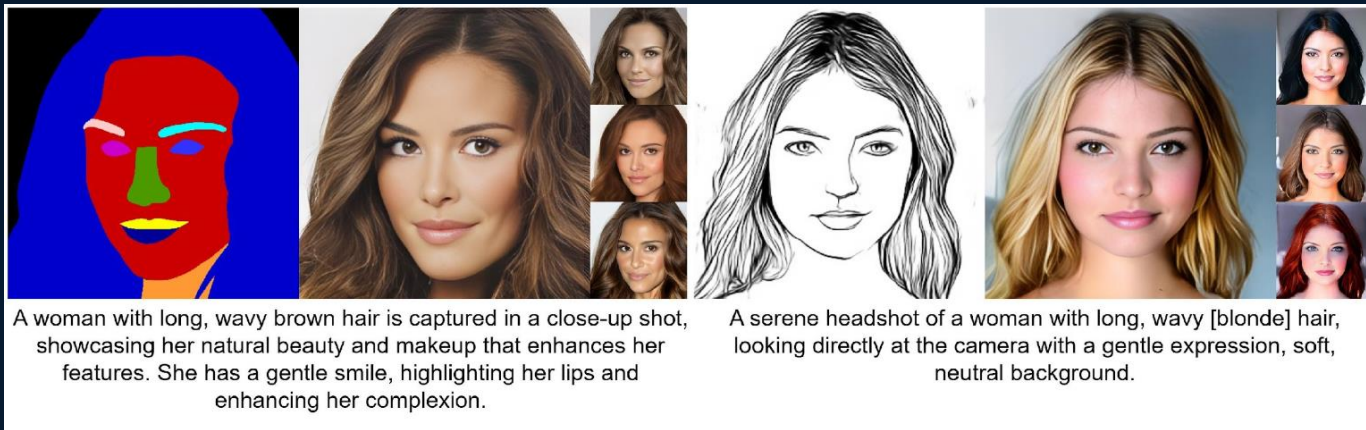


MMFace-DiT

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



IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2026

Bharath Krishnamurthy and Ajita Rattani

The Challenge: Modal Dominance

Current Paradigm:

- ControlNet or stitch-together approaches are the standard.



- Rely on appended ad-hoc control modules.



The Problem:

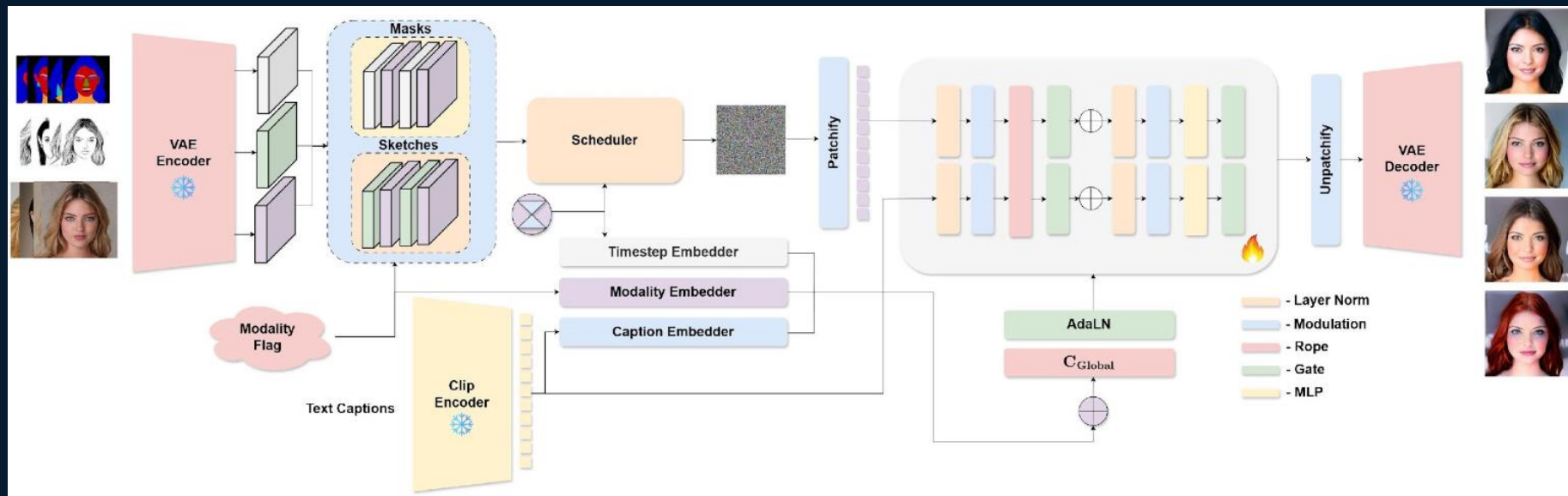
- Modalities conflict rather than synergize.



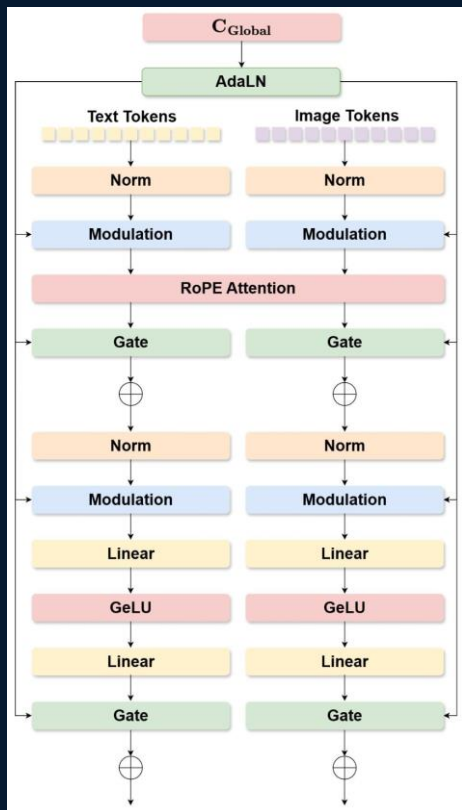
- Strong spatial priors (e.g., dense sketches) overpower subtle text cues.



MMFace-DiT Generation Pipeline



Dual-Stream Architecture



Deep Fusion via RoPE

Attention:

- Image and Text tokens run in parallel
- 2D axial RoPE for image patches
- 1D sequential RoPE for text
- Bidirectional, continuous cross-attention

Unified Conditioning & Adaptation

1. Modality Embedder

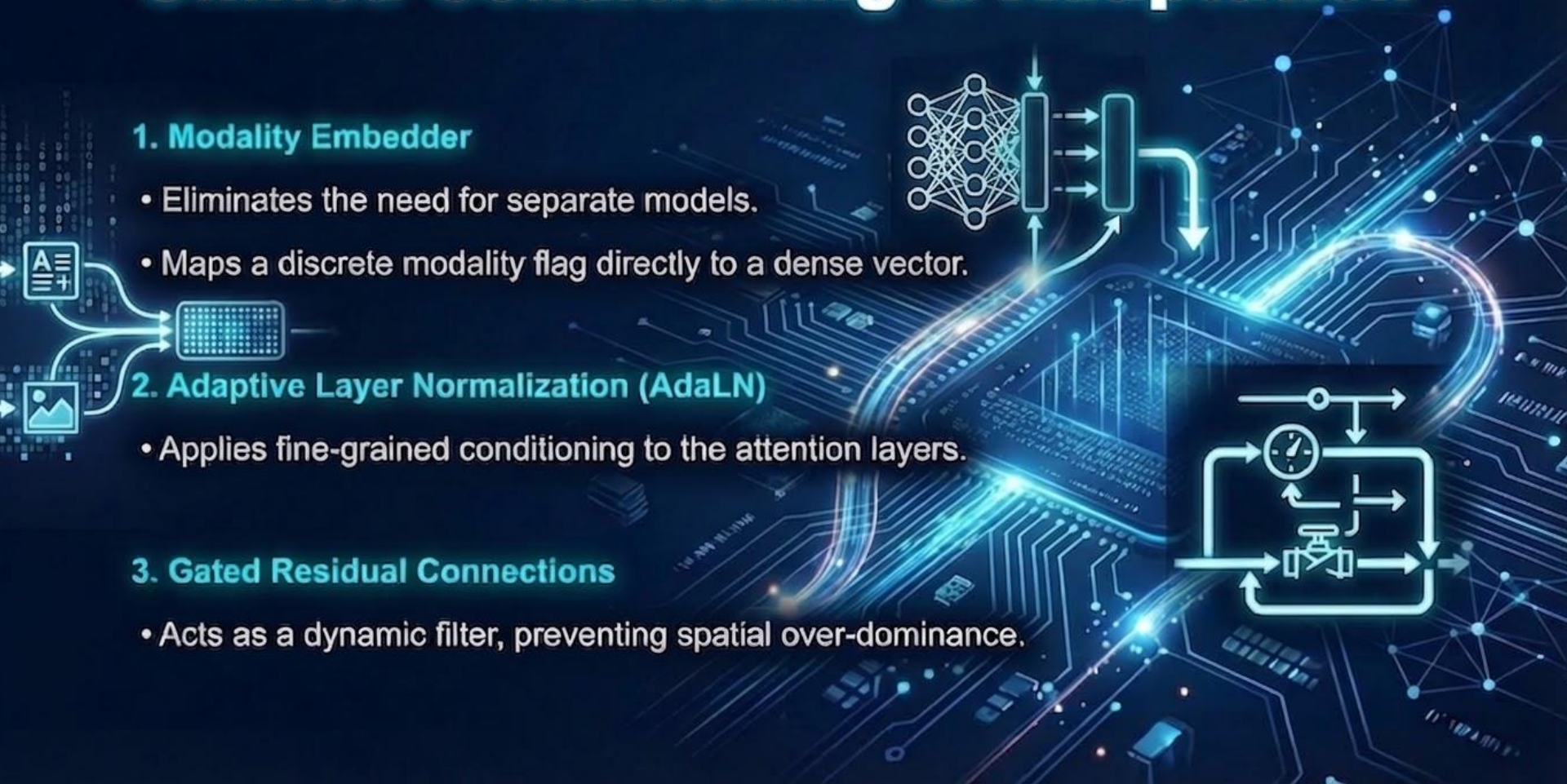
- Eliminates the need for separate models.
- Maps a discrete modality flag directly to a dense vector.

2. Adaptive Layer Normalization (AdaLN)

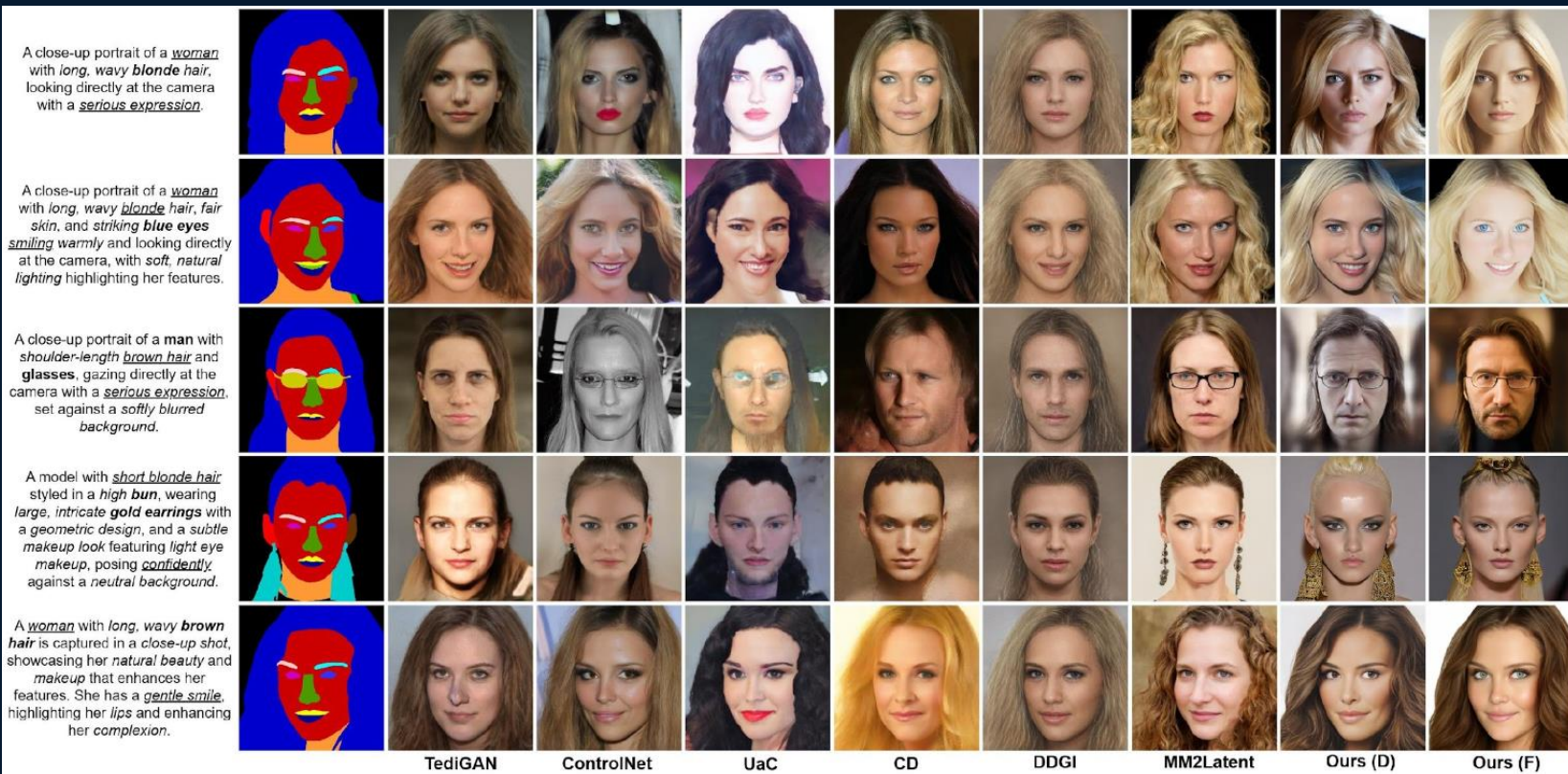
- Applies fine-grained conditioning to the attention layers.

3. Gated Residual Connections

- Acts as a dynamic filter, preventing spatial over-dominance.

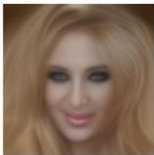


Text + Mask Results



Text + Sketch Results

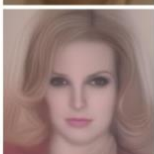
A close-up portrait of a woman with long, dark brown hair and green eyes, wearing subtle makeup and a neutral expression, smiling warmly against a plain white background.



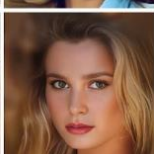
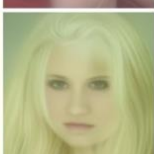
A professional headshot of an older man with short, graying hair, wearing a dark suit and light blue shirt, speaking with a gentle smile, set against a softly lit background.



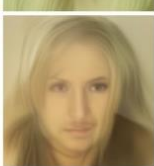
A close-up portrait of a woman with shoulder-length wavy blonde hair, wearing a light blue blouse, gazing directly at the camera with a gentle expression, soft, warm lighting creating a subtle, natural glow.



A close-up portrait of a woman with long, wavy blonde hair and light blue eyes gazing directly at the camera with a subtle expression, soft, natural lighting enhancing her features.



A man with short, graying hair and a serious expression, wearing a dark blue shirt, is captured in a close-up shot against a brick wall background.



TediGAN

ControlNet

UaC

DDGI

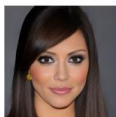
MM2Latent

Ours (D)

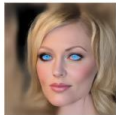
Ours (F)

Quantitative Excellence & Ablations

A woman with long, straight dark brown hair, wearing subtle makeup including eyeliner and mascara, and gold earrings, is smiling softly against a grey background.



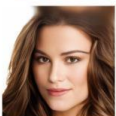
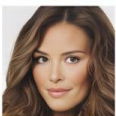
A woman with shoulder-length blonde hair styled in loose waves, wearing subtle makeup, blue eyes, and a soft expression



A young man with short, dark brown hair styled in a casual manner, wearing a light blue dress shirt, smiling gently, against a dark backdrop



A woman with long, wavy brown hair is captured in a close-up shot, showcasing her natural beauty and makeup that enhances her features. She has a gentle smile, highlighting her lips and enhancing her complexion.



A man with short, curly dark hair and bright eyes is smiling warmly, wearing a red collared shirt. He has a clean-shaven face and a gentle expression. The background is plain and slightly blurred.



Flux

Qwen

SD2

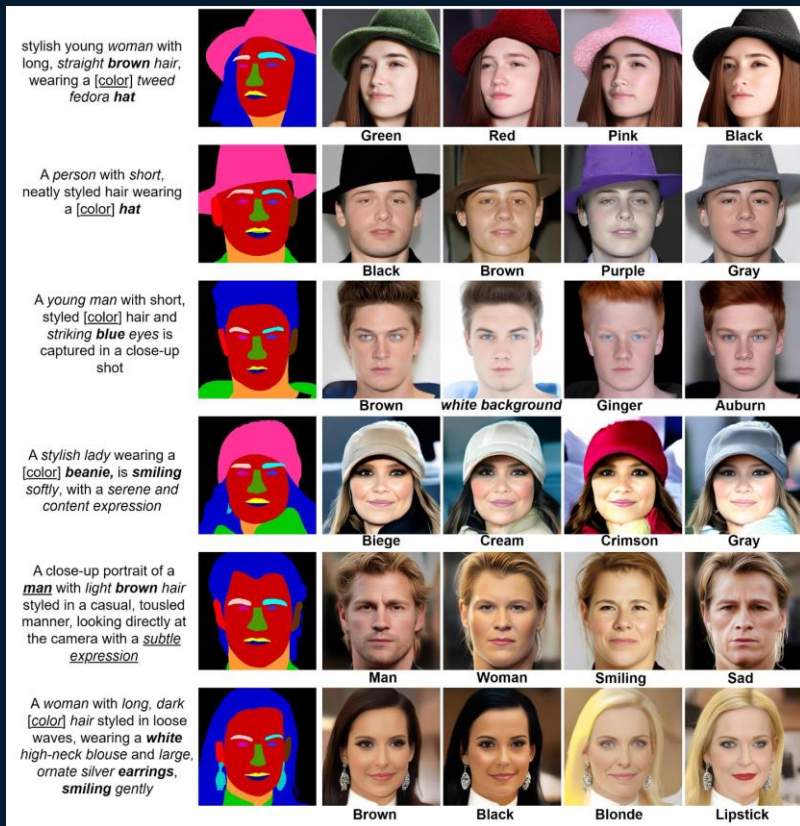
SD3

SDXL

State-of-the-art Performance:

- 40% improvement across all metrics
- Best prompt adherence (LLM Score)
- Superior FID (16.63) and mIoU
- Ablations confirm FLUX VAE value

Conclusion



Key Takeaways:

- Solves Modal Dominance elegantly
- 1M new high-quality captions
- Single unified architecture
- Thank you!