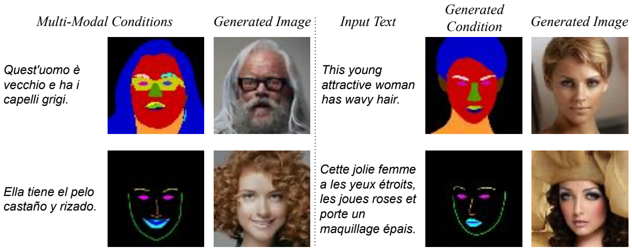
M^{3}Face: A Unified Multi-Modal Multilingual Framework for Human Face Generation and Editing

Mohammadreza Mofayezi[1] Reza Alipour[1] Mohammad Ali Kakavand[1] Ehsaneddin Asgari[2]

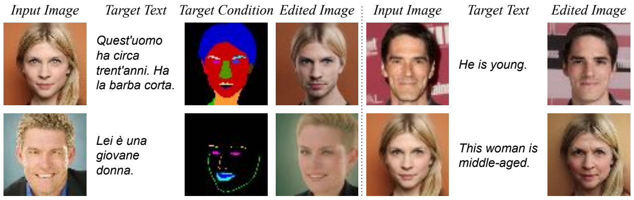
[1] Sharif University of Technology [2] Qatar Computing Research Institute

Abstract

*Human face generation and editing represent an essential task in computer vision and the digital world. Recent studies have shown remarkable progress in multi-modal face generation and editing, for instance, using face segmentation to guide the generation process. However, it may be challenging for some users to create these conditioning modalities manually. Thus, we introduce **M^{3}Face**, a unified multi-modal multilingual framework for controllable face generation and editing. This framework enables users to generate or edit controlling modalities automatically using only text input. Specifically, it can generate semantic segmentation or facial landmarks, which are then used to create face images. We conduct extensive qualitative and quantitative experiments to showcase our framework's face generation and editing capabilities. Additionally, we propose the **M^{3}CelebA Dataset**, a large-scale multi-modal multilingual face dataset containing high-quality images, semantic segmentations, facial landmarks, and different captions for each image in multiple languages.*


Multi-Modal Conditions
Generated Image
Input Text
Generated Condition
Generated Image
Quest'uomo è vecchio e ha i capelli grigi.
This young attractive woman has wavy hair.
Ella tiene el pelo castaño y rizado.
Cette jolie femme a les yeux étroits, les joues roses et porte un maquillage épais.

(a) Face Generation



(b) Face Editing

Figure 1. We introduce **M^{3}Face** for controllable multi-modal multilingual face generation and editing. **(a) Face Generation** can be done with both multi-modal conditions or a single text input. We generate face images consistent with the input text and other given modalities. **(b) Face Editing** can also be done with text, mask, landmarks, or a combination of them.

1. Introduction

The field of image generation and editing has witnessed remarkable progress, propelled by advancements in generative models [6, 15, 23, 34] and large-scale datasets [20, 21, 34]. These advancements are supposed to substantially contribute to the forthcoming generation of animations, movies, and video games [10, 17]. The creation of human characters, particularly characterized by their facial features, is one of the fundamental elements in the field of Generative AI. The innovations in this domain have given rise to sophisticated techniques capable of creating and modifying human facial images with unprecedented realism and versatility. Beyond their synthesis capability, these methods exhibit proficiency in handling diverse modalities, such as text [4, 13, 16, 22, 27–29], semantic segmentation masks [28, 32, 38], facial landmarks [31, 37, 38], or a combination of them [12, 34].

Using multiple modalities to guide the generation of face images is a powerful approach. This allows users to precisely control and customize various facial features in the generated images. However, it poses some challenges; for instance, creating specific modalities manually, such as semantic segmentation, can be a complex task for most users. The intricate process of generating these conditions from scratch can be a potential barrier to the ease of use of multimodal approaches.

In addition to the advancements in multi-modal generation and editing, the exploration of multilingual features has become more prominent [3, 30, 35]. Acknowledging the global diversity of languages and cultures, researchers are now incorporating linguistic elements into the generation and editing processes. This multilingual perspective transcends language barriers and fosters a more inclusive and accessible approach to face image generation and editing.

In this paper, we introduce **M^{3}Face**, a unified multi-modal multilingual framework for controllable face generation and editing. While previous methods do well on face generation and editing with different modalities, users may not have the initial conditions in real-world scenarios. Our framework addresses these issues in two significant ways. Firstly, it simplifies the generation of necessary conditions for image generation, such as semantic segmentation, by a text prompt. The user can continue generating or editing these conditions interactively. The framework then incorporates this text input and generated conditions to generate face images utilizing our specialized face ControlNet [38] model, or edit face images with our adapted editing method based on Imagic [16], making it easier for users to control and generate the desired images. Furthermore, we propose the **M^{3}CelebA Dataset** built upon the original CelebA [21] dataset. Our dataset contains over 150,000 images of faces (5× larger than the Multi-Modal CelebA-HQ [34] dataset) paired with high-quality multilingual captions. Secondly, *M^{3}Face* is trained using multiple languages existing in *M^{3}CelebA Dataset*, which makes it accessible to users all over the world. To summarize, our contributions are as follows:

- We introduce **M^{3}Face**, a unified multi-modal multilingual framework for controllable face generation and editing. Unlike existing multi-modal approaches, our method generates necessary conditions directly from text prompts, offering an option to use additional conditioning modalities without requiring them. Furthermore, this approach disentangles content and structure, enabling more precise and consistent results. It also supports multiple languages, broadening its applicability.
- We propose the **M^{3}CelebA Dataset**, a large-scale multi-modal and multilingual face dataset containing high-quality images, semantic segmentations, facial landmarks, and different multilingual captions for each image.
- We achieve state-of-the-art qualitative and quantitative results in face generation and editing using facial landmarks and semantic segmentation.

2. M^{3}Face

We introduce **M^{3}Face**, a unified framework for controllable multilingual human face generation and editing. Within this framework, we initially generate additional modalities from the input text and then utilize them for controllable face generation and editing. Fig. 2 shows an overview of the framework.

2.1. Face Generation Pipeline

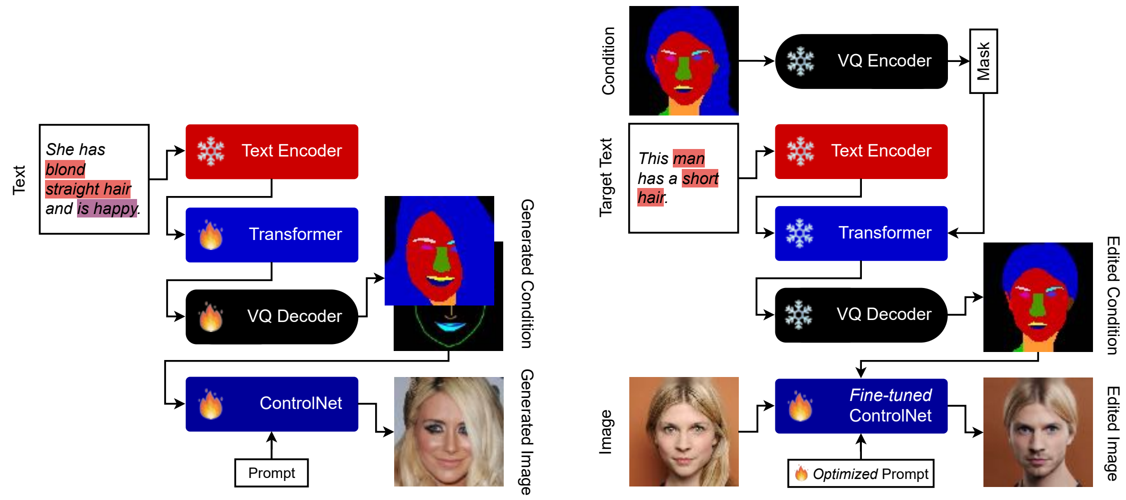
For face generation, we begin with generating facial landmarks or semantic segmentation from the given text input using a masked transformer model, inspired from [4, 24]. The masked transformer uses a U-VIT [11] architecture. It generates segmentation mask and landmark image tokens based on the given text input. To decode these tokens, we utilize a VQ-GAN [7] with an encoder and a decoder, which maps images into a sequence of tokens using a learned codebook. We fine-tune this VQ-GAN specifically for tokenizing segmentation masks and landmarks to ensure high accuracy in the generated intermediate representations. To support multiple languages, our pipeline integrates the M-CLIP (LaBSE) [3, 8] text encoder. We chose M-CLIP over T5 [26] since the pre-trained model used for fine-tuning utilized the CLIP-L14 [25] text encoder, allowing us to maintain consistency and leverage the same pre-trained weights. After fine-tuning the VQ-GAN and changing the text encoder, we fine-tune the transformer model to generate masks and landmarks.

Once a segmentation mask or landmark is generated, we utilize a face ControlNet [38] to generate the final face images from the intermediate results. Specifically, we train two distinct ControlNet models: one for generating images based on landmark conditions and another for generating images based on mask conditions.

2.2. Face Editing Pipeline

Face editing can be done with text, masks, landmarks, or a combination of them. We first edit the segmentation mask or landmarks using the inpainting technique with the mask/landmark transformer. Unlike general image editing, where details might be lost with inpainting or the identity of an image might be changed, masks and landmarks serve as fundamental structural elements of a face. Using inpainting to edit these elements will not have these issues because these images are not very highly detailed, and a portion of these structures usually can provide sufficient information for reconstructing the original structure.

Then, we utilize the Imagic [16] method, which we have extended to support image editing with multi-modal inputs, to edit face images with our face ControlNet models. This involves optimizing the embedding of the input prompt and subsequently fine-tuning the UNet in the base Stable Diffusion component of our ControlNet model. We then directly input the edited condition into the ControlNet model. Unlike Collaborative Diffusion [12], which requires multiple optimization steps to apply this method for different modalities, we perform the edit in a single step. This approach makes multi-modal (or uni-modal) editing faster and results in a better outcome.



She has blond straight hair and is happy.

This man has a short hair.

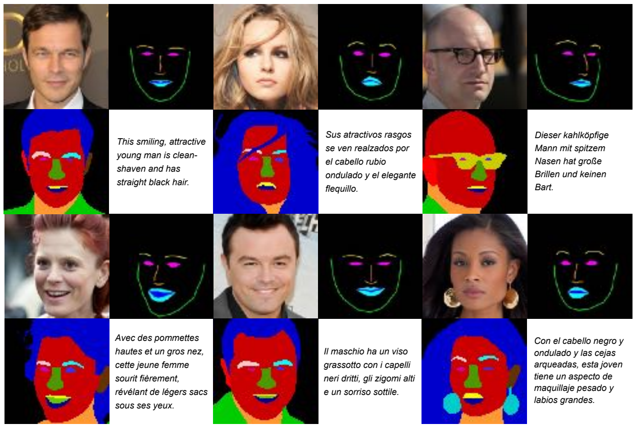
(a) Face Generation **(b) Face Editing**

Figure 2. **M^{3}Face Framework.** For **(a) Face Generation**, we first generate facial landmarks or semantic segmentation masks from a given text input with our mask/landmark generator [4]. We then utilize our ControlNet [38] model to generate face images from the intermediate results. For **(b) Face Editing**, we first edit the given condition using the inpainting technique with the mask/landmark generator. Next, we optimize the embedding of the text input and subsequently fine-tune the ControlNet model following the Imagic [16] method. Finally, we directly input the edited condition and optimized prompt to the fine-tuned ConrolNet. Face editing can be done with text, mask, landmarks, or a combination of them.

3. M^{3}CelebA Dataset

We propose the **M^{3}CelebA Dataset**, a large-scale multi-modal multilingual face dataset based on the original CelebA [21] dataset. Our dataset contains more than 150K high-quality face images with semantic segmentation, facial landmarks, and multilingual captions.

In Fig. 13, we present an overview of the dataset generation pipeline. Initially, we align and crop the original CelebA images using the facial landmark annotations provided in the dataset. These images are then upscaled to 512 × 512 resolution using the Real-ESRGAN [33] model. For facial landmarks, we utilize the `dlib` [18] library, and for semantic segmentation, we use the DML-CSR [39] model. For generating the captions, we use the 40 annotated attributes in the original CelebA dataset. We create three captions for each image utilizing a fine-tuned GPT3.5 [2] model, followed by translation with the SeamlessM4T [1] model. Finally, we filter our dataset by both human and automatic evaluation and remove images with occlusions, blurriness, or extreme poses. Some samples from the dataset are shown in Fig. 3. More details about the dataset are available in Sec. H.



This smiling, attractive young man is clean-shaven and has straight black hair.

Sus atractivos rasgos se ven realzados por el cabello rubio ondulado y el elegante flequillo.

Dieser kahlköpfige Mann mit spitzem Nasen hat große Brillen und keinen Bart.

Avec des pommettes hautes et un gros nez, cette jeune femme sourit fièrement, révélant de légers sacs sous ses yeux.

Il maschio ha un viso grassotto con i capelli neri dritti, gli zigomi alti e un sorriso sottile.

Con el cabello negro y ondulado y las cejas arqueadas, esta joven tiene un aspecto de maquillaje pesado y labios grandes.

Figure 3. **M^{3}CelebA Dataset.** 512 × 512 images from the M^{3}CelebA dataset as well as the generated facial landmarks and semantic segmentation. Three multilingual captions are available for each image.

4. Experiments

Qualitative Comparison. In this section, we show our face generation and editing results with both multi-modal and text-only conditions. In Fig. 4, you can see the text-guided face generation and editing capabilities of our method. A semantic segmentation or facial landmarks condition will be generated for each input text prompt. Then, the face image can be generated with multi-modal conditions. We can see that both the generated conditions and the face images are consistent with the input prompt, and the model can generate or edit details such as emotions and

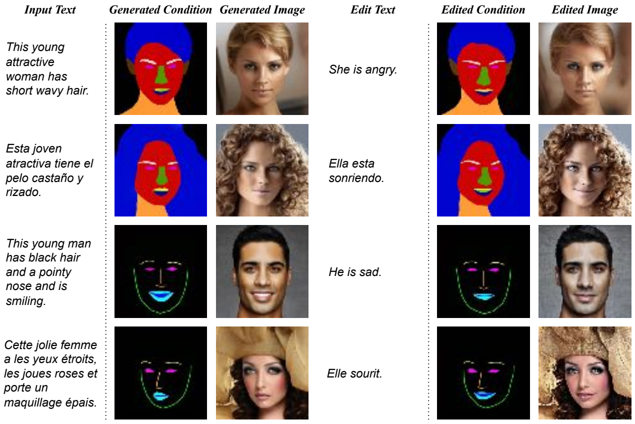

Input Text
Generated Condition
Generated Image
Edit Text
Edited Condition
Edited Image
This young attractive woman has short wavy hair.
She is angry.
Esta joven atractiva tiene el pelo castaño y rizado.
Ella esta sonriendo.
This young man has black hair and a pointy nose and is smiling.
He is sad.
Cette jolie femme a les yeux étroits, les joues roses et porte un maquillage épais.
Elle sourit.

Figure 4. **Text-Guided Face Generation and Editing Results.** We can generate and edit face images with a single text prompt. We first generate semantic segmentation or facial landmarks according to the given text prompt. Then, we generate face images from these intermediate conditions.

Method	FID ↓	Text ↑	Mask ↑	Human ↑
TediGAN	58.49	-[1]	0.90	19.02
CollabDiff	38.20	24.80	0.90	23.64
Ours	**30.16**	**27.86**	**0.93**	**57.34**
SD-v1.5	67.29	26.82	-	8.22
GCDP	53.45	27.15	-	10.32
Ours (Mask)	39.14	27.28	-	23.93
Ours (Landmark)	**36.31**	**27.65**	-	**57.53**

Table 1. **Quantitative Results of Face Generation.** Our method generates images with higher quality and more consistency with multi-modal conditions compared to TediGAN and Collaborative Diffusion. In generating faces from text input, our results surpass Stable Diffusion and GCDP. Ours (X) means that given the text input, we first generate the condition X, and then generate a face image from text and X.

face details. For face generation, we compare our results with multi-modal methods [12, 34] in Fig. 9. TediGAN [34] often fails to generate face images consistent with the given mask condition due to its two-step generation method. Collaborative Diffusion [12] also cannot generate details such as earrings, make-up, or specific beards like a goatee.

For face editing, we compare our results with both text-guided [13, 22] and multi-modal [12] approaches. Fig. 10 shows the results. Collaborative Diffusion fails to follow the target mask and text prompt due to its two-step editing method. In Fig. 10 (b), we can see the text-guided editing results. While Null-Text Inversion [22] preserves the identity of the face very well, it struggles with editing specific attributes like beard and is sensitive to the target prompt. Talk-to-Edit [13] has a problem with most in-the-wild images and cannot make edits to them. In successful attempts, it does not preserve the identity very well. Beyond these, we tried to compare with face editing methods using facial landmarks. FReeNet [37] seems to be very overfitted to specific poses and images, and the generated faces are either not changed or not reconstructed at all.

Method	Mask/Landmark ↑	Text ↑	Human ↑
CollabDiff	0.91	0.04	19.80
Ours (Mask)	**0.94**	**0.09**	**80.20**
Ours (Landmark)	0.87	0.10	-

Table 2. **Quantitative Results of Face Editing.** Our face editing outperforms Collaborative Diffusion in terms of segmentation consistency and directional CLIP similarity.

Quantitative Comparison. We compare our face generation results with both multi-modal (top rows) and text-guided (bottom rows) approaches in Tab. 1. We outperform these methods in all quantitative metrics, showcasing the capabilities of our method. We also compare multi-modal face editing results with Collaborative Diffusion in Tab. 2.

5. Conclusion

In this work, we propose a novel framework for multi-modal multilingual human face generation and editing called **M^{3}Face**. With this framework, users can generate face images and necessary conditions by a text prompt, offering the option to use additional conditioning modalities but not requiring them, and at the same time disentangle content and structure. We also introduce the **M^{3}CelebA Dataset** which contains high-quality face images, semantic segmentations, facial landmarks, and multilingual captions for each image. We conduct extensive experiments to demonstrate our framework's capabilities in face generation and editing with different modalities.

Our experiments demonstrate M^{3}Face's effectiveness. However, we observed a limitation with the mask/landmark generator, which may produce invalid colors. This challenges color consistency and accuracy, prompting the exploration of alternative architectures in future research. Another limitation of our framework is that the quality of generated images is heavily impacted by the Stable Diffusion backbone in the ControlNet model. Recent studies have introduced more robust SD models such as the Smooth Diffusion [9], which might improve face generation results and better preserve unedited content in face editing. These investigations are left for future studies.

---

[1] TediGAN fine-tunes its model with the CLIP loss, and it eventually overfits on the input prompt. Thus, calculating the CLIP Score for this method is misleading.

References

[1] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023. 3

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[3] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, 2022. 1, 2

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 1, 2, 3

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics, 2020. 2

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 2

[8] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding, 2022. 2

[9] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. *arXiv preprint arXiv:2312.04410*, 2023. 4

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[11] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 2

[12] Ziqi Huang, Kelvin C.K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 4

[13] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021. 1, 4, 2, 6

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 6

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[16] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 1, 2, 3

[17] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 1

[18] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 3

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[20] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 6

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 3, 6

[22] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 1, 4, 2

[23] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7591–7600, 2023. 1, 2

[24] Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. amused: An open muse reproduction, 2024. 2, 1

[25] 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*, pages 8748–8763, 2021. 2

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. 2

[27] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative ad-

versarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 1

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[30] Michael Saxon and William Yang Wang. Multilingual conceptual coverage in text-to-image models. *arXiv preprint arXiv:2306.01735*, 2023. 1

[31] Hao Tang and Nicu Sebe. Total generate: Cycle in cycle generative adversarial networks for generating human faces, hands, bodies, and natural scenes. *IEEE Transactions on Multimedia*, 24:2963–2974, 2021. 1

[32] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 1

[33] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021. 3

[34] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 1, 2, 4, 6

[35] Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. Altdiffusion: A multilingual text-to-image diffusion model, 2023. 1, 2

[36] Dongxu Yue, Qin Guo, Munan Ning, Jiaxi Cui, Yuesheng Zhu, and Li Yuan. Chatface: Chat-guided real face editing via diffusion latent space manipulation. *arXiv preprint arXiv:2305.14742*, 2023. 1, 2

[37] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5326–5335, 2020. 1, 4

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2, 3

[39] Qi Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *Computer Vision and Pattern Recognition*, 2022. 3

M^3Face: A Unified Multi-Modal Multilingual Framework for Human Face Generation and Editing

Supplementary Material

We provide supplementary material for our paper including more results and a detailed explanation of our dataset in this document. We also release our code and dataset at https://huggingface.co/m3face.

A. Related Work

Face Generation. Various image generation approaches have been explored, employing diverse architectures and algorithms to enhance the quality and controllability of image generation. GAN-based methods can generate images with text by leveraging their latent space. TediGAN [34] added controllability to the generation by utilizing different modalities, which led to more stable results. Stable Diffusion [28] achieved improved qualitative results with an easier learning process in image generation. Similar success has also been found in various related tasks, including face generation. Collaborative Diffusion [12] uses pre-trained uni-modal diffusion models to utilize semantic segmentation maps as guidance in image generation. GCDP [23] uses a Gaussian-categorical diffusion process to learn the joint image-semantic distribution and generate image-semantic pairs with text.

Face Editing. Recent works in face image editing have aimed to improve the quality and accuracy of this complex task. Talk-to-Edit [13] employs a location-specific semantic field within the GAN latent space to determine the edit distance. Similarly, ChatFace [36] learns high-level semantic information about images and uses a diffusion model to facilitate editing within this space. Null-Text Inversion [22] is based on Stable Diffusion and uses Null-Text optimization for editing real images. Some approaches integrate semantic segmentation as a guide for editing face images. Collaborative Diffusion [12] uses semantic segmentation in image editing alongside text input in two diffusion models and then aggregates the results. Furthermore, methods that utilize facial landmarks to address pose and spatial properties have also been explored. FReeNet [37] uses an encoder-decoder architecture to edit input images based on target landmarks, while C^2GAN [31] uses two GAN models to incorporate facial landmarks into the editing process. While these methods benefit from different modalities to achieve stability and quality, they often present challenges to user experience due to the complex process of creating inputs like segmentation maps. In contrast, our method simplifies this process by automatically generating these modalities from input text and then utilizing them in the subsequent stages of the pipeline, thereby enhancing both efficiency and user-friendliness.

B. Experimental Setup

Data Preparation. We use a combination of the CelebA-HQ and M^3CelebA datasets for training. To improve consistency in face segmentation, we adjust the segmentation colors to be symmetric; for example, both eyebrows are assigned the same color. For training the mask/landmark transformer, we use random flip for data augmentation. Additionally, when training ControlNet, we use multiple data augmentation techniques, including random flips, changing color, brightness, and contrast of condition images. To account for potential invalid color generation by the mask/landmark generator, we also introduce random color changes in the condition images. This ensures that the ControlNet model remains robust and effective within our pipeline. We evaluate all methods on the test set of our dataset.

Training Details. For training the mask/landmark generator, we use an open-source implementation of Muse called aMUSEd [24] and its pre-trained models. First, we fine-tune the VQ-GAN on three types of images: face segmentation, facial landmarks, and face portrait. Then, we fine-tune the transformer model for face segmentation, facial landmarks, and face portrait generation. For this purpose, we add instructions for each task to the image captions. After completing 15K training steps on these tasks, we perform fine-tuning for 5K additional steps on two individual models: one for face segmentation generation and the other for facial landmarks generation. During the training of the transformer model, the percentage of masked latent tokens is sampled from a cosine masking schedule. For training ControlNet, we use a variant of Stable Diffusion that supports multiple languages [35]. We train the ControlNet model for ten epochs with a total batch size of 4 and gradient accumulation of 16, using a learning rate of 5×10^{-5}. We train two individual models, one by conditioning on face segmentation and another by conditioning on facial landmarks. To make the model more accurate on using condition images, 25% of the training samples are provided without a caption and only with a simple prompt to ”generate a high-quality portrait of a face”.

Face Editing. For face editing with our multi-modal Imagic method, we use Adam [19] optimizer with a learning rate of 10^{-3} for optimizing the text embedding and 5×10^{-5} for fine-tuning the UNet parameters. We optimize the embedding for 500 and fine-tune the model for 1000 steps.

B.1. Baselines

We compare our proposed framework with state-of-the-art approaches for text-guided and multi-modal face generation and editing. **Collaborative Diffusion** [12] is an LDM-based multi-modal face generation and editing method. They introduce a dynamic diffuser to combine different uni-modal pre-trained diffusion models. Similar to us, they utilize the Imagic method for multi-modal face editing. **TediGAN** [34] is a StyleGAN-based face generation method. They utilize StyleGAN's latent space to project the embedding of different modalities into it. The performs style mixing to achieve multi-modal face generation. **GCDP** [23] is an Imagen-based text-guided image-mask pair generation method. They propose a Gaussian-categorical diffusion process that simultaneously generates images and corresponding semantic segmentations. **Stable Diffusion** [28] is a latent text-to-image diffusion model. We compare face generation results with Stable-Diffusion-v1-5, which was initialized with the weights of the Stable-Diffusion-v1-2 checkpoint and subsequently fine-tuned with more steps. We chose this version instead of v2-1 because it performs better on face images. **Talk-to-Edit** [13] performs interactive face editing via dialog. They model a location-specific semantic field. However, it is limited to several predefined attributes. Similar to this work, ChatFace [36] conducts text-driven face editing in the semantic latent space of a diffusion model. This work has no public implementation; thus we do not include them in the comparisons. **Null-Text Inversion** [22] is a Stable Diffusion-based method for editing real images by using Null-text inversion, i.e., modifying the unconditional text embedding rather than the input text embedding.

B.2. Evaluation Metrics

FID. The quality of generated images is evaluated through Frechet Inception Distance, which measures the feature representation's distance between generated images and real images. Lower FID implies better sample quality.

CLIP Score. The CLIP Score is the cosine similarity between the normalized image and text embeddings. A higher score usually indicates higher consistency between the generated image and the text prompt. We utilized the MCLIP-XLM model [3, 5] to assess the scores, as we needed to compare models with captions in diverse languages.

Directional CLIP Similarity. It measures the consistency of the change between the two images (in CLIP [25] space) with the change between the two image captions. This metric is used for evaluating image editing, and the higher the directional CLIP similarity, the better it is.

Segmentation and Landmark Consistency. Segmentation Consistency is the pixel-wise accuracy against the ground-truth segmentation. Landmark Consistency is the distance between ground-truth landmarks and the generated face landmarks.

Human Evaluation. We conducted a user study to assess our framework's performance. For face generation, 200 multi-modal conditions were randomly selected from the M^{3}CelebA dataset's test split, and images were generated based on these conditions. Evaluators chose the best image based on photo realism and consistency with the conditions. For face editing, evaluators assessed identity preservation and the same criteria as in the face generation study using input and edited images along with edited conditions.

C. Ablation Study

Multilingual Diffusion. We study the effect of the diffusion backbone in the ControlNet model and training on multiple languages as discussed in Sec. B. We compare using AltDiffusion [35] as a multilingual backbone and the vanilla Stable Diffusion in Tab. 3. It shows that while the multilingual model has a broader use case, the overall performance may decrease. Additionally, it demonstrates that our superiority over methods using the vanilla Stable Diffusion does not come from the backbone model. We also report quantitative results of face generation with multilingual captions in Tab. 4. Our method demonstrates consistent performance across different languages, highlighting its ability to understand multilingual prompts.

Backbone	FID ↓	Text ↑	Mask ↑
AltDiffusion	30.16	27.86	0.93
SD-v1.5	28.08	27.35	0.93

Table 3. **Effect of the Diffusion Backbone.** While AltDiffusion brings multilingual knowledge, it does not improve the overall performance of our pipeline.

Method	FID ↓	Text ↑	Mask ↑
English	30.16	27.86	0.93
Spanish	30.51	27.31	0.93
French	33.66	27.26	0.93
Italian	32.93	27.35	0.93
German	31.77	27.15	0.93

Table 4. **Quantitative Results of Multilingual Face Generation.** Our method can generate high-quality face images with captions in different languages.

In Tab. 5, we show the effect of augmenting condition images with color perturbations and replacing a fraction of training captions with a simple prompt. Both make face generation more robust and accurate using the condition images.

Method	FID ↓	Text ↑	Mask ↑
M^{3}Face	30.16	27.86	0.93
w.o. Augmentation	31.25	27.62	0.92
w.o. Caption	31.37	27.40	0.91

Table 5. **Ablation on Training.** Using color change on condition images as augmentation and replacing the caption with a simpler one improves the quality and accuracy on condition images.

Face Editing. We study the effect of the α parameter used in the Imagic method. Given an image and a target text prompt, we encode the target text and get the initial text embedding e_{tgt}, then optimize it to reconstruct the input image, obtaining e_{opt}. We then fine-tune the UNet model to improve fidelity to the input image while fixing e_{opt}. Finally, we interpolate between e_{opt} and e_{tgt} with an α coefficient to generate the edit result. In Fig. 5, the effect of this α coefficient and the number of UNet layers for fine-tuning is shown. The strength of the edit can be controlled with α. Fine-tuning all of the UNet UpBlock parameters provides more consistent and controllable edits despite requiring more GPU memory compared to fine-tuning on fewer layers.

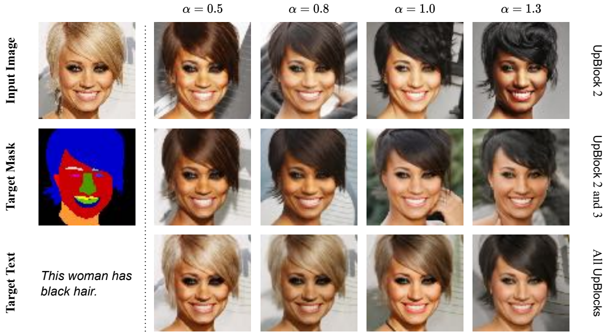


Figure 5. **Effect of Alpha and Number of Layers.** We can control the strength of textual editing with the alpha parameter. With more alpha values, we achieve higher face manipulation. While fine-tuning on fewer layers requires less GPU memory, we can have more consistent and controllable edits by fine-tuning all of the UNet UpBlock parameters.

D. More Face Generation and Editing Results

We provide more face genration and editing results in this section. In Fig. 6, we provide multi-modal face generation results with semantic segmentation and facial landmarks conditions. We can generate diverse face images consistent with the given modalities as shown in Fig. 7. In Fig. 8, we show comprehensive multi-modal face editing results. We see that the identity of faces is well-preserved and the generated faces are consistent with both the text and other conditions.

E. Qualitative Comparison Results

We show qualtitative comparisons with baselines for face generation and editing in Fig. 9 and Fig. 10.

F. More Mask/Landmark Generation Results

We present more results for mask/landmark generation given text inputs in Fig. 11. We can see that the generated images are diverse and coherent with the given text description.

G. Zero-Shot Face Generation

In this section, we show the capabilities of the M^{3}Face pipeline in zero-shot face generation. We test captions in languages that are not in the training data, as well as specific human names. The results are shown in Fig. 12.

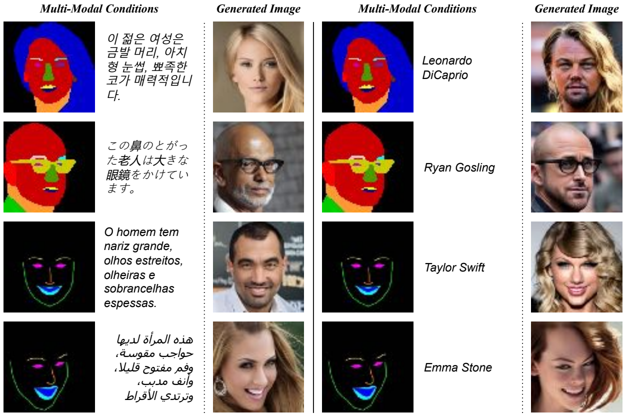

Multi-Modal Conditions
Generated Image
Multi-Modal Conditions
Generated Image
이 젊은 여성은 금발 머리, 아치형 눈썹, 뾰족한 코가 매력적입니다.
この鼻のとがった老人は大きな眼鏡をかけています。
O homem tem nariz grande, olhos estreitos, olheiras e sobrancelhas espessas.
هذه المرأة لديها حواجب مقوسة، وفم مفتوح قليلا، وأنف مدبب، وترتدي الأقراط
Leonardo DiCaprio
Ryan Gosling
Taylor Swift
Emma Stone

Figure 12. **Zero-Shot Generation Results.** Our pipeline can make zero-shot face generations with languages or prompts that are not in the training data.

H. Details of the M^{3}CelebA Dataset

In this section, we explain our method for generating captions and the details and some statistics about our M^{3}CelebA dataset.

We used the few-shot technique with GPT-3.5 Turbo [2] to generate captions for images based on their attributes. We then picked the top 1500 captions by human evaluation, which were subsequently used to fine-tune the GPT-3.5 Turbo. We fine-tuned the model for five epochs using the default training parameters. The fine-tuned model was then used to generate captions for all images. We show this process in Fig. 16.

In Tab. 6, we compare available CelebA-based datasets. The CLIP Score of M^{3}CelebA captions is 3.1% higher than

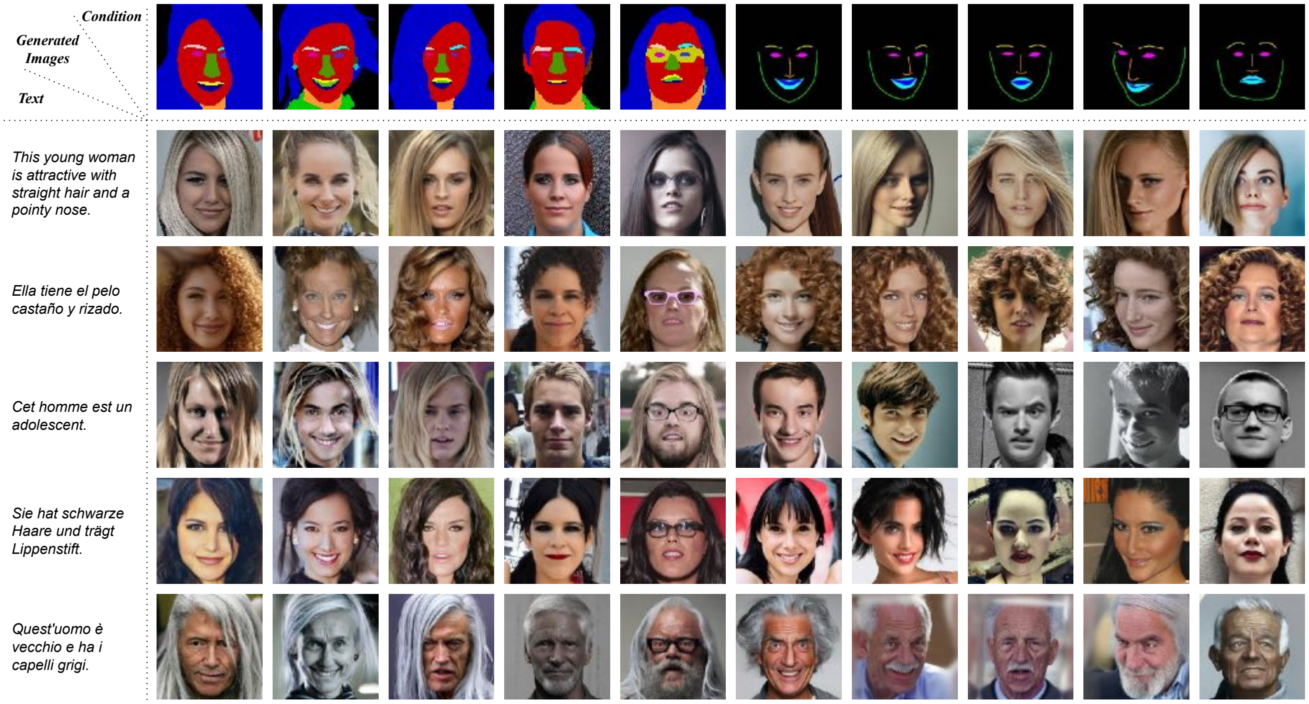

Condition
Generated Images
Text
This young woman is attractive with straight hair and a pointy nose.
Ella tiene el pelo castaño y rizado.
Cet homme est un adolescent.
Sie hat schwarze Haare und trägt Lippenstift.
Quest'uomo è vecchio e ha i capelli grigi.

Figure 6. **Face Generation Results.** Our method generates realistic images based on the input prompt and the conditioning modality. We can generate faces consistent with semantic segmentation and facial landmarks. It also captures difficult attributes in the input prompt or the segmentation, such as glasses, hair color and style, and different face directions.

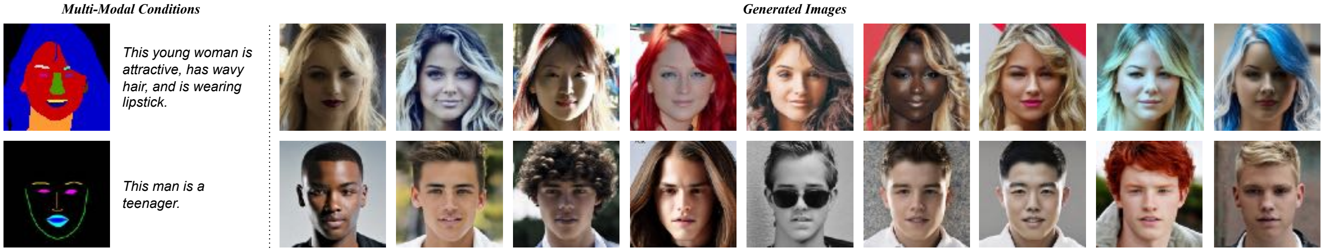

Multi-Modal Conditions
Generated Images
This young woman is attractive, has wavy hair, and is wearing lipstick.
This man is a teenager.

Figure 7. **Diversity of Face Generation.** Diverse face images can be generated from a single text input and a conditioning modality. We can see a diverse set of features such as hair colors, styles, skin tones, glasses, and hats.

the Multi-Modal CelebA-HQ [34] captions, showing that the quality of our generated captions is better. We show the distribution of different attributes in the dataset in Fig. 14. In Fig. 15, we compare the distribution of several important attributes to the CelebA-HQ dataset. We see improvements in several attributes such as Male/Female, Eyeglasses, and Black Hair. The number of face images with "Attractive" attributes has decreased, contributing to reducing bias in the dataset.

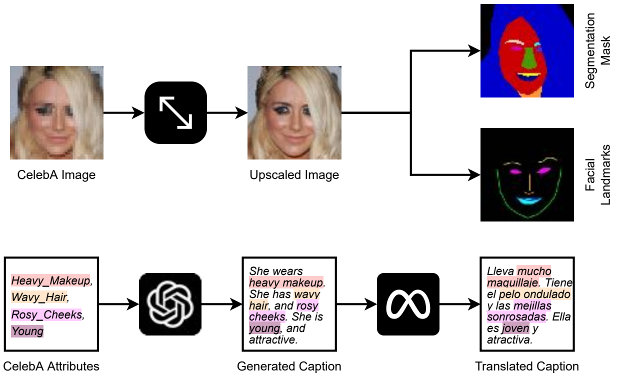
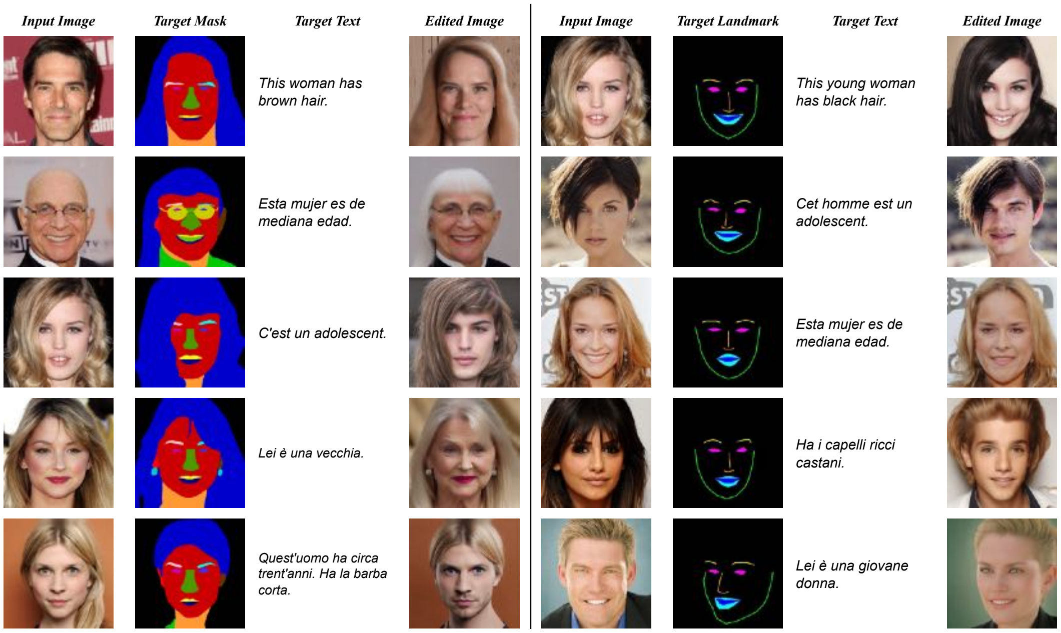

Segmentation Mask
CelebA Image
Upscaled Image
Facial Landmarks
Heavy_Makeup, Wavy_Hair, Rosy_Cheeks, Young
She wears heavy makeup. She has wavy hair, and rosy cheeks. She is young, and attractive.
Lleva mucho maquillaje. Tiene el pelo ondulado y las mejillas sonrosadas. Ella es joven y atractiva.
CelebA Attributes
Generated Caption
Translated Caption

Figure 13. **Dataset Generation Pipeline.** We first align, crop, and upscale the original CelebA images. We then generate the facial landmarks and semantic segmentation for each image. For generating the captions, we use the 40 CelebA attributes and utilize the GPT3.5 model for generation and the SeamlessM4T model for translation.



This woman has brown hair.

Esta mujer es de mediana edad.

C'est un adolescent.

Lei è una vecchia.

Quest'uomo ha circa trent'anni. Ha la barba corta.

This young woman has black hair.

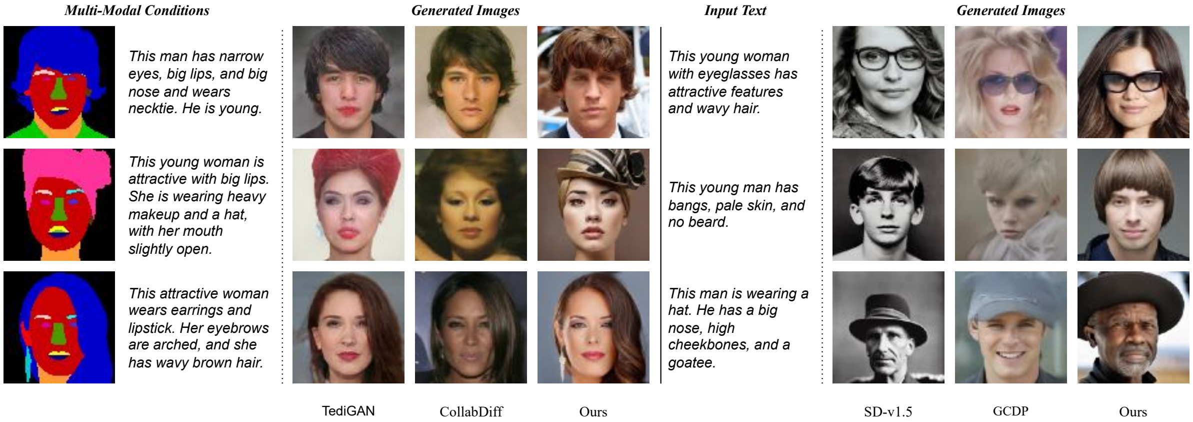
Cet homme est un adolescent.

Esta mujer es de mediana edad.

Ha i capelli ricci castani.

Lei è una giovane donna.

Figure 8. **Face Editing Results.** We can edit face images with both semantic segmentation and facial landmarks. The results are consistent with the conditioning and the text prompt. We achieve better results in details such as hair color and style.



This man has narrow eyes, big lips, and big nose and wears necktie. He is young.

This young woman is attractive with big lips. She is wearing heavy makeup and a hat, with her mouth slightly open.

This attractive woman wears earrings and lipstick. Her eyebrows are arched, and she has wavy brown hair.

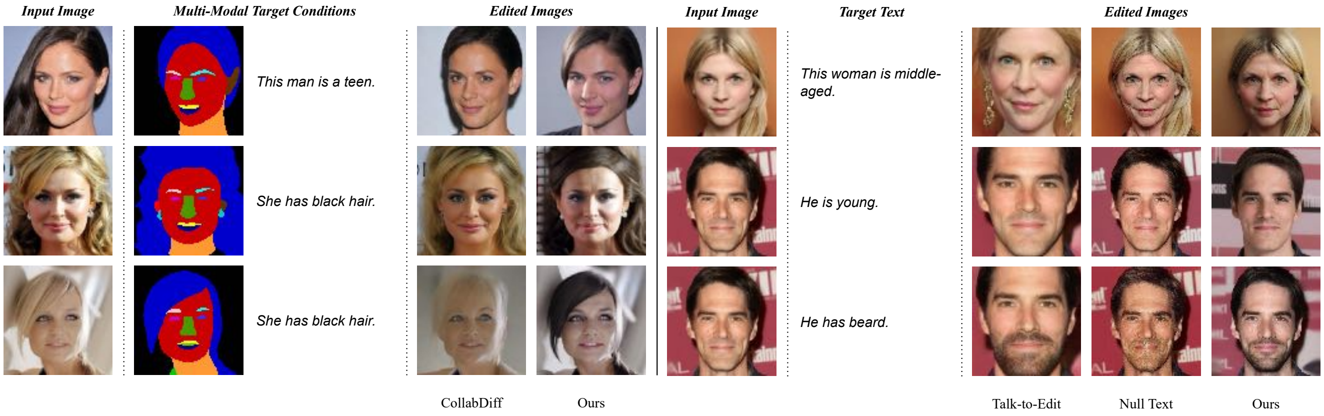
This young woman with eyeglasses has attractive features and wavy hair.

This young man has bangs, pale skin, and no beard.

This man is wearing a hat. He has a big nose, high cheekbones, and a goatee.

(a) Multi-Modal Face Generation **(b) Text-Guided Face Generation**

Figure 9. **Qualitative Comparison of Face Generation.** In **(a) Multi-Modal Face Generation**, TediGAN fails to follow both the target prompt and segmentation mask in most examples because of its two-step generation method. On the other hand, Collaborative Diffusion often cannot generate details such as earrings, make-up, or a detailed beard. In **(b) Text-Guided Face Generation**, only our method can generate detailed images with hight quality.



This man is a teen.

She has black hair.

She has black hair.

This woman is middle-aged.

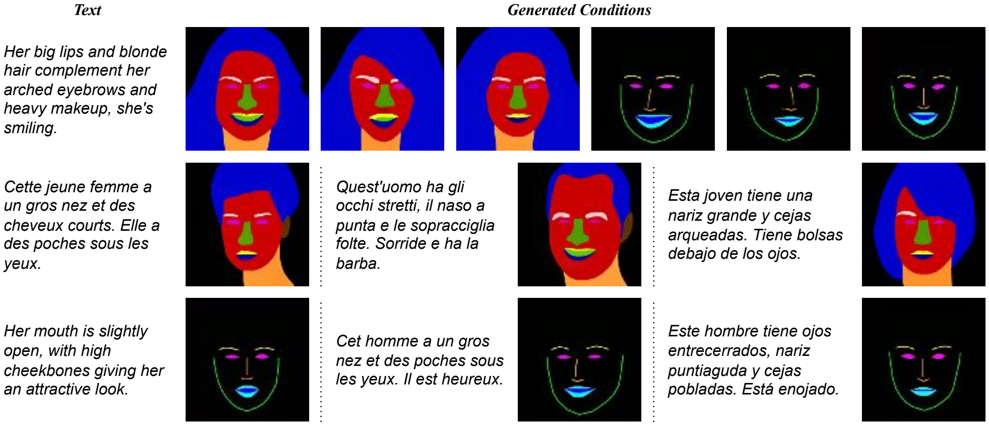
He is young.

He has beard.

(a) Multi-Modal Face Editing

(b) Text-Guided Face Editing

Figure 10. **Qualitative Comparison of Face Editing.** In **(a) Multi-Modal Face Editing**, our method significantly outperforms Collaborative Diffusion. While Collaborative Diffusion fails to follow both the target prompt and segmentation mask in examples that require editing details such as the hair color or style, our method manipulates faces accurately and maintains the original face identity. In **(b) Text-Guided Face Editing**, our method addresses limitations observed in existing approaches. Talk-to-Edit exhibits inconsistent performance, often failing to apply edits effectively on in-the-wild images. Null-Text Inversion maintains the original face identity well but is too sensitive to the edit prompt and sometimes fails to edit attributes accurately.



Her big lips and blonde hair complement her arched eyebrows and heavy makeup, she's smiling.

Cette jeune femme a un gros nez et des cheveux courts. Elle a des poches sous les yeux.

Quest'uomo ha gli occhi stretti, il naso a punta e le sopracciglia folte. Sorride e ha la barba.

Esta joven tiene una nariz grande y cejas arqueadas. Tiene bolsas debajo de los ojos.

Her mouth is slightly open, with high cheekbones giving her an attractive look.

Cet homme a un gros nez et des poches sous les yeux. Il est heureux.

Este hombre tiene ojos entrecerrados, nariz puntiaguda y cejas pobladas. Está enojado.

Figure 11. **More Mask/Landmark Genertation Results.** Our mask/landmark generator can create diverse and accurate conditions for the face generator model.

Name	Size	Segmentation Map	Landmark	Caption	Multilingual
CelebA [21]	202599	✗	✗	✗	✗
CelebA-HQ [14]	30000	✗	✗	✗	✗
CelebAMask-HQ [20]	30000	✓	✓	✗	✗
CelebA-Dialog [13]	202599	✗	✗	✓	✗
MM-CelebA-HQ [34]	30000	✓	✓	✓	✗
M^3CelebA (Ours)	173314	✓	✓	✓	✓

Table 6. **Comparison of Different CelebA-Based Datasets.** We compare different characteristics of datasets created based on the CelebA [21] dataset.

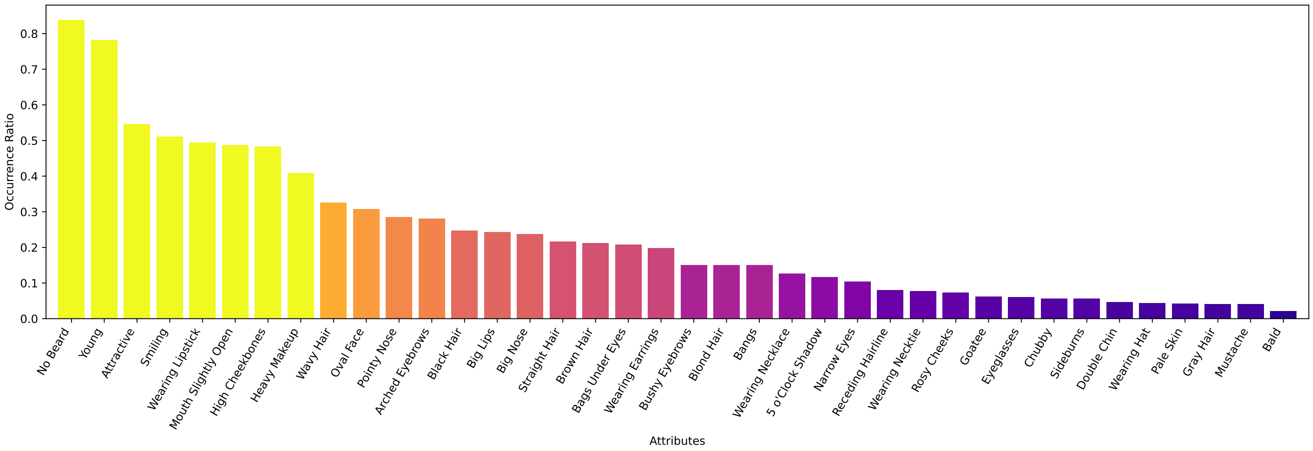


Figure 14. **Attribute Distribution in Dataset.** Occurrence ratio of different face attributes in M^{3}CelebA dataset.

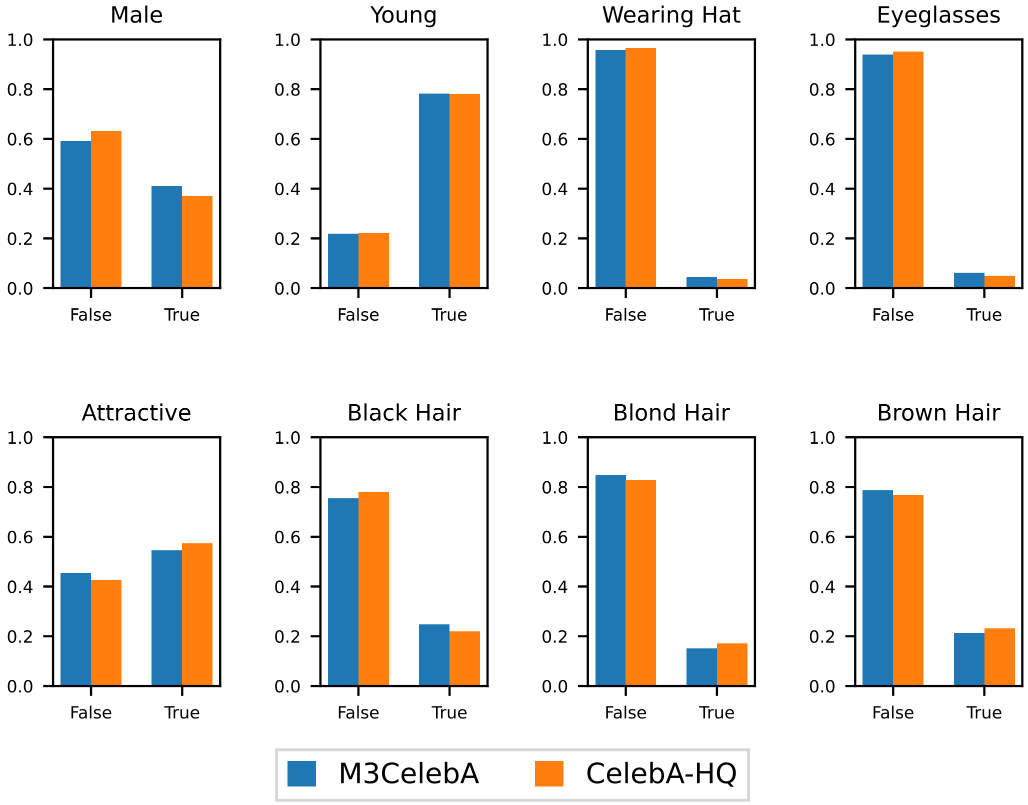


Figure 15. **Attribute Distribution Comparison.** We compare the occurrence ratio of different face attributes in CelebA-HQ and M^{3}CelebA datasets. Our dataset is more balanced in terms of Male/Female and Non-Attractive/Attractive attributes.

Few Shot Caption Generation

```
[system]:
Generate 3 captions in JSON format for a face image from its attributes.

[user]:
Wearing Lipstick, Smiling, Blond Hair, Young, Wavy Hair, Female, Oval Face

[assistant]:
{"caption1": "This woman is wearing lipstick, smiling with blond wavy hair. She has a young oval face.", "caption2": "With
a young oval face, this woman is wearing lipstick, smiling with wavy blonde hair.", "caption3": "The female with a young
oval face is wearing lipstick, has wavy blond hair, and is smiling."}

[user]:
Wearing Necktie, Bags Under Eyes, Double Chin, Straight Hair, Male
```

```
{"caption1": "This man wears a necktie and has bags under his eyes. He has a double chin and straight hair.", "caption2":
"The man has bags under his eyes and a double chin. He wears a necktie and has straight hair.", "caption3": "With a necktie
and straight hair, this man has bags under his eyes and a double chin."}
```

Model Fine-tuning

```
[system]:
Generate 3 captions in JSON format for a face image from its attributes.

[user]:
Wearing Necktie, Bags Under Eyes, Double Chin, Straight Hair, Male

[assistant]:
{"caption1": "This man wears a necktie and has bags under his eyes. He has a double chin and straight hair.", "caption2":
"The man has bags under his eyes and a double chin. He wears a necktie and has straight hair.", "caption3": "With a necktie
and straight hair, this man has bags under his eyes and a double chin."}
```

Dataset Caption Generation

```
[system]:
Generate 3 captions in JSON format for a face image from its attributes.

[user]:
Wearing Lipstick, Wearing Earrings, High Cheekbones, Female, Heavy Makeup, Big Lips, Pale Skin
```

```
{"caption1": "This woman is wearing lipstick and earrings. She has high cheekbones and heavy makeup.", "caption2": "The
female has big lips, pale skin, and high cheekbones. She wears lipstick and earrings.", "caption3": "With high cheekbones
and pale skin, this woman wears lipstick and has heavy makeup."}
```

Figure 16. **Caption Generation Process.** The prompts used in the few-shot and fine-tuning processes are shown in this figure.