peng,
1051 Xiaoshuai Sun, and Rongrong Ji. 2024. Fast text-to-3D-aware face generation
1052 and manipulation via direct cross-modal mapping and geometric regularization.
In *ICML*. 60605–60625.
- 1053
- 1054
- 1055
- 1056
- 1057
- 1058
- 1059
- 1060
- 1061
- 1062
- 1063
- 1064
- 1065
- 1066
- 1067
- 1068
- 1069
- 1070
- 1071
- 1072
- 1073
- 1074
- 1075
- 1076
- 1077
- 1078
- 1079
- 1080
- 1081
- 1082
- 1083
- 1084
- 1085
- 1086
- 1087
- 1088
- 1089
- 1090
- 1091
- 1092
- 1093
- 1094
- 1095
- 1096
- 1097
- 1098
- 1099
- 1100
- 1101
- 1102
- [61] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte.
2022. Plug-and-Play Image Restoration With Deep Denoiser Prior. *TPAMI* 44, 10
(2022), 6360–6376.
- [62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional
control to text-to-image diffusion models. In *ICCV*. 3836–3847.
- [63] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu
Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. 2022. General facial
representation learning in a visual-linguistic manner. In *CVPR*. 18697–18709.
- [64] Yutong Zhou. 2021. Generative adversarial network for text-to-face synthesis
and manipulation. In *ACM MM*. 2940–2944.

1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160