MMFace-DiT: A Dual-Stream Diffusion Transformer for High-Fidelity Multimodal Face Generation

Supplementary Material

1. Implementation Details

Our Dual-Stream MMFace-DiT follows a DiT-XL [7] configuration with 1.345 billion parameters, including 28 transformer blocks, a hidden size of 1152, and 16 attention heads. The model operates with a patch size of 2 in the latent space of the 16-channel FLUX VAE, forming a 32-channel input tensor (16 for the noisy latent and 16 for the spatial condition). Key to our architecture is a shared attention mechanism implementing *2D Rotary Position Embedding (RoPE)*, which uses a base period (θ = 10,000) to encode positional information in the query and key tensors of both streams. Dynamic conditioning is driven by a *Modality Embedder*, a learnable lookup layer that converts the discrete spatial condition (mask or sketch) into a dense vector. This vector enriches the global conditioning signal, from which our AdaLN scheme derives twelve distinct parameter vectors per block to control the shift, scale, and gate of both the attention and MLP sub-layers. Training proceeds progressively using either a DDPM objective with Min-SNR weighting or a Rectified Flow Matching (RFM) objective. We first train from scratch at 256×256 resolution for 300 epochs with a batch size of 32 per GPU, a learning rate of 1×10^{-4}, then fine-tune at 512×512 for 50 epochs with an effective batch size of 16 (via 2-step gradient accumulation) per GPU and a learning rate of 1×10^{-6}. To enable classifier-free guidance, we apply a 5% dropout probability to both the text and spatial conditioning inputs during training. We employ the 8-bit AdamW optimizer with a cosine learning rate scheduler and ensure memory efficiency using `bfloat16` mixed precision, full gradient checkpointing, and an Exponential Moving Average (EMA) of model weights with a decay of 0.9999. The complete details for all the hyperparameters are provided in Table 1.

2. Training Objectives and Inference

Our MMFace-DiT architecture is compatible with two distinct generative training paradigms: Denoising Diffusion Probabilistic Modeling (DDPM) [4] and Rectified Flow Matching (RFM) [5]. Both approaches leverage the same core model and conditioning mechanisms but differ in their optimization objective and sampling procedure. To rigorously evaluate the efficacy and generalization of our method, all inference and sampling procedures described herein are conducted exclusively on the official CelebA-HQ test split of $6,000$ images, a held-out partition on which our model was not trained.

Algorithm 1 MMFace-DiT Training with Diffusion Objective

Input: DiT model ϵ_θ, VAE Encoder $\mathcal{E}_{vae}$, CLIP Text Encoder $\mathcal{E}_{text}$, Noise Scheduler $\mathcal{S}$
Require: Dataset D of triplets (x, c_{sp}, p) and modality flag m
repeat
  Sample a batch (x, c_{sp}, p, m) from D
  $z_0 \leftarrow \mathcal{E}_{vae}(x)$; $z_c \leftarrow \mathcal{E}_{vae}(c_{sp})$ ▷ Encode images to latents
  $c_{pooled}, c_{seq} \leftarrow \mathcal{E}_{text}(p)$ ▷ Encode text prompt
  Sample timestep $t \sim \mathcal{U}\{1, ..., T\}$ and noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  $z_t \leftarrow \mathcal{S}$.add_noise$(z_0, \epsilon, t)$ ▷ Apply forward diffusion process
  $z_{in} \leftarrow$ concat(z_t, z_c) ▷ Concatenate noisy latent and spatial condition
  Form global conditioning C_{global} from t, c_{pooled}, m
  $\epsilon_{pred} \leftarrow \epsilon_\theta(z_{in}, t, C_{global}, c_{seq})$ ▷ Predict noise from all conditions
  Calculate Min-SNR weight $w(t)$ based on scheduler alphas at timestep t
  $\mathcal{L} \leftarrow w(t) \cdot \mathbb{E}[\|\epsilon - \epsilon_{pred}\|^2]$ ▷ Compute weighted MSE loss
  Update model parameters θ using gradient descent on $\mathcal{L}$
until converged

2.1. DDPM Objective

Under the DDPM framework, the model ϵ_θ is trained to predict the noise ϵ added to a clean latent z_0 at a given timestep t. The model is conditioned on a concatenated input tensor z_{in} (containing the noised latent z_t and the spatial condition latent z_c), a global conditioning signal C_{global} (derived from the timestep t, pooled text embedding c_{pooled}, and modality flag m), and the text sequence embedding c_{seq}. We use a Min-SNR weighting strategy [2] to stabilize training by re-weighting the loss at each timestep. The complete training and inference procedures, which employ the efficient DPM-Solver Multistep scheduler [6] with Classifier-Free Guidance (CFG) [3], are detailed in Algorithm 1 and Algorithm 2, respectively.

2.2. RFM Objective

In the RFM paradigm, the model v_θ learns to predict the constant-velocity vector $v = z_1 - z_0$ that connects a noise

Hyperparameter	Value (Stage 1: 256x256)	Value (Stage 2: 512x512 Fine-tuning)
Model Architecture		
Hidden Size (D)	1152	1152
Depth (Transformer Blocks)	28	28
Attention Heads	16	16
Patch Size	2	2
Input Channels	32 (16 for latent, 16 for condition)	32 (16 for latent, 16 for condition)
MLP Ratio	4.0	4.0
RoPE Theta (θ)	10,000	10,000
External Components		
VAE	FLUX VAE (16-channel)	FLUX VAE (16-channel)
Text Encoder	CLIP (from SD 2.1-base)	CLIP (from SD 2.1-base)
Training Strategy		
Training Objectives	DDPM (Min-SNR) / RFM	DDPM (Min-SNR) / RFM
Epochs	300	50
Steps	440700	146900
Batch Size (per GPU)	32	8
Gradient Accumulation Steps	1	2
Effective Batch Size	$\mathbf{32 \times 2}$	$\mathbf{8 \times 2 \times 2}$
Optimizer (8-bit AdamW)		
Learning Rate	1×10^{-4}	1×10^{-6}
LR Scheduler	Cosine Decay	Cosine Decay
LR Warmup Steps	200	100
Adam β_1	0.9	0.9
Adam β_2	0.999	0.999
Weight Decay	0.01	0.01
Adam ϵ	1×10^{-8}	1×10^{-8}
Regularization & Efficiency		
Max Gradient Norm	0.5	0.5
Conditioning Dropout	5% (Text, Mask, Sketch)	5% (Text, Mask, Sketch)
EMA Decay	0.9999	0.9999
Mixed Precision	‘bfloat16‘	‘bfloat16‘
Gradient Checkpointing	Enabled	Enabled
Dataloader Workers	4	4

Table 1. Detailed hyperparameters for the two-stage progressive training of our MMFace-DiT model. Stage 1 trains the model from scratch at 256x256 resolution, and Stage 2 fine-tunes the resulting checkpoint at 512x512 resolution.

sample $z_0 \sim \mathcal{N}(0, \mathbf{I})$ to a data sample z_1. The model is trained on an interpolated latent $z_t = (1-t)z_0 + tz_1$ and receives the same set of conditioning inputs as in the DDPM setup (z_c, C_{global}, and c_{seq}) to predict the target velocity. Inference involves integrating the predicted velocity field from $t = 0$ to $t = 1$ using an ODE solver like the Euler method, also guided by CFG. The specific training and inference steps are outlined in Algorithm 3 and Algorithm 4.

3. VLM-Powered Data Enrichment Pipeline

A core contribution of our work is the creation of a large-scale, high-quality, and diverse set of textual annotations for the FFHQ and CelebA-HQ datasets. The performance of modern controllable generative models is critically dependent on the quality of their training data [1]. We identified two primary data-related bottlenecks in the field: 1) the complete lack of captions for the 70,000 high-resolution

Algorithm 2 MMFace-DiT Inference with Diffusion Objective

Input: Prompt p, spatial condition c_{sp}, modality flag m, guidance scale ω
Require: Trained DiT model ϵ_θ, VAE Encoder/Decoder $\mathcal{E}_{vae}, \mathcal{D}_{vae}$, CLIP Encoder $\mathcal{E}_{text}$, Scheduler $\mathcal{S}$
$z_c \leftarrow \mathcal{E}_{vae}(c_{sp})$ ▷ Encode spatial condition
$c^{cond}_{pooled}, c^{cond}_{seq} \leftarrow \mathcal{E}_{text}(p)$ ▷ Encode conditional prompt
$c^{uncond}_{pooled}, c^{uncond}_{seq} \leftarrow \mathcal{E}_{text}(\text{" "})$ ▷ Encode unconditional (null) prompt
Sample initial latent $z_T \sim \mathcal{N}(0, \mathbf{I})$
Set scheduler timesteps $\mathcal{T} = \{T, T-1, ..., 1\}$
for t in $\mathcal{T}$ **do**
  $z^{in}_t \leftarrow [z_t; z_t]$ ▷ Duplicate latent for CFG
  $z_{in} \leftarrow \text{concat}(z^{in}_t, [z_c; z_c])$ ▷ Concatenate with spatial condition
  Form global conditioning $C^{cond}_{global}, C^{uncond}_{global}$
  $\epsilon_{pred} \leftarrow \epsilon_\theta(z_{in}, t, [C^{uncond}_{global}; C^{cond}_{global}], [c^{uncond}_{seq}; c^{cond}_{seq}])$
  $\epsilon_{uncond}, \epsilon_{cond} \leftarrow \text{split}(\epsilon_{pred})$
  $\epsilon_{final} \leftarrow \epsilon_{uncond} + \omega \cdot (\epsilon_{cond} - \epsilon_{uncond})$ ▷ Apply guidance
  $z_{t-1} \leftarrow \mathcal{S}.\text{step}(\epsilon_{final}, t, z_t)$ ▷ Scheduler denoising step
end for
$x_{out} \leftarrow \mathcal{D}_{vae}(z_0)$ ▷ Decode final latent to image
return x_{out}

Algorithm 3 MMFace-DiT Training with Rectified Flow Matching Objective

Input: DiT model v_θ, VAE Encoder $\mathcal{E}_{vae}$, CLIP Text Encoder $\mathcal{E}_{text}$
Require: Dataset D of triplets (x, c_{sp}, p) and modality flag m
repeat
  Sample a batch (x, c_{sp}, p, m) from D
  $z_1 \leftarrow \mathcal{E}_{vae}(x)$; $z_c \leftarrow \mathcal{E}_{vae}(c_{sp})$ ▷ Encode images to data latents
  $c_{pooled}, c_{seq} \leftarrow \mathcal{E}_{text}(p)$ ▷ Encode text prompt
  Sample noise latent $z_0 \sim \mathcal{N}(0, \mathbf{I})$
  $v_{target} \leftarrow z_1 - z_0$ ▷ Define the target velocity vector
  Sample time $t \sim \mathcal{U}[0, 1]$
  $z_t \leftarrow (1-t)z_0 + tz_1$ ▷ Create interpolated latent on the flow path
  $z_{in} \leftarrow \text{concat}(z_t, z_c)$ ▷ Concatenate interpolated latent and spatial condition
  Form global conditioning C_{global} from t, c_{pooled}, m
  $v_{pred} \leftarrow v_\theta(z_{in}, t, C_{global}, c_{seq})$ ▷ Predict velocity from all conditions
  $\mathcal{L} \leftarrow \mathbb{E}[\|v_{target} - v_{pred}\|^2]$ ▷ Compute MSE loss
  Update model parameters θ using gradient descent on $\mathcal{L}$
until converged

images in the FFHQ dataset, and 2) the limited semantic richness of existing captions for CelebA-HQ.

To overcome these limitations, we designed a sophisticated, two-stage automated annotation pipeline detailed below.

3.1. Stage 1: VLM-Based Caption Generation

The Multi-Prompt Strategy. The first stage uses the powerful **InternVL3** Vision-Language Model [9] to generate a base set of captions. Instead of using a single generic prompt, we developed a systematic multi-prompt strategy. For each image, we query the VLM with ten uniquely engineered prompts, each designed to elicit a different style and focus of information. This careful prompt engineering ensures that our final dataset captures a wide range of attributes. The prompts include:

- **Few-shot Descriptive Prompts:** These ask for a standard, concise description, providing examples to guide the model's output format (e.g., "A professional headshot of a woman with medium-length curly brown hair...").
- **Structured Demographic Templates:** These prompts explicitly ask the VLM to fill in perceived demographic details, capturing information often missing from simple descriptions (e.g., "A [photo style] of a [age group] [gender] with [racial/ethnic appearance] features...").
- **Keyword-focused Prompts:** These request comma-separated keywords, ideal for capturing salient objects and attributes (e.g., "woman, long blonde hair, smiling, red dress...").
- **Detail-Oriented Prompts:** These focus on fine-grained features like accessories or specific makeup details.

This strategy ensures that our dataset contains a rich and diverse set of textual descriptions for every single image, forming a robust foundation for the next stage.

3.2. Stage 2: LLM-Based Post-Processing and Augmentation

The Two-Stage Refinement. The raw text generated by the VLM often contains instructional artifacts, prompt remnants, or subtle factual inconsistencies (hallucinations). To address this, we developed a rigorous two-stage post-processing pipeline.

First, a rule-based cleaning script is applied to every raw caption. This script performs programmatic sanitation by stripping common instructional prefixes and suffixes (e.g., "Generate a caption for the image above:", "Max 75 tokens."), removing unfilled template placeholders (e.g., "[clothing/accessories]"), and standardizing punctuation and capitalization.

Algorithm 4 MMFace-DiT Inference with Rectified Flow Matching Objective

Input: Prompt p, spatial condition c_{sp}, modality flag m, guidance scale ω, number of steps N
Require: Trained DiT model v_θ, VAE Encoder/Decoder $\mathcal{E}_{vae}, \mathcal{D}_{vae}$, CLIP Encoder $\mathcal{E}_{text}$
$z_c \leftarrow \mathcal{E}_{vae}(c_{sp})$ ▷ Encode spatial condition
$c^{cond}_{pooled}, c^{cond}_{seq} \leftarrow \mathcal{E}_{text}(p)$ ▷ Encode conditional prompt
$c^{uncond}_{pooled}, c^{uncond}_{seq} \leftarrow \mathcal{E}_{text}(\text{" "})$ ▷ Encode unconditional (null) prompt
Sample initial latent $z_0 \sim \mathcal{N}(0, \mathbf{I})$
Set step size $\Delta t \leftarrow 1/N$
Let $z \leftarrow z_0$
for $i = 0$ to $N - 1$ **do**
    $t \leftarrow i \cdot \Delta t$ ▷ Current time
    $z^{in} \leftarrow [z; z]$ ▷ Duplicate latent for CFG
    $z_{in} \leftarrow \text{concat}(z^{in}, [z_c; z_c])$ ▷ Concatenate with spatial condition
    Form global conditioning $C^{cond}_{global}, C^{uncond}_{global}$
    $v_{pred} \leftarrow v_\theta(z_{in}, t, [C^{uncond}_{global}; C^{cond}_{global}], [c^{uncond}_{seq}; c^{cond}_{seq}])$
    $v_{uncond}, v_{cond} \leftarrow \text{split}(v_{pred})$
    $v_{final} \leftarrow v_{uncond} + \omega \cdot (v_{cond} - v_{uncond})$ ▷ Apply guidance to velocity
    $z \leftarrow z + v_{final} \cdot \Delta t$ ▷ Euler method ODE step
end for
$x_{out} \leftarrow \mathcal{D}_{vae}(z)$ ▷ Decode final latent at t=1 to image
return x_{out}

Second, we leverage a powerful Large Language Model, **Qwen3** [8], to conduct the final, intelligent post-processing. This LLM stage serves a crucial dual role:

1. **Refinement and Hallucination Mitigation:** The LLM is tasked with reviewing and rephrasing the cleaned VLM captions to improve their grammatical structure and coherence. By using the full set of cleaned captions for an image as context, the LLM can identify and correct subtle factual inconsistencies that a rule-based script would miss, thereby mitigating hallucinations.
2. **Generation and Gap-Filling:** In cases where the initial VLM prompting and rule-based cleaning did not result in ten unique, high-quality captions, the LLM is prompted to generate novel captions. It is strictly constrained to use only the factual information present in the existing valid captions for that image, ensuring that the new captions are factually consistent while increasing the diversity of the phrasing.

Finally, this entire pipeline strictly enforces that every caption is below the 77-token limit compatible with CLIP's context length. This process results in a dense, high-fidelity, and diverse textual annotation layer. The final dataset, containing 10 high-quality captions for each of the 100,000 images from FFHQ and CelebA-HQ (totaling 1 million captions), is being released publicly to benefit the research community.

4. Qualitative Visualizations

We provide extensive qualitative results to complement our quantitative analysis. Fig. 1 and Fig. 2 highlight the model's capacity for fine-grained, disentangled attribute control by varying a single word in the text prompt while keeping the spatial condition fixed. Fig. 3 and Fig. 4 compare the high-quality outputs from the Diffusion and Flow training paradigms, demonstrating the model's effectiveness with both objectives. Additionally, Fig. 5 and Fig. 6 provide the visual evidence for our VAE ablation study, illustrating the superior perceptual quality and artifact reduction achieved with the Flux VAE compared to other backbones. Finally, Fig. 7 and Fig. 8 demonstrate the substantial impact of our VLM-powered data enrichment, illustrating that our semantically rich captions yield superior photorealism and detail compared to the original sparse annotations.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2

[2] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7441–7451, 2023. 1

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[5] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 1

[6] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022. 1

[7] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 1

[8] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 4

[9] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and

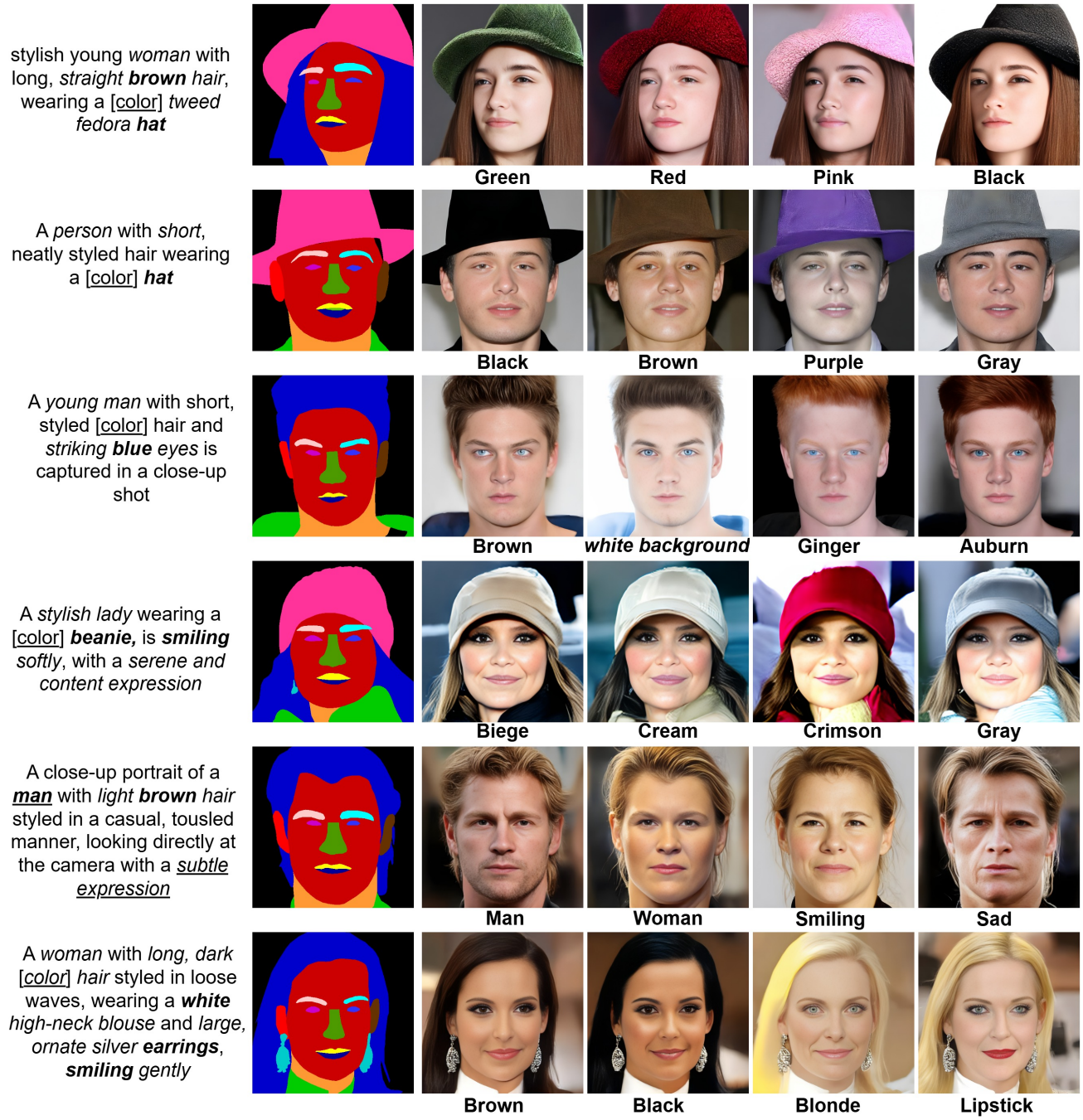

Figure 1. **Demonstration of Disentangled Fine-Grained Attribute Control.** Our MMFace-DiT exhibits exceptional disentangled control over the synthesis process. Each row is generated from a single, fixed segmentation mask, where we systematically vary a single keyword in the text prompt. The model accurately synthesizes diverse attributes—including color (hats, hair), expression (smiling, sad), gender, and even semantic concepts like background details, showcasing our model's advanced capability for precise, text-guided semantic generation.

test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3

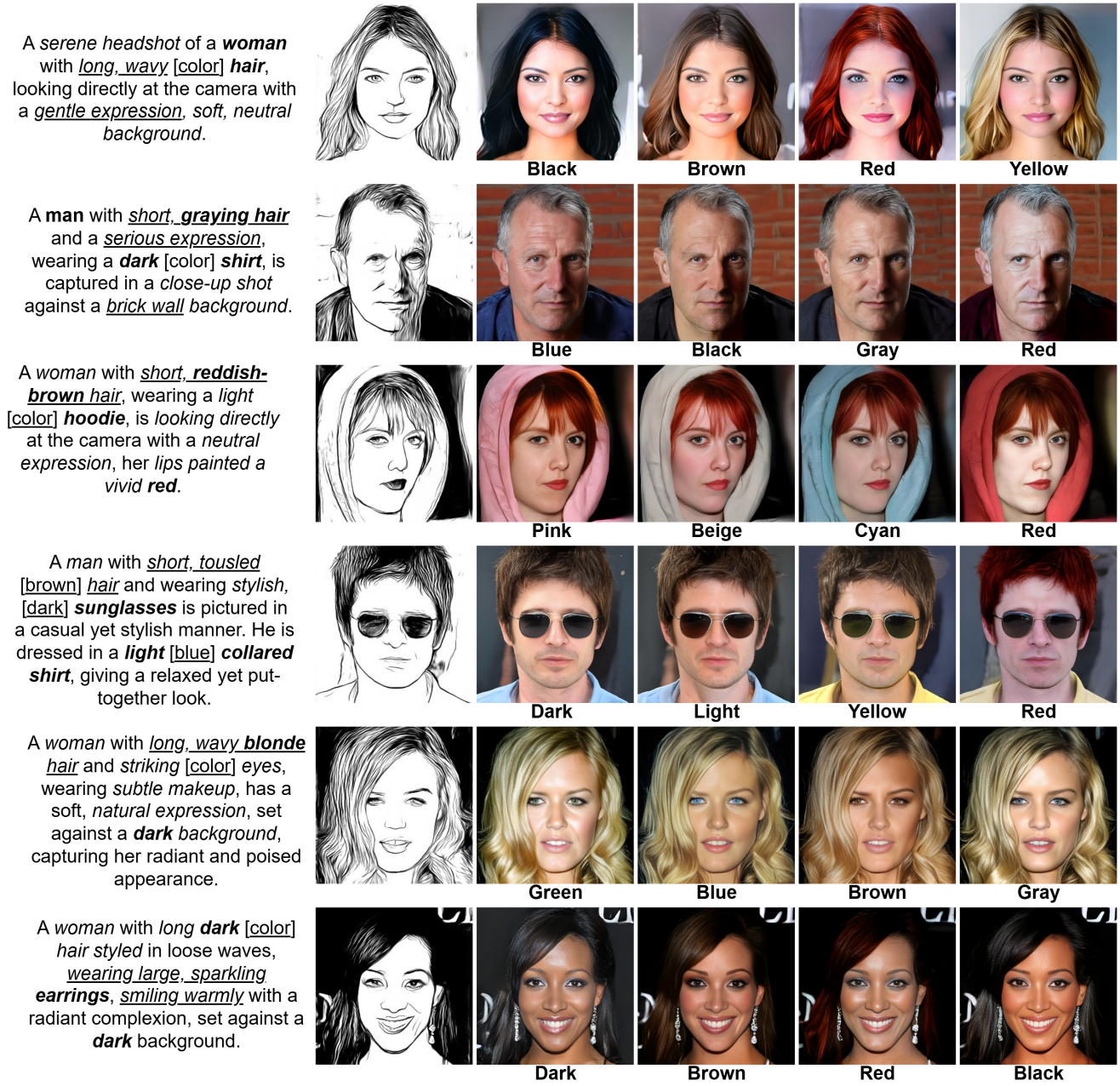

Figure 2. **Disentangled Attribute Control via Sketch-Conditioned Generation.** MMFace-DiT exhibits fine-grained disentanglement in multimodal synthesis when guided by sketch-based spatial priors. Each row is generated from a single fixed sketch while systematically varying a single textual attribute (e.g., hair color, shirt color, eye color). The model precisely follows the specified text-based edits while preserving identity, expression, and geometric consistency dictated by the sketch. This demonstrates MMFace-DiT's capability for precise semantic integration with strong geometric priors.

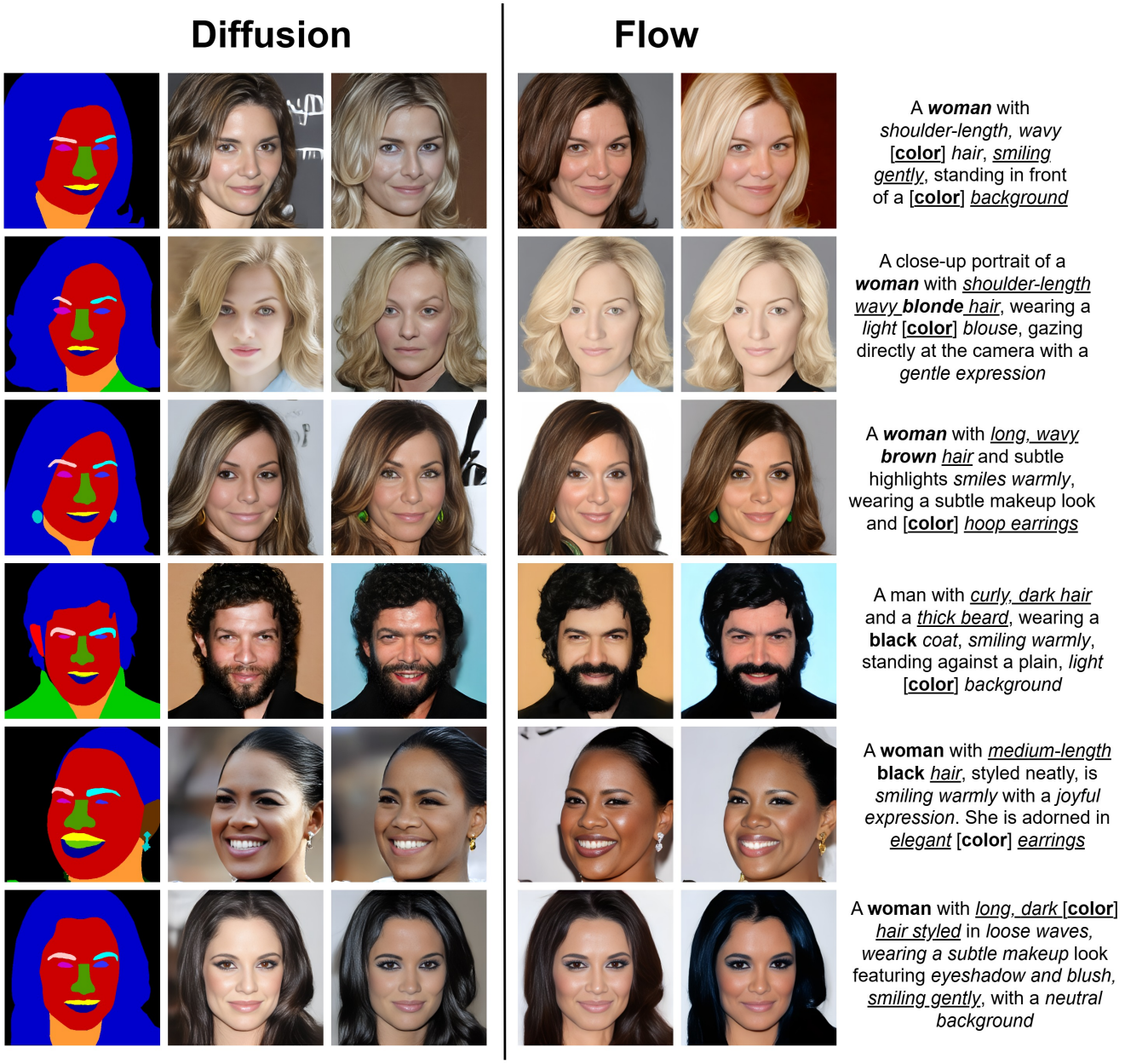

Figure 3. **Mask-Conditioned Synthesis with Diffusion and Flow Paradigms.** This figure showcases the high-quality performance of our MMFace-DiT model when trained under both Diffusion and Rectified Flow (Flow) objectives. Each example is generated from an identical segmentation mask (far left) and text prompt (far right), demonstrating the model's robust ability to synthesize diverse and realistic portraits that align with both spatial and semantic guidance. Both training paradigms yield excellent results, successfully interpreting complex attributes like hair style, expression, and accessories. Notably, the Flow-based model often exhibits a particularly refined level of photorealism, producing images with remarkably consistent lighting, skin texture, and fine-grained detail.

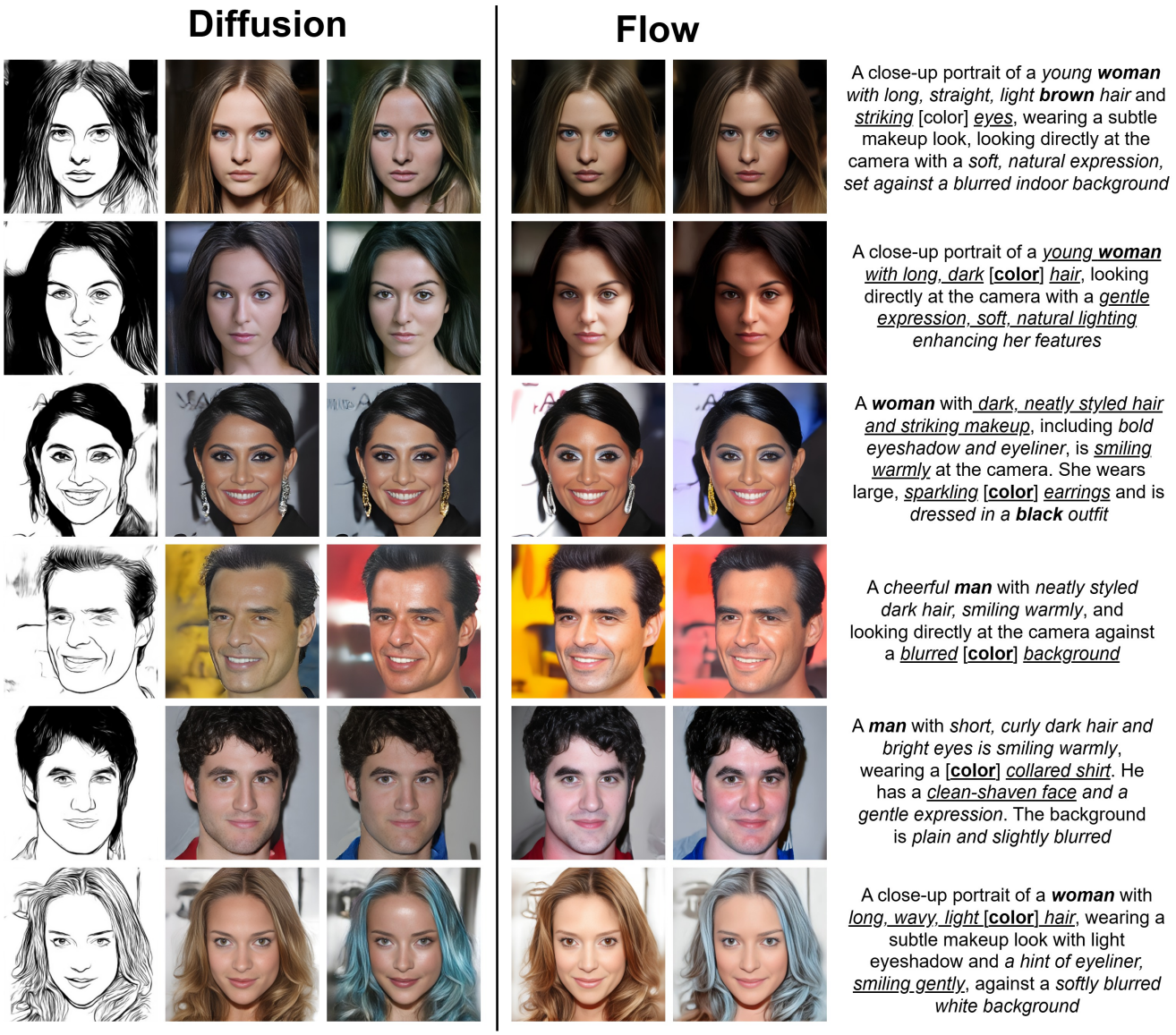

Figure 4. **Sketch-Conditioned Synthesis with Diffusion and Flow Paradigms.** This figure highlights the versatility of the MMFace-DiT architecture in translating artistic sketches into photorealistic faces, comparing results from both the Diffusion and Flow training frameworks. Conditioned on the same input sketch (far left) and text prompt (far right), both models successfully preserve the core identity, pose, and expression of the sketch while integrating the specified textual attributes. This demonstrates the model's strong multimodal capabilities regardless of the training objective. The Flow-based generations, in particular, show a strong proficiency in maintaining structural fidelity to the sketch, resulting in outputs that seamlessly blend the artistic input with photorealistic rendering.

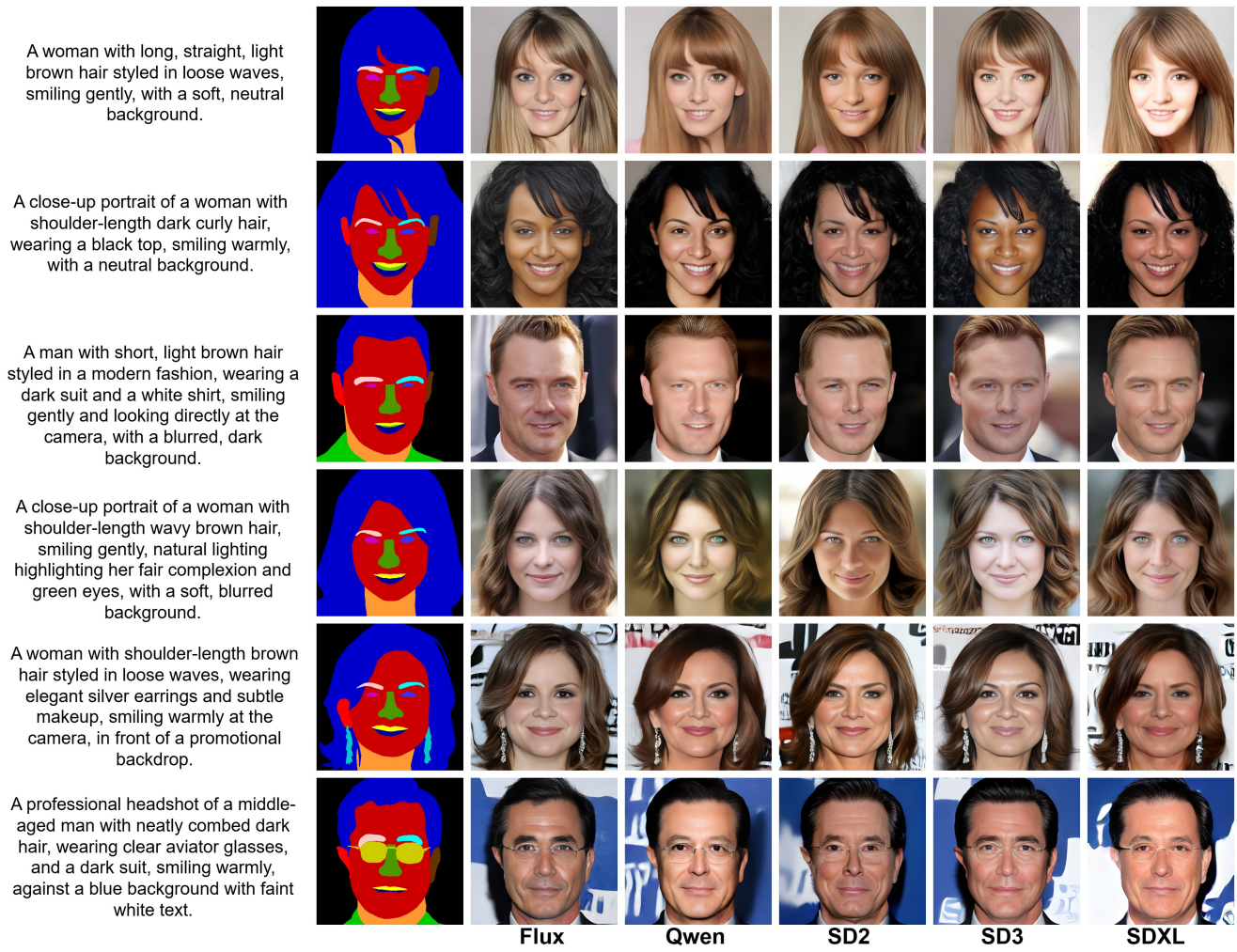

Figure 5. **Qualitative VAE Ablation with Mask-Conditioning.** This figure provides a visual comparison of five different VAE backbones integrated into our MMFace-DiT framework, using segmentation masks as spatial guidance. For each row, every model receives the identical text prompt and mask. The results illustrate a clear trade-off between statistical fidelity and perceptual realism. While models like **SD3** produce sharp outputs, they often introduce an artificial glossiness and oversaturate skin tones. **SD2** and **SDXL** can lead to desaturated or less vibrant results. In contrast, the **Flux** VAE consistently delivers the most balanced and photorealistic portraits, excelling in color accuracy, natural skin texture, and fine-detail preservation. These qualitative findings strongly establish **Flux** as the superior backbone for generating high-fidelity, artifact-free portraits.

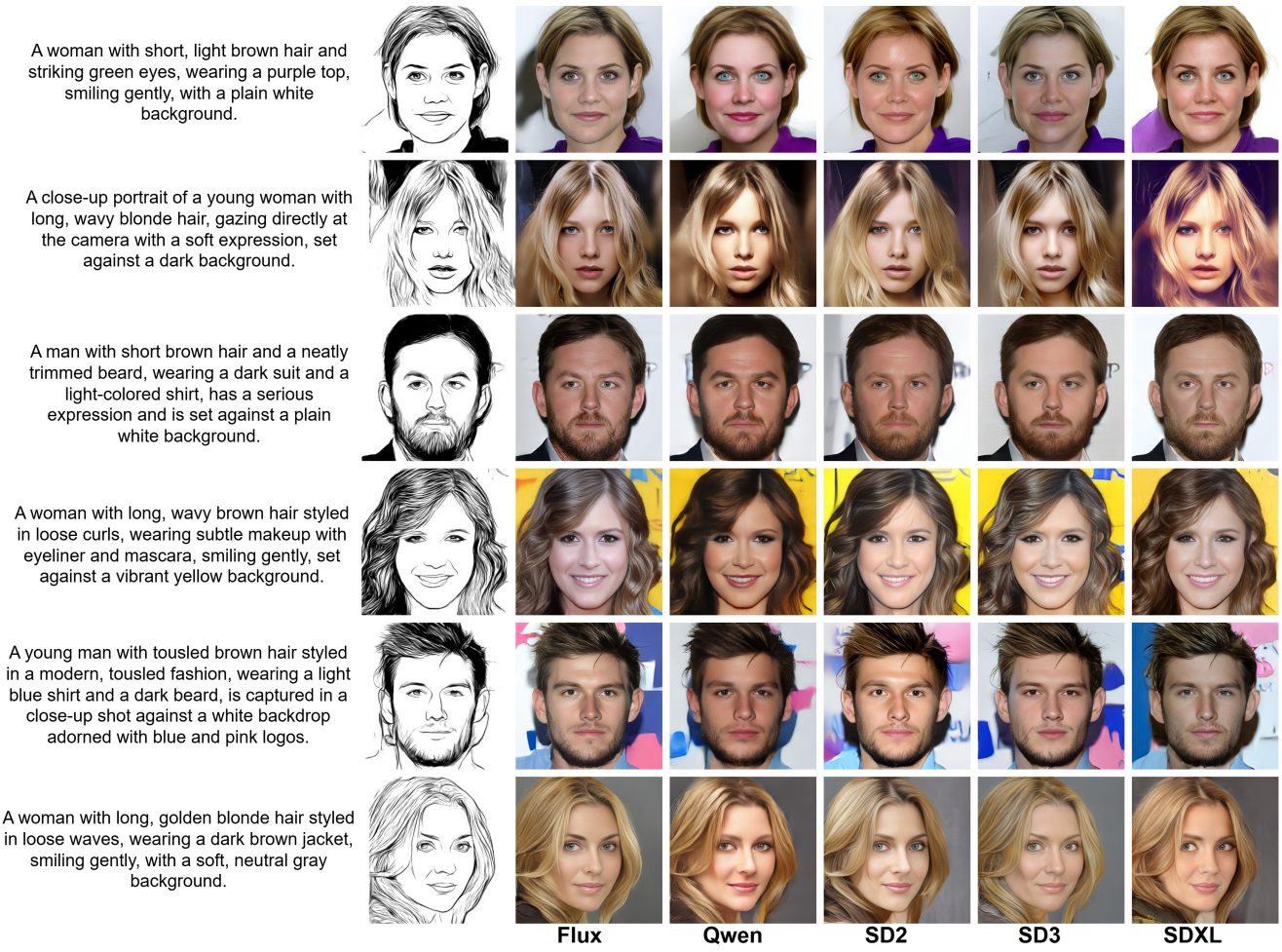

Figure 6. **Qualitative VAE Ablation with Sketch-Conditioning.** This figure showcases the performance of the five VAE backbones when conditioned on artistic sketches. Each model generates an image from the same input sketch and text prompt, testing its ability to preserve geometric structure while synthesizing photorealistic detail. The comparison highlights key differences in semantic adherence and realism. For example, several models fail to accurately render specified features like the *striking green eyes* in the first row. Although **SD3** captures details sharply, its outputs can appear airbrushed. The **Flux** model demonstrates a superior synthesis capability, faithfully translating the sketch's identity and expression while accurately integrating nuanced textual details. This visual evidence aligns with our quantitative results, where **Flux** achieved the best perceptual quality (LPIPS) for sketch-conditioned generation.

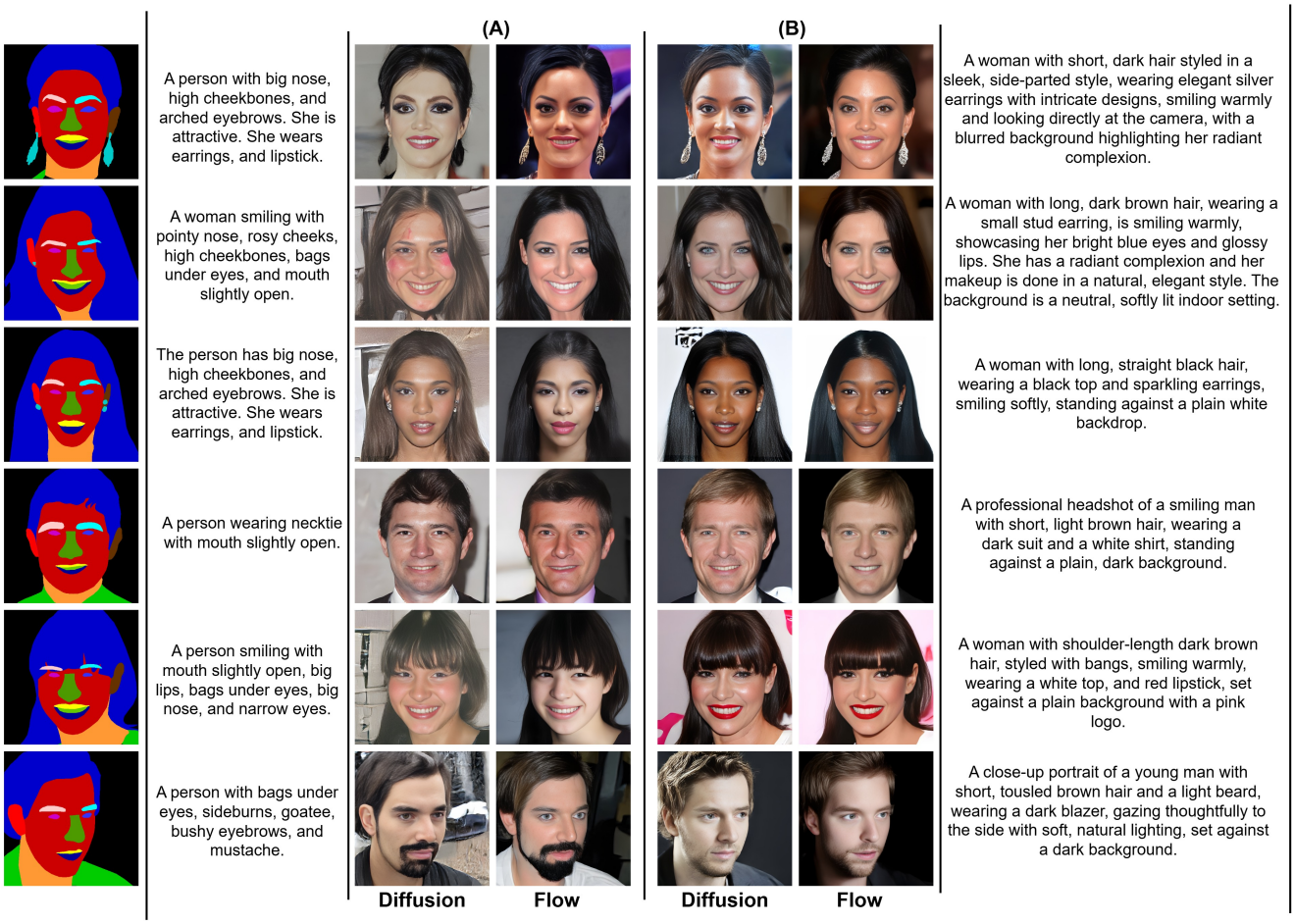

Figure 7. **Efficacy of Rich Textual Conditioning in Mask-Guided Synthesis.** This figure illustrates the substantial qualitative gain achieved through our VLM-powered data enrichment. Column **(A)** relies on the original, sparse annotations, which frequently result in flat lighting or visual artifacts (e.g., the unnatural skin texture in row 2). Conversely, Column **(B)** utilizes our comprehensive descriptions on the exact same segmentation masks. The enriched prompts empower the model to generate intricate accessories and textures that were previously absent, such as *elegant silver earrings* (row 1) or a *dark suit* (row 4). Furthermore, specific stylistic attributes like *red lipstick* (row 5) and environmental context like a *softly lit indoor setting* (row 2) are rendered with high fidelity, demonstrating that detailed semantic guidance is essential for resolving ambiguity in mask-to-image generation.

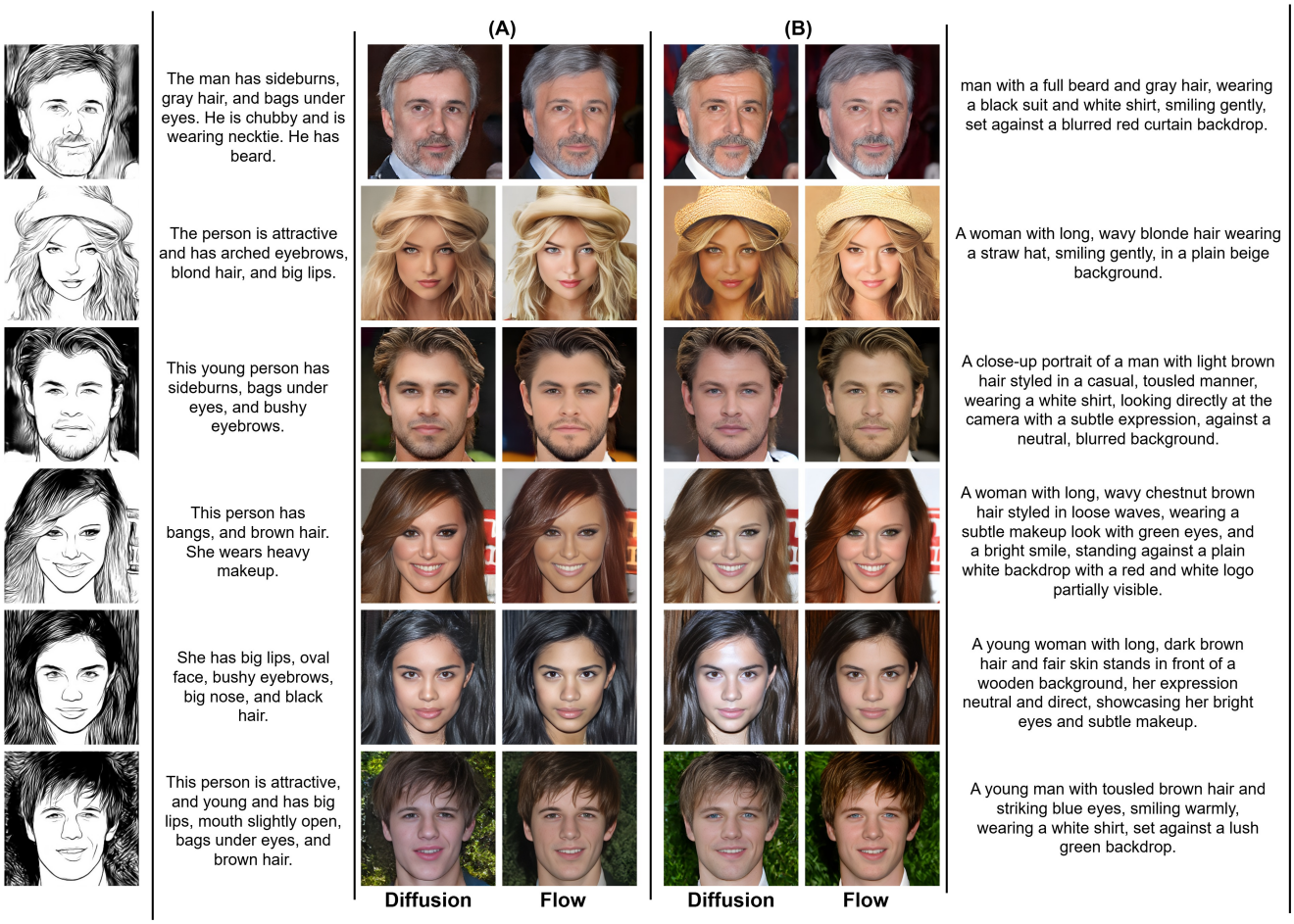

Figure 8. **Efficacy of Rich Textual Conditioning in Sketch-Based Synthesis.** This figure evaluates the impact of our VLM-enriched annotations on sketch-to-image generation. Column **(A)** presents results using the original, brief captions, which typically yield generic attributes and unspecified settings. In contrast, Column **(B)** utilizes our comprehensive descriptions with the exact same input sketches. The enriched prompts enable the model to render precise semantic details—ranging from specific environmental contexts like a *plain beige background* (row 2) or *lush green backdrop* (row 6), to fine-grained facial features such as *striking blue eyes* (row 6) and material textures like a *straw hat* (row 2). Crucially, this semantic enrichment significantly improves photorealism and scene composition while strictly adhering to the structural constraints provided by the input sketch.