

EXPLICIT CONDITIONAL CONSISTENCY DIFFUSION: TOWARDS PRECISE SEMANTIC ALIGNMENT IN MUL- TIMODAL FACE GENERATION

Anonymous authors

Paper under double-blind review

ABSTRACT

With the collaborative guidance of multimodal conditions (e.g., semantic masks as structural visual guidance and text descriptions as linguistic guidance), diffusion models have significantly improved the controllability of face generation. However, existing methods rely solely on noise learning or flow matching to implicitly model the relationship between latent representations and multimodal features, making it difficult to fully capture their semantic associations and resulting in sub-optimal conditional consistency in the generated outputs. To overcome this limitation, we propose EC²Face, a novel facial generation method based on an explicit conditional consistency diffusion framework. Our approach introduces a temporally modulated conditional consistency guidance mechanism in the pixel space, explicitly driving precise semantic alignment between the latent representations and multimodal conditions. In addition, to address the poor response to long-tailed attributes in mask conditions, we design a long-tail adaptive learning strategy. It dynamically assigns differentiated weights to spatial locations through gradient reweighting, enhancing the model’s ability to perceive rare attributes and effectively mitigating model bias. Extensive experiments demonstrate that EC²Face significantly outperforms other competing methods across most evaluation metrics, particularly exhibiting an improvement of over 9.0% in mask accuracy for regions with rare attributes.

1 INTRODUCTION

Facial generation Ning et al. (2023); Melnik et al. (2024); Wang et al. (2025a), as a core task in the fields of artificial intelligence and computer vision, has long been a focal point of research attention. In recent years, multimodal facial generation has gradually emerged as a research hotspot due to its exceptional controllability advantages. Unlike traditional unimodal generation methods, multimodal facial generation Meng et al. (2025); Sowmya & Meeradevi (2024); Kim et al. (2024b) integrates multi-source modality data, such as textual descriptions (e.g., “blonde hair, blue eyes, smiling expression”) and semantic masks (e.g., segmentation masks for facial organ regions), thereby enabling fine-grained control over the generation process.

In the early stages, research on multimodal facial generation was primarily based on Generative Adversarial Networks (GANs) Du et al. (2023a); Meng et al. (2025). However, constrained by factors such as limited expressive capacity in the latent space and unstable training, these methods commonly suffered from issues such as insufficient generation diversity and mode collapse. Diffusion models Po et al. (2024); He et al. (2025); Chang et al. (2025) have achieved a revolutionary breakthrough in the field of image generation, thanks to their unique “forward noising - reverse denoising” progressive learning mechanism. Their flexible conditional injection mechanism Zhang et al. (2023b); Mou et al. (2024a); Peng et al. (2024) has opened up a new technical pathway for the controllability of multimodal facial generation, making them the prevailing technical framework in current research.

Nevertheless, diffusion models primarily rely on noise learning or flow matching Rombach et al. (2022b); Dao et al. (2023) for implicit modeling, making it challenging to fully capture the semantic correlations between latent representations and multimodal conditions. Additionally, the original

optimization mechanism cannot effectively tackle the long-tailed attribute distribution issue in mask conditions, leading to inadequate responses for some rare attributes (e.g., earrings, necklaces).

To address the aforementioned challenges, this paper introduces EC²Face—a novel approach grounded in an explicit conditional consistency diffusion framework. This method significantly enhances the semantic alignment accuracy of synthesized faces through two key improvements. First, we introduce a temporally modulated conditional consistency guidance mechanism in the pixel space. This mechanism explicitly drives the latent representations to maintain precise semantic alignment with multimodal conditions. The specially designed temporal modulation function can circumvent the adverse effects caused by the unreliability of reverse estimation at large time steps. Second, to tackle the modeling bias issue induced by the long-tailed attribute distribution in mask conditions, we propose a long-tailed adaptive learning strategy. This strategy dynamically applies differential weights to spatial positions during the flow matching process Dao et al. (2023) based on pre-computed attribute frequencies, thereby enhancing the sensitivity of the diffusion model to rare attributes and further strengthening the multimodal conditional consistency of synthesized faces.

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

1. We propose EC²Face, a novel multimodal facial generation framework with explicit conditional consistency guidance, which addresses the limitations inherent in diffusion models that rely solely on latent space modeling.
2. We devise a multimodal conditional consistency constraint in the pixel space based on latents reverse estimation. This constraint explicitly steers latent representations towards achieving precise semantic alignment with multimodal conditions. Furthermore, we introduce a temporal dynamic modulation function that can effectively circumvent the constraint bias caused by the unreliability of reverse estimation at large time steps.
3. We propose a long-tail adaptive learning strategy, which enhances the model’s sensitivity to rare attributes by assigning differential weights to different spatial locations during the flow matching computation process, thereby effectively mitigating the modeling bias of diffusion models towards long-tail semantic attributes.

Comprehensive qualitative and quantitative evaluations demonstrate that EC²Face significantly outperforms existing baseline methods across multiple evaluation metrics, particularly in multimodal conditional consistency.

2 RELATED WORKS

2.1 FACE GENERATION

Early research on face generation primarily focused on unimodal scenarios, employing only text Pinkney & Li (2022); Park et al. (2023); Zhang et al. (2024) or semantic masks Tan et al. (2021); Wei et al. (2022); Wang et al. (2024) as single input conditions. Among them, representative methods based on Generative Adversarial Networks (GANs), such as the StyleGAN series Karras et al. (2019; 2020) and its text-driven extension StyleCLIP Patashnik et al. (2021), have demonstrated the immense potential of latent space modeling, achieving not only highly realistic generation results but also enabling semantic image editing through latent space manipulations. Despite the significant progress these methods have made in generating realistic and diverse facial images, they are inherently limited by their unimodal nature, exhibiting significant limitations: text-driven approaches often suffer from deviations in spatial layout control, whereas mask-driven methods find it challenging to flexibly manipulate appearance attributes like makeup and accessories. Consequently, unimodal methods face challenges in achieving simultaneous and precise control over both spatial structure and attribute configuration.

To overcome these limitations, multimodal face generation has gradually emerged as a research hotspot, enhancing the controllability of generated faces by integrating the complementary strengths of heterogeneous modalities. For instance, methods such as TediGAN Xia et al. (2021), PixelFace+ Du et al. (2023b), CoDiffusion Huang et al. (2023), and MM2Latent Meng et al. (2025) combine textual descriptions with visual conditions (e.g., masks, sketches), enabling precise multi-dimensional control over identity, pose, and attributes. Similarly, hybrid strategies that fuse diffusion models

with GANs Kim et al. (2024a); Nair et al. (2023) further enhance cross-modal semantic alignment capabilities. However, current methods still face challenges in achieving precise semantic alignment between synthesized faces and multimodal conditions (e.g., masks and text), particularly for modeling long-tailed or small-area attributes such as earrings.

2.2 DIFFUSION MODELS

In the field of generative modeling, recent research has demonstrated a paradigm shift from Generative Adversarial Networks (GANs) Salimans et al. (2016); Arjovsky et al. (2017) to diffusion models Ho et al. (2020); Song et al. (2021); Rombach et al. (2022a). This transition is primarily attributed to the significant advantages of diffusion models in terms of training stability and output diversity. Notably, DDPM Ho et al. (2020) and DDIM Song et al. (2021) have laid the technical foundation for diffusion-based generation, while LDM Rombach et al. (2022a) has substantially enhanced image generation efficiency through latent space diffusion strategies. In the realm of controllable generation, the introduction of techniques such as CFG Ho & Salimans (2022), ControlNet Zhang et al. (2023a), T2I-Adapter Mou et al. (2024b), and IP-Adapter Ye et al. (2023) has established diffusion models as the mainstream approach for conditional image generation. Moreover, incorporating visual transformer architectures has allowed us to overcome the perceptual constraints inherent in traditional U-Net models, significantly enhancing the quality of generated images Peebles & Xie (2023); Wang et al. (2025b); Tan et al. (2024); Li et al. (2023). However, existing methods still predominantly rely on noise learning Ho et al. (2020) or flow matching Dao et al. (2023) to implicitly explore the associations between latent variables and multimodal features during multimodal collaborative image generation, resulting in insufficient fine-grained consistency between generated images and multimodal conditions.

To overcome these limitations, we propose a novel multimodal face generation method termed EC²Face. By incorporating a conditional consistency loss function tailored to multimodal conditions in pixel space, our approach effectively compensates for the suboptimal generation quality stemming from implicit latent space modeling. Additionally, we refine the traditional flow matching Dao et al. (2023) learning strategy through differential spatial weight allocation, significantly enhancing the diffusion model’s accuracy in modeling long-tailed attributes.

3 METHOD

3.1 PRELIMINARIES

Diffusion Transformer. The Diffusion Transformer (DiT) Peebles & Xie (2023) stands as one of the cutting-edge architectures in the current generation field. It replaces the traditional U-Net backbone with stacked transformer blocks, successfully breaking through the limitations of local perception. The core advantage of this architecture lies in the fact that multimodal feature tokens can engage in deep interactions through the multi-head attention mechanism, with its mathematical formalization expressed as:

$$\text{attn}(q, k, v) = \text{softmax} \left(\frac{qk^\top}{\sqrt{d_h}} \right) v. \quad (1)$$

Here, the query vector q , the key vector k , and the value vector v are all generated through linear projections of the concatenated multimodal tokens, while d_h denotes the dimensionality of a single attention head.

Flow matching. Flow Matching Dao et al. (2023), as an optimization method for generative models based on continuous normalized flows, exhibits significant advantages in generation efficiency and stability compared to the noise prediction paradigm relied upon by traditional diffusion models. Its core lies in constructing a smooth flow field $\{z_t\}_{t \in [0,1]}$ from a noise distribution $\epsilon \sim \mathcal{N}(0, 1)$ to a real data distribution $z \sim p_{\text{data}}(z)$, where t serves as a continuous time variable.

The model enhances the generative process by learning the velocity function $v_\theta(z_t, t)$ of the flow field, ensuring that the flow field at any moment t satisfies the following differential equation:

$$\frac{dz_t}{dt} = v_\theta(z_t, t). \quad (2)$$

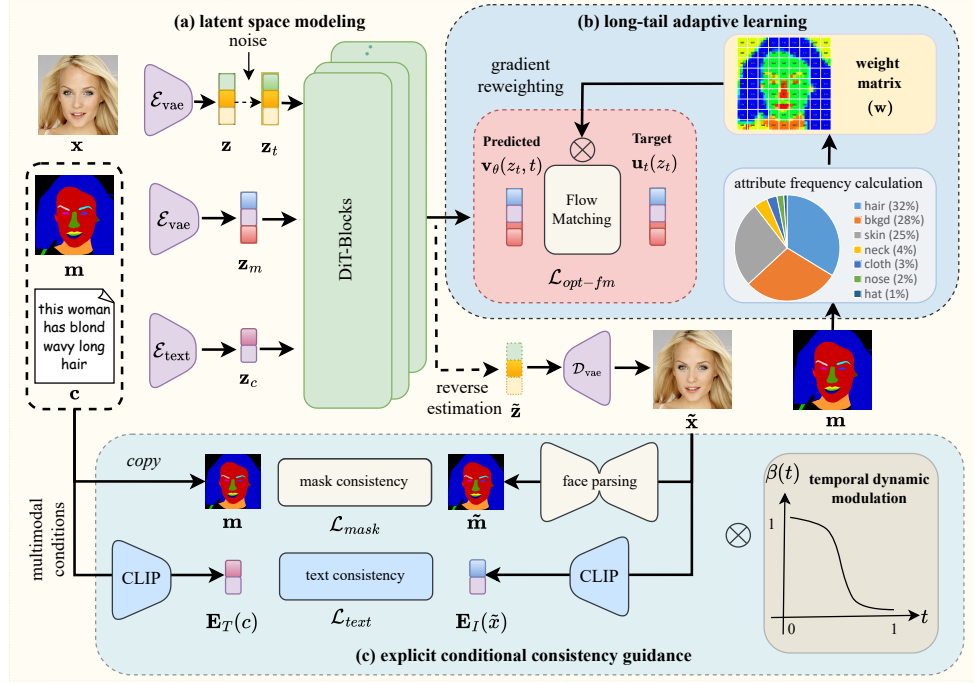


Figure 1: The overall training framework of EC²Face.

The optimization objective of flow matching is to minimize the discrepancy between the predicted velocity field $v_\theta(z_t, t)$ and the target velocity field $\mu_t(z_t)$. The loss function is defined as:

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{t \sim \text{Uniform}[0,1], \epsilon \sim \mathcal{N}(0,1), z \sim p_{\text{data}}} [\|v_\theta(z_t, t) - \mu_t(z_t)\|_2^2]. \quad (3)$$

Here, $z_t = (1 - t) \cdot z + t \cdot \epsilon$ represents the intermediate state at time t along the flow path between the noise ϵ and the real sample z , and $\mu_t(z_t) = \epsilon - z$ denotes the target velocity field.

3.2 OVERALL FRAMEWORK

Figure 1 illustrates the overall training framework of EC²Face. This framework, an extension of FLUX.1-dev Black-Forest-Labs (2024), is designed to precisely synthesize high-fidelity facial images through the collaborative driving of mask and text. During the training process, we project both text and mask information into a latent space shared with noisy images, thereby eliminating discrepancies between representations of heterogeneous modalities. To maintain the framework’s simplicity, we reuse the Variational Autoencoder (VAE) to process mask images. Specifically, the noisy image token z_t , mask token z_m , and text token z_c form a joint token sequence, which is fed into stacked transformer modules for attention-based interaction. The query vector q , key vector k , and value vector v can be computed according to Equation 4:

$$q = [w_q^c z_c; w_q z_t; w_q^m z_m], \quad k = [w_k^c z_c; w_k z_t; w_k^m z_m], \quad v = [w_v^c z_c; w_v z_t; w_v^m z_m]. \quad (4)$$

In our training framework, we have implemented two core improvement measures to strengthen the semantic consistency between synthetic facial images and multimodal conditions. Firstly, we constructed an explicit conditional consistency guidance mechanism in the pixel space. This mechanism can directly facilitate precise semantic alignment between latent representations and multimodal conditions. Secondly, to address the modeling bias issue of diffusion models when handling long-tailed attributes, we innovatively proposed a long-tail adaptive learning strategy. This strategy effectively enhances the model’s ability to model rare attributes, thereby significantly improving the controllable accuracy of synthetic facial images.

3.3 EXPLICIT CONDITIONAL CONSISTENCY GUIDANCE

Existing DiTs models predominantly employ flow-matching learning strategies within the latent space. However, such implicit modeling approaches struggle to enable latent representations to accurately capture the semantic information of multimodal features (e.g., text c and mask m). To address this challenge, this study proposes introducing explicit constraint functions compatible with multimodal conditions in the pixel space to directly drive the modeling process of diffusion transformers. Specifically, during the training phase, for each noisy latent z_t and the model-predicted velocity field $v_\theta(z_t, t)$, the initial latent variable $\tilde{z}$ is derived through reverse estimation (as shown in Equation 5), and then an estimated facial image $\tilde{x}$ is generated via the VAE decoder. Based on $\tilde{x}$, the consistency loss between the facial image and multimodal conditions can be quantitatively computed in the pixel space.

$$\tilde{z} = z_t - t \cdot v_\theta(z_t, t). \quad (5)$$

Text Consistency Loss. We employ the CLIP model Radford et al. (2021) to measure the semantic consistency between the facial image $\tilde{x}$ and the textual condition c . This model is equipped with a pre-trained text encoder $\mathbf{E}_T$ and a pre-trained image encoder $\mathbf{E}_I$, which can efficiently extract features from both text and images. Based on the extracted features, we utilize the cosine loss function to quantify the discrepancy between the generated facial image $\tilde{x}$ and the given textual condition c . The specific formula is as follows:

$$\mathcal{L}_{\text{text}} = 1 - \frac{\mathbf{E}_I(\tilde{x}) \cdot \mathbf{E}_T(c)}{\|\mathbf{E}_I(\tilde{x})\| \cdot \|\mathbf{E}_T(c)\|}. \quad (6)$$

Mask Consistency Loss. We employ the pre-trained facial parsing model FaRL Zheng et al. (2022) to process the facial image $\tilde{x}$, thereby obtaining its predicted semantic segmentation map $\tilde{m}$. Based on this predicted outputs and the given mask condition m , we calculate the mask consistency loss by taking a weighted combination of the multi-label cross-entropy loss and the Dice coefficient loss Milletari et al. (2016). The specific definition can be found in Equation 7.

$$\begin{aligned} \mathcal{L}_{\text{mask}} &= \lambda \cdot \mathcal{L}_{\text{CE}}(\tilde{m}, m) + (1 - \lambda) \cdot \mathcal{L}_{\text{Dice}}(\tilde{m}, m) \\ &= \lambda \cdot \left[- \sum_i (m_i \log \tilde{m}_i + (1 - m_i) \log(1 - \tilde{m}_i)) \right] \\ &\quad + (1 - \lambda) \cdot \left(1 - \frac{2 \sum_i m_i \tilde{m}_i}{\sum_i m_i^2 + \sum_i \tilde{m}_i^2 + \delta} \right), \end{aligned} \quad (7)$$

where $\lambda \in [0, 1]$ is the balancing coefficient (set to 0.5 empirically in our experiments), and $\delta > 0$ denotes a smoothing parameter introduced to ensure numerical stability.

Temporal Dynamic Modulation. The acquisition of facial image $\tilde{x}$ hinges on the latent reverse estimation outlined in Equation 5 (from z_t to $\tilde{z}$). Experiments reveal that the reliability of this reverse estimation is not uniform across time steps; as shown in Figure 2, it declines as the time step t increases. To avoid the adverse effects caused by the unreliability of reverse estimation at large time steps, we designed a temporal dynamic modulation function, the specific form of which is presented in Equation 8. This function can dynamically adjust the intensity of the conditional consistency constraint according to the variation of the timestep t .

$$\beta(t) = \frac{1}{1 + \exp(\gamma_s \cdot (t - t_c))}. \quad (8)$$

The rationale is straightforward: When t is small, the reverse estimation is highly reliable, so we should strengthen its influence. Conversely, when t is large, the reverse estimation becomes less reliable, necessitating a reduction in the constraint intensity. The coefficients γ_s and t_c control the steepness of this function and the position of its center point, respectively. We empirically set $\gamma_s = 10$ and $t_c = 0.4$ in our experiments.



Figure 2: Reliability of reverse estimation.

Multimodal Conditional Consistency Loss. Upon applying temporal dynamic modulation, we can calculate an overall multimodal conditional consistency loss in the pixel space, with its specific formulation depicted in Equation 9. Here, α_1 and α_2 serve as balancing coefficients to adjust the contribution weights of different loss terms.

$$\mathcal{L}_{\text{mcc}} = \beta(t) \cdot (\alpha_1 \cdot \mathcal{L}_{\text{text}} + \alpha_2 \cdot \mathcal{L}_{\text{mask}}). \quad (9)$$

3.4 LONG-TAIL ADAPTIVE LEARNING

Long-Tail Distribution of Semantic Attributes. We observed a significant imbalance in the distribution of semantic attributes within mask images (refer to Figure 5 in Appendix A for details). Statistical analysis of the widely - used MM-CelebA-HQ dataset Lee et al. (2020) revealed that semantic attributes exhibit a typical long-tailed distribution in terms of both average pixel proportion and frequency of occurrence in training samples. However, existing methods fail to effectively address this issue, resulting in rare attributes such as earrings and necklaces being overlooked during the generation process. To tackle this challenge, we propose a long-ailed adaptive learning strategy that enhances the diffusion model’s perception of rare attributes through a gradient reweighting mechanism.

Gradient Reweighting. In the flow-matching computation within the latent space, we employ a method of assigning differential weights to spatial positions to enhance the model’s perception of rare attributes. Specifically, during the data loading phase, we pre-compute the distribution of semantic attributes within the mask m . Let n_c denote the total number of attribute categories, and $\{T_c\}_{c=1}^{n_c}$ represent the aggregated pixel count for each attribute category. Then, the elements of the differential weight matrix $w \in \mathbb{R}^{h \times w}$ are calculated according to the following formula:

$$w(i, j) = \frac{\sum_{c=1}^n T_c}{n_c \cdot T_{m(i, j)} + 1} \quad (10)$$

where $m(i, j)$ indicates the attribute category label at the spatial position (i, j) . The underlying motivation of this approach is to increase the weights of regions with rare attributes and decrease the weights of regions with high-frequency attributes, thereby achieving a relative balance among different attributes. The calculation of the flow-matching loss after gradient reweighting is shown in Equation 11:

$$\mathcal{L}_{\text{opt-fm}} = \mathbb{E}_{t \sim \text{Uniform}[0, 1], \epsilon \sim \mathcal{N}(0, 1), z \sim p_{\text{data}}} [\|w \cdot (v_{\theta}(z_t, m, c, t) - \mu_t(z_t))\|_2^2]. \quad (11)$$

3.5 TRAINING STRATEGY

LoRA Fine-tuning. During the training phase, to reduce computational overhead and parameter redundancy in model training, we employ a lightweight adaptation architecture. Specifically, we conduct low-rank adaptation fine-tuning (LoRA) Hu et al. (2021) solely on the visual embedding layer of the base model and some linear layers within the transformer blocks. Compared with full-model fine-tuning, this approach introduces an increase of less than 0.1% in the number of new parameters. The final loss function of this model comprises two components: firstly, the flow-matching loss with gradient reweighting; secondly, the multimodal conditional consistency loss introduced in the pixel space:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{opt-fm}} + \mathcal{L}_{\text{mcc}} \\ &= \mathcal{L}_{\text{opt-fm}} + \beta(t) \cdot (\alpha_1 \cdot \mathcal{L}_{\text{text}} + \alpha_2 \cdot \mathcal{L}_{\text{mask}}). \end{aligned} \quad (12)$$

Stochastic Condition Dropout. To endow the model with the ability to handle both multimodal collaborative-driven and unimodal-driven facial generation simultaneously, we specifically introduce a stochastic condition dropout strategy during the training process. More precisely, at the data loading stage, for each individual sample, this strategy independently and randomly drops the text condition c and the mask condition m with a pre-defined probability p ($p = 0.1$ in our experiments). The detailed procedure is demonstrated in Equation 13.

$$c/m = \begin{cases} \phi & \text{with probability } p, \\ c/m & \text{with probability } 1 - p. \end{cases} \quad (13)$$

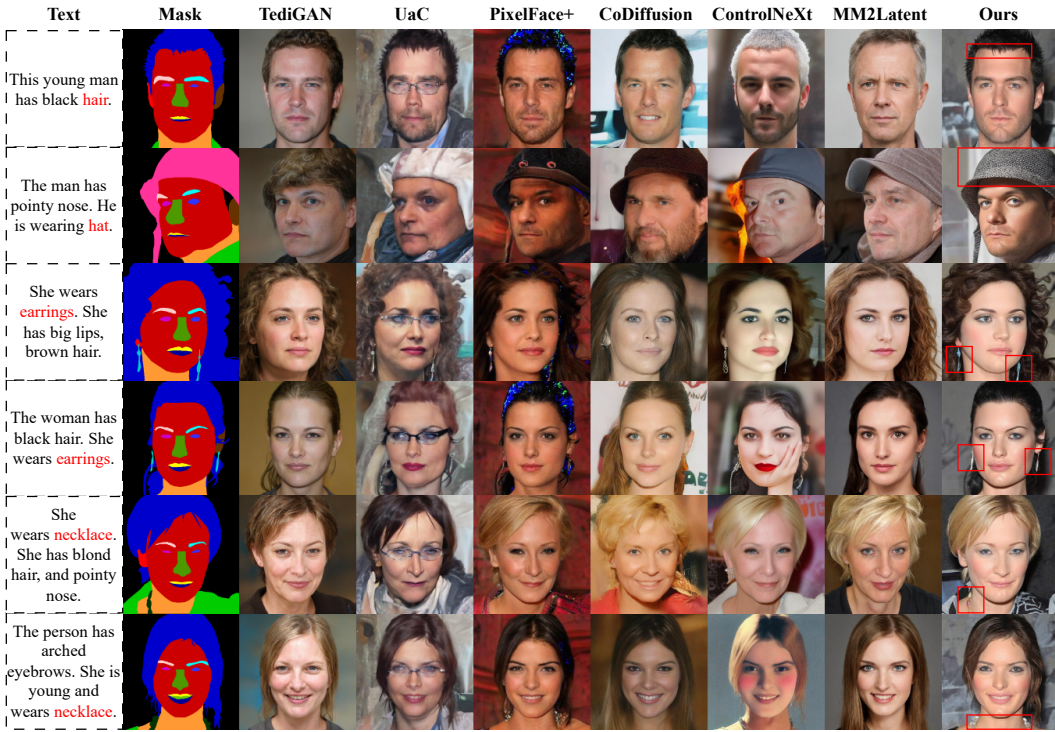


Figure 3: Visual comparison of multimodal face generation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. To comprehensively evaluate the performance of the proposed EC²Face model, we selected the MM-CelebA-HQ Lee et al. (2020), a large-scale benchmark dataset widely recognized in the field of multimodal facial synthesis. This dataset comprises 30,000 high-resolution, diverse RGB facial images, each accompanied by 10 distinct textual descriptions and professionally annotated semantic masks. These masks precisely partition the face into 19 attribute categories, encompassing both facial features (e.g., skin, eyes, mouth) and accessories (e.g., earrings, necklaces). In our experiments, the MM-CelebA-HQ dataset was divided into training and test sets at a ratio of 9:1, which were respectively utilized for model training and performance evaluation. In addition, we employed a pre-trained facial parser Zheng et al. (2022) to process the FFHQ-Text dataset Zhou (2021), obtaining facial masks. Furthermore, we process the FFHQ-Text dataset using a pre-trained face parser to obtain facial masks, based on which we construct the MM-FFHQ dataset. This new benchmark allows us to further evaluate the zero-shot generalization capability of EC²Face.

Evaluation Metrics. We employ widely recognized evaluation metrics to measure the performance of our model. Specifically, we utilize the CLIP score Radford et al. (2021) to assess the consistency between facial images and given textual conditions, while leveraging the accuracy metric to calculate the average accuracy of various attributes in facial images in mask conditions. Regarding image quality assessment, we adopt the Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018) metric to quantify the perceptual similarity between generated images and real images, and employ the Natural Image Quality Evaluator (NIQE) Mittal et al. (2012) to evaluate the naturalness quality of the synthesized facial images.

Implementation Details. All experiments were conducted on a server equipped with eight NVIDIA A100 GPUs (each with 80GB memory). We employed the AdamW optimizer with an initial learning rate of 1×10^{-4} . During training, each GPU processed a mini-batch of four samples, achieving an effective global batch size of 32 through multi-GPU parallel computation. The model underwent 5,000 optimization iterations with gradient accumulation enabled to ensure stable training conver-

gence. The hyperparameters in the loss function were set as $\alpha_1 = 0.5$ and $\alpha_2=0.5$. For reproducibility, we fixed the random seed to 42 across all experiments during inference. The sampling process utilized a Flow-Matching Euler Discrete scheduler Esser et al. (2024) with 28 discretization steps.

4.2 QUANTITATIVE COMPARISON

To thoroughly analyze the disparities in generation accuracy across different attribute regions, during the evaluation of mask-conditional consistency, we not only report the mean accuracy for overall attributes (OAA) but also separately calculate the mean accuracies for high-frequency attributes (HFA), normal-frequency attributes (NFA), and rare-frequency attributes (RFA) (refer to Appendix A for detailed attribute categorization).

Table 1 presents the quantitative comparison between the proposed EC²Face method and state-of-the-art approaches on the MM-CelebA-HQ benchmark dataset. The experiments demonstrate that our method significantly outperforms the competing methods in terms of both text and mask conditional consistency. Notably, in terms of the generation accuracy of rare attributes, our method nearly doubles that of the second-best method. Meanwhile, it also achieves state-of-the-arts performance in generated image quality. Furthermore, Table 2 showcases the zero-shot generalization experimental results on the MM-FFHQ dataset, further validating the superiority and robustness of the EC²Face method.

Table 1: Quantitative comparison with state-of-the-art methods on the MM-CelebA-HQ dataset.

Methods	Multimodal Conditional Consistency					Visual Quality	
	Text (% , $\uparrow$)	Mask (% , $\uparrow$)				LPIPS ($\downarrow$)	NIQE ($\downarrow$)
		OAA	HFA	NFA	RFA		
TediGAN Xia et al. (2021)	23.90	66.58	91.14	69.18	0.61	0.49	5.19
UaC Nair et al. (2023)	25.52	65.37	79.54	68.27	18.19	0.57	6.47
PixelFace+ Du et al. (2023b)	26.16	82.11	93.96	86.80	27.81	0.56	6.24
CoDiffusion Huang et al. (2023)	24.51	61.51	85.34	62.83	5.26	0.58	4.46
ControlNeXt Peng et al. (2024)	25.88	80.93	91.67	84.94	33.40	0.54	4.54
MM2Latent Meng et al. (2025)	24.61	65.62	85.75	69.18	2.27	0.51	3.74
EC²Face (Ours)	27.05	88.64	96.56	90.18	62.78	0.51	3.55

Table 2: Comparison of zero-shot generalization performance on the MM-FFHQ dataset.

Methods	Multimodal Conditional Consistency					Visual Quality	
	Text (% , $\uparrow$)	Mask (% , $\uparrow$)				LPIPS ($\downarrow$)	NIQE ($\downarrow$)
		OAA	HFA	NFA	RFA		
TediGAN Xia et al. (2021)	25.03	64.63	91.40	66.25	0.56	0.64	5.28
UaC Nair et al. (2023)	26.97	67.61	82.52	70.23	20.73	0.67	6.39
PixelFace+ Du et al. (2023b)	26.60	83.36	93.72	88.24	30.99	0.67	5.93
CoDiffusion Huang et al. (2023)	23.03	58.35	82.80	59.31	3.22	0.68	6.79
ControlNeXt Peng et al. (2024)	27.80	71.63	86.93	74.72	20.95	0.62	3.76
MM2Latent Meng et al. (2025)	26.37	63.25	84.97	66.00	1.93	0.64	3.83
EC²Face (Ours)	28.40	88.28	96.98	90.02	59.63	0.51	3.67

4.3 QUALITATIVE COMPARISON

Figure 3 presents a visual comparison of facial images synthesized by EC²Face and current state-of-the-art methods under the collaborative guidance of mask and text conditions. As clearly depicted

in the figure, the facial images generated by EC²Face achieve more precise semantic alignment between the given text and mask conditions, as exemplified by the hair-end details in the first row, the hat shape in the second row, the earring styles in the third and fourth rows, and the necklace features in the fifth and sixth rows. This indicates that our method demonstrates significant advantages in fine-grained control, such as response to rare attributes, while ensuring high visual fidelity. For more visualization examples, please refer to the Appendix.

4.4 ABLATION STUDIES

Effectiveness of Long-tail Adaptive Learning. We conducted ablation experiments to evaluate the effectiveness of the long-tail adaptive learning strategy (LTAL). The quantitative results in Table 3 demonstrate that this strategy significantly enhances the mask alignment accuracy for rare attribute regions, with the average accuracy increasing from 52.04% to 61.1%. Further analysis reveals that the strategy yields minimal improvement for high-frequency attributes, while boosting the average accuracy of normal-frequency attributes by 0.79%. Figure 4 provides a visual comparison that intuitively illustrates the improvement in accuracy for rare-frequency attributes.

Effectiveness of Explicit Conditional Consistency Guidance. Table 4 presents the results of item-by-item ablation experiments conducted on explicit conditional consistency guidance (ECCG). The experiments demonstrate that applying either the text consistency constraint $\mathcal{L}_{\text{text}}$ or the mask consistency constraint $\mathcal{L}_{\text{mask}}$ individually can enhance the model’s evaluation scores under corresponding conditions. Furthermore, when both constraints are concurrently applied, all metrics pertaining to multimodal conditional consistency demonstrate notable improvements, thereby providing conclusive evidence for the effectiveness of the explicit consistency guidance approach.

5 CONCLUSION

This study proposes a novel facial synthesis method named EC²Face, grounded on an explicit-consistency-guided diffusion transformer architecture, which enables mask-text multimodal collaborative facial generation. Relying on latents reverse estimation, we introduce a multimodal condition consistency guidance mechanism in the pixel space. This mechanism explicitly drives the latent representations to achieve precise semantic alignment with the multimodal conditions, effectively overcoming the suboptimal generation issues caused by relying solely on implicit modeling in the latent space. Meanwhile, we design a temporal modulation function to avoid constraint biases resulting from the unreliability of reverse estimation at large time steps. Additionally, we propose a long-tail adaptive learning strategy. This strategy, by gradient reweighting during flow matching, mitigates the modeling bias of diffusion models towards rare attributes, leading to an improvement of over 9.0% in average accuracy.

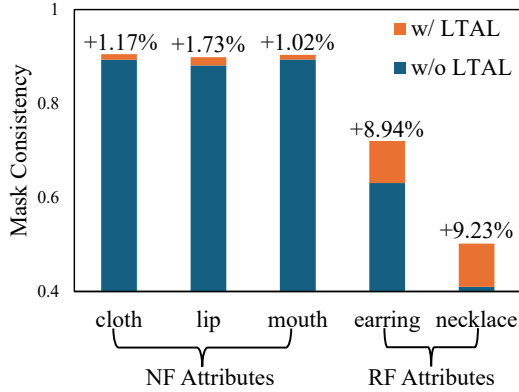


Figure 4: Accuracy enhancement of rare attributes through long-tail adaptive learning.

Table 3: Ablation experiment on the effectiveness of the proposed long-tail adaptive learning.

LTAL	Mask (% , $\uparrow$)			
	OAA	HFA	NFA	RFA
✗	86.14	95.13	88.62	52.04
✓	87.64	95.18	89.41	61.12 (+9.08)

Table 4: Ablation experiment on the effectiveness of explicit conditional consistency constraints.

ECCG		Text (% , $\uparrow$)	Mask (% , $\uparrow$)			
$\mathcal{L}_{\text{text}}$	$\mathcal{L}_{\text{mask}}$		OAA	HFA	NFA	RFA
✗	✗	26.12	87.64	95.18	89.41	61.62
✗	✓	26.26	88.59	96.50	90.14	62.68
✓	✗	27.22	87.48	95.36	89.16	60.60
✓	✓	27.05	88.64	96.56	90.18	62.78

ETHICAL STATEMENT

The multimodal face generation method proposed in this study is capable of high-fidelity facial image generation based on natural language descriptions and semantic masks, with anticipated applications in areas such as virtual character creation and secure identity verification. The positive motivation of this research lies in providing technological support for public safety and social well-being. However, we are fully aware that if this technology is misused, it could be employed to forge or disseminate fake facial images. Therefore, we strongly advocate for the implementation of rigorous human verification and professional supervision mechanisms, and emphasize that the model's output results should not be regarded as conclusive evidence. We encourage the research community to apply this technology in a responsible manner, ensuring its consistent compliance with ethical and legal standards.

REPRODUCIBILITY STATEMENT

We have included the complete source code for model implementation, training, and testing in the supplementary materials. We sincerely invite other researchers to replicate our experimental results based on this code and further expand our work, thereby jointly advancing technological development in the field of multimodal face generation.

THE USE OF LARGE LANGUAGE MODELS

In the course of this research and in preparing the manuscript, we utilized Large Language Models (LLMs) DeepSeek-AI (2024) in a limited capacity. Specifically, during the manuscript preparation phase, an LLM was used to assist in refining the wording and improving the clarity of the English prose. Its role in this capacity was strictly limited to enhancing sentence structure, grammar, and the overall flow of the text. Beyond this, LLMs were not involved in the research design, data collection, evaluation, or the generation of core scientific ideas. All substantive content, methodologies, and conclusions are entirely the original work of the authors.

REFERENCES

- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pp. 214–223, 2017.
- Black-Forest-Labs. FLUX. <https://github.com/black-forest-labs/flux>, 2024.
- Ziyi Chang, George A Koulouris, Hyung Jin Chang, and Hubert PH Shum. On the design fundamentals of diffusion models: A survey. *Pattern Recognition*, pp. 111934, 2025.
- Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023.
- DeepSeek-AI. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- Xiaoxiong Du, Jun Peng, Yiyi Zhou, Jinlu Zhang, Siting Chen, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Pixelface+: Towards controllable face generation and manipulation with text descriptions and segmentation masks. In *ACM International Conference on Multimedia*, pp. 4666–4677, 2023a.
- Xiaoxiong Du, Jun Peng, Yiyi Zhou, Jinlu Zhang, Siting Chen, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Pixelface+: Towards controllable face generation and manipulation with text descriptions and segmentation masks. In *ACM MM*, pp. 4666–4677, 2023b.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

- Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, pp. 1–8, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *CVPR*, pp. 6080–6090, 2023.
- Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pp. 4401–4410, 2019.
- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pp. 8110–8119, 2020.
- Jihyun Kim, Changjae Oh, Hoseok Do, Soohyun Kim, and Kwanghoon Sohn. Diffusion-driven gan inversion for multi-modal face image generation. In *CVPR*, pp. 10403–10412, 2024a.
- Jihyun Kim, Changjae Oh, Hoseok Do, Soohyun Kim, and Kwanghoon Sohn. Diffusion-driven gan inversion for multi-modal face image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10403–10412, 2024b.
- Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5549–5558, 2020.
- Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pp. 22511–22521, 2023.
- Andrew Melnik, Maksim Miasayedzenkau, Dzianis Makaravets, Dzianis Pirshutuk, Eren Akbulut, Dennis Holzmann, Tarek Renusch, Gustav Reichert, and Helge Ritter. Face generation and editing with stylegan: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (5):3557–3576, 2024.
- Debin Meng, Christos Tzelepis, Ioannis Patras, and Georgios Tzimiropoulos. MM2Latent: Text-to-facial image generation and editing in GANs with multimodal assistance. In *European Conference on Computer Vision*, pp. 88–106. Springer, 2025.
- Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.
- Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.
- Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, pp. 4296–4304, 2024a.
- Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, volume 38, pp. 4296–4304, 2024b.
- Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. Unite and conquer: Plug & play multi-modal synthesis using diffusion models. In *CVPR*, pp. 6070–6079, 2023.

- Xin Ning, Fangzhe Nan, Shaohui Xu, Lina Yu, and Liping Zhang. Multi-view frontal face image generation: A survey. *Concurrency and Computation: Practice and Experience*, 35(18):e6147, 2023.
- Minho Park, Jooyeol Yun, Seunghwan Choi, and Jaegul Choo. Learning to generate semantic layouts for higher text-image correspondence in text-to-image synthesis. In *ICCV*, pp. 7591–7600, 2023.
- Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pp. 2085–2094, 2021.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.
- Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.
- Justin N. M. Pinkney and Chuan Li. clip2latent: Text driven sampling of a pre-trained stylegan using denoising diffusion and clip. In *BMVC*, pp. 1–12, 2022.
- Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit Bermano, Eric Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. In *Computer Graphics Forum*, volume 43, pp. e15063. Wiley Online Library, 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022a.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022b.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29:1–9, 2016.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, pp. 1–20, 2021.
- BJ Sowmya and Seems Shedole Meeradevi. Generative adversarial networks with attentional multimodal for human face synthesis. *Indonesian Journal of Electrical Engineering and Computer Science*, 33(2):1205–1215, 2024.
- Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *CVPR*, pp. 7962–7971, 2021.
- Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.
- Chaoyang Wang, Xiangtai Li, Lu Qi, Henghui Ding, Yunhai Tong, and Ming-Hsuan Yang. Semflow: Binding semantic segmentation and image synthesis via rectified flow. In *NeurIPS*, volume 37, pp. 138981–139001, 2024.
- Jiayu Wang, Yue Yu, Jingjing Chen, Qi Dai, and Yu-Gang Jiang. FaceA-Net: Facial Attribute-Driven ID Preserving Image Generation Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7736–7743, 2025a.
- Zhaoqing Wang, Xiaobo Xia, Runnan Chen, Dongdong Yu, Changhu Wang, Mingming Gong, and Tongliang Liu. Lavin-dit: Large vision diffusion transformer. In *Computer Vision and Pattern Recognition Conference*, pp. 20060–20070, 2025b.

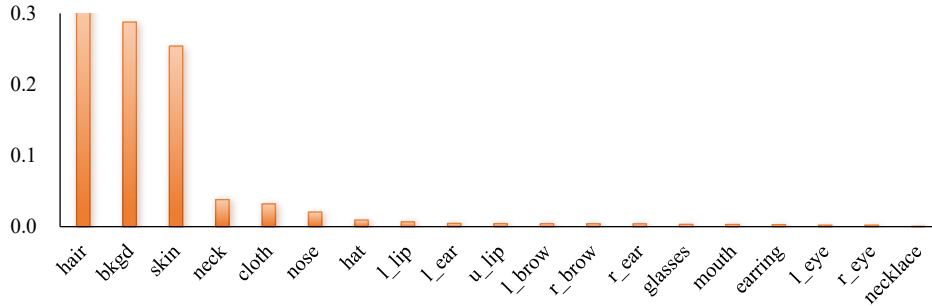
- Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. E2style: Improve the efficiency and effectiveness of stylegan inversion. *TIP*, 31: 3267–3280, 2022.
- Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, pp. 2256–2265, 2021.
- Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- Jinlu Zhang, Yiyi Zhou, Qiancheng Zheng, Xiaoxiong Du, Gen Luo, Jun Peng, Xiaoshuai Sun, and Rongrong Ji. Fast text-to-3d-aware face generation and manipulation via direct cross-modal mapping and geometric regularization. In *ICML*, pp. 60605–60625, 2024.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pp. 3836–3847, 2023a.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023b.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.
- Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18697–18709, 2022.
- Yutong Zhou. Generative adversarial network for text-to-face synthesis and manipulation. In *ACM International Conference on Multimedia*, pp. 2940–2944, 2021.

A PRE-STATISTICAL ANALYSIS AND SUBSET PARTITIONING OF SEMANTIC ATTRIBUTES IN MASK CONDITIONS.

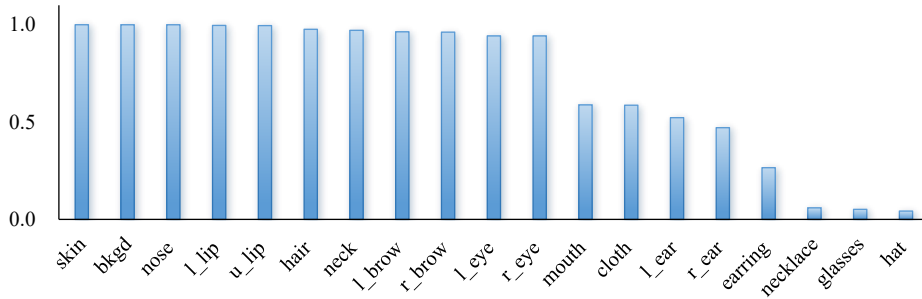
The mask images in the MM-CelebA-HQ dataset Lee et al. (2020) predefine 19 typical facial semantic attributes. Figure 5 presents statistical analyses of the average pixel proportion of each attribute within the images and their occurrence frequencies across the dataset, revealing a pronounced long-tailed distribution of semantic attributes—manifesting as severe imbalances in both pixel proportion and occurrence frequency. However, existing methods generally lack optimization strategies tailored to this distributional characteristic, resulting in suboptimal generation performance for rare-frequency attributes.

To systematically evaluate the generation quality across different attribute regions, this study categorizes the 19 semantic attributes into three groups based on the statistical results: High-Frequency Attributes (HFA), Normal-Frequency Attributes (NFA), and Rare-Frequency Attributes (RFA). The classification criteria are as follows: an attribute is assigned to the HFA group if it exhibits both high average pixel proportion and high occurrence frequency; to the RFA group if it demonstrates low values in both metrics; and the remaining attributes are classified as NFA. The specific compositions of the three attribute groups are detailed as follows:

- **HFA** (High-Frequency Attributes): background, hair, skin, neck.
- **NFA** (Normal-Frequency Attributes): cloth, nose, l_lip, u_lip, l_ear, r_ear, l_brow, r_brow, mouth, l_eye, r_eye, hat, glasses.
- **RFA** (Rare-Frequency Attributes): earring, necklace.



(a) Average pixel proportion of attributes



(b) Frequency distribution of attributes

Figure 5: The long-tailed distribution of semantic attributes in mask conditions.

B MORE EXAMPLES OF MULTIMODAL FACIAL GENERATION.

Figure 6 presents more cases of multimodal facial synthesis. On the left side of each row are the given text and mask conditions, while on the right side are the generated results of TediGAN Xia et al. (2021), UaC Nair et al. (2023), PixelFace+ Du et al. (2023b), CoDiffusion Huang et al. (2023), ControlNeXt Peng et al. (2024), MM2Latent Meng et al. (2025), and our proposed EC²Face in



Figure 6: Visual comparison of additional multimodal facial synthesis samples.

sequence. Compared with other competing methods, EC²Face exhibits superior performance in detail control, with the areas of significant improvement marked by red boxes in the figure. This indicates that, given the text and mask conditions, EC²Face can generate facial images that are more semantically aligned with the two modalities (text and image mask) and can more clearly respond to the conditional signals of rare attributes (e.g., earrings, necklaces).

C COMPATIBILITY OF UNIMODAL-DRIVEN FACIAL GENERATION

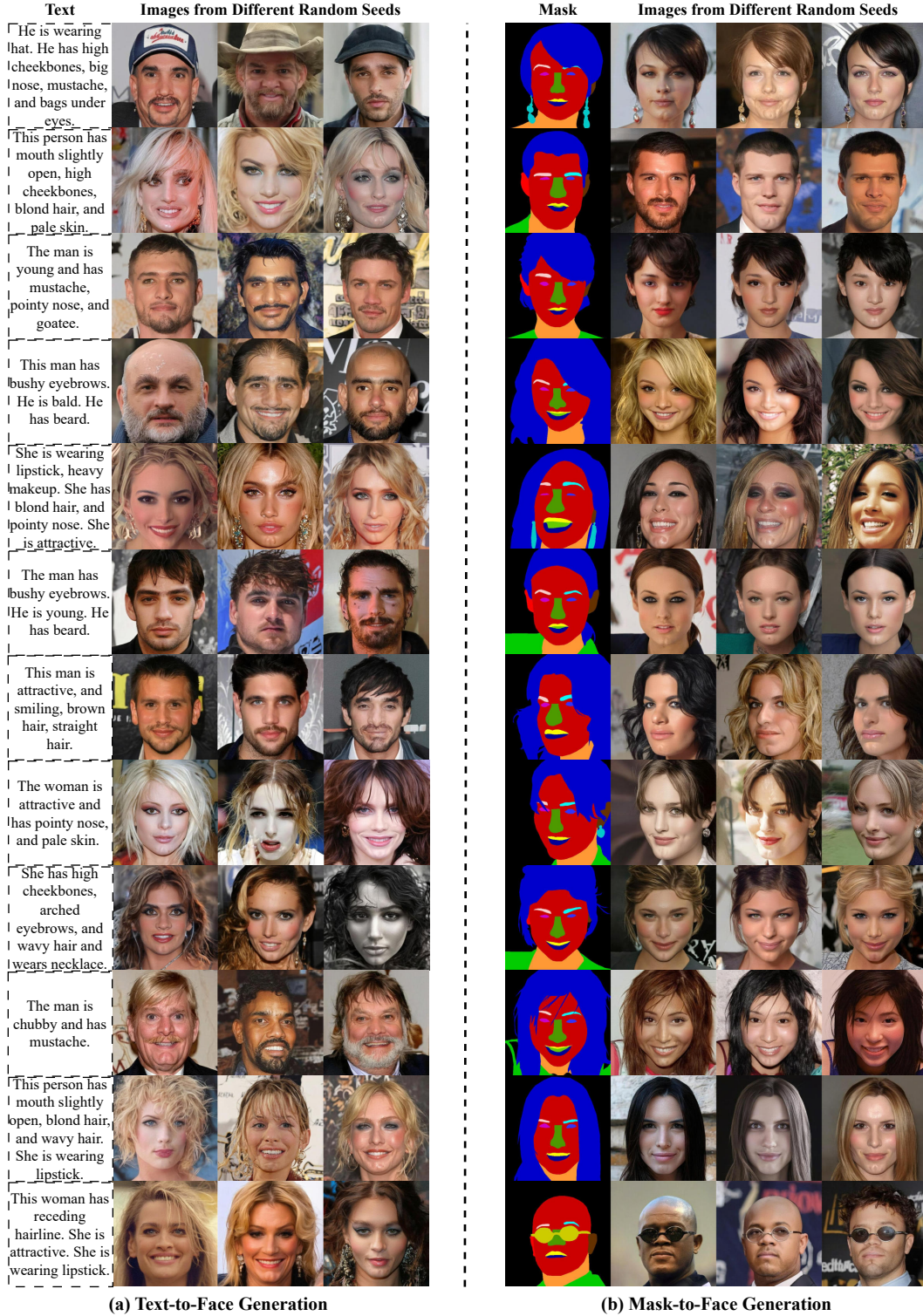
EC²Face simultaneously supports facial generation tasks under both multimodal collaborative and unimodal (text or mask) conditions, without requiring any modification to the model architecture. This capability stems from two core design principles:

1. Explicit conditional consistency guidance in pixel space: By strengthening semantic alignment between latent representations and input conditions, this mechanism ensures generated results strictly adhere to the specified controllable conditions, regardless of whether they are unimodal or multimodal.
2. Stochastic conditional dropout during training: By simulating unimodal input scenarios, this strategy equips the model with robustness against missing modalities, enabling flexible adaptation to diverse conditional combinations during generation.

Experimental validation, as illustrated in Figure 7, demonstrates that EC²Face maintains high-quality facial generation even when provided solely with textual descriptions or mask images. By varying random seeds, the model preserves output diversity while ensuring both semantic consistency and visual fidelity.

D GENERATIVE DIVERSITY UNDER CONDITIONAL CONSISTENCY.

We conducted an in-depth exploration of the multimodal generation diversity of EC²Face. In the experiments, with a fixed set of text and mask conditions, we employed different random seeds to drive the model to generate facial images, and some representative samples are shown in Figure 8. The experimental results demonstrate that while ensuring precise semantic alignment between the generated results and the given conditions, EC²Face also exhibits remarkable diversity in unspecified attribute dimensions, covering aspects such as identity traits, hair color, skin tone, earring styles, and hat appearances. This characteristic offers broad application prospects in practical scenarios like data augmentation and interactive creation.

Figure 7: Compatibility of unimodal-driven face generation using EC²Face.

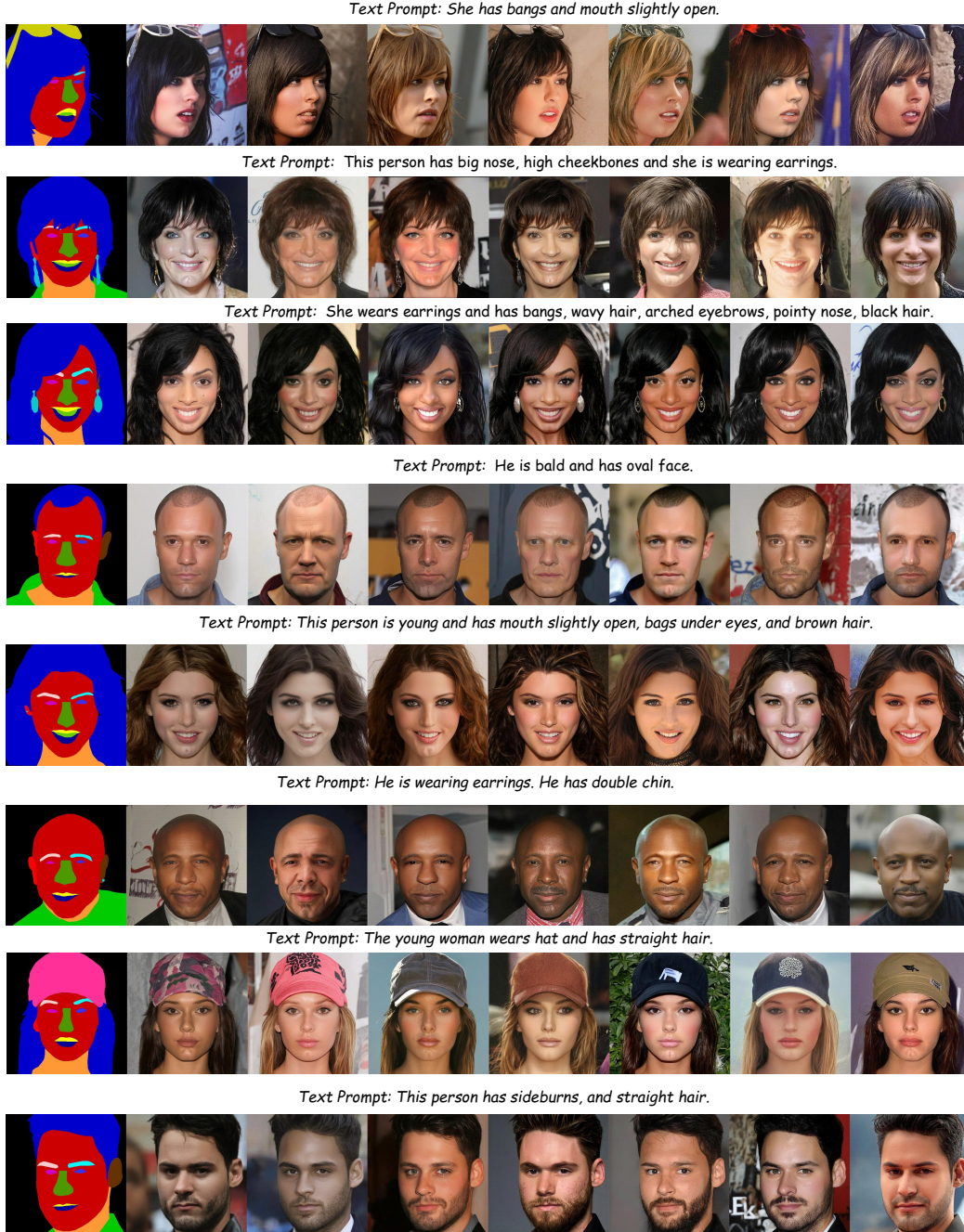


Figure 8: Facial generation diversity under multimodal consistency.