# FEBench : Multi-dimensional Facial Editing Benchmark

Dongsoo Shin, Shihyung Park, Seohee Kim
Seoul National University
{dongsoo, psh0416, ksh0529}@snu.ac.kr

## Abstract

*Facial image editing demands high fidelity and semantic precision due to its impact on identity perception, yet existing benchmarks fall short in evaluating these unique challenges. We propose FEBench, a comprehensive benchmark tailored for text-guided facial editing. FEBench covers a diverse set of editing tasks—including nasal bridge augmentation, chin and jawline reshaping, shadow removal, and emotion modification—across multiple editing paradigms: mask-based, prompt-based, and instruction-based. It introduces a multi-dimensional evaluation framework encompassing editing fidelity, non-target consistency, and overall image quality supported by multiple metrics. Through extensive experiments with four representative editing models, we demonstrate FEBench's effectiveness in enabling holistic and comparative analysis of facial editing capabilities.*

## 1. Introduction

Facial image editing has emerged as a crucial task in various domains such as beauty enhancement, cosmetic surgery simulation, and digital entertainment, due to the demand for high-fidelity and user-controllable facial transformations. Among the approaches proposed to address this task, text-guided editing methods [1–3] have garnered increasing interest due to their intuitive and user-friendly interfaces, enabling users to express desired edits through natural language descriptions.

While text-guided editing provides a convenient interface, facial image editing inherently demands a high degree of precision. Even minor alterations can significantly affect a person's perceived identity—potentially resulting in the edited face appearing as a different individual altogether(e.g. Change in skin texture or skin tone) [4]. Such discrepancies may lead to misidentification in digital entertainment, thereby disrupting viewer immersion; cause critical misunderstandings in cosmetic surgery planning; and reduce user satisfaction even in casual or recreational use cases.

Therefore, achieving high accuracy and precision in facial image editing is essential, as it entails challenges that fundamentally differ from those encountered in general image editing tasks. Such challenges highlight the necessity for dedicated benchmarks tailored to the specific demands of facial editing. Nevertheless, existing benchmarks [5–8] have primarily focused on general text-guided image editing, and to date, no benchmark has been specifically tailored to the unique challenges of face-specific editing.

Consequently, there remains a significant gap in the development of dedicated evaluation protocols tailored to facial editing, a task which demands higher granularity and semantic precision. The absence of such benchmarks limits the ability to rigorously evaluate and compare the performance of facial editing models in a consistent and reproducible manner.

Therefore we introduce FEBench, a benchmark designed explicitly for evaluating text-guided facial editing models. Unlike prior benchmarks, FEBench supports diverse types of editing models—including mask-based editing, target prompt editing, and instruction-based editing—thereby enabling a more holistic assessment across various editing scenarios. Our benchmark evaluates facial edits that are commonly desired in everyday life, and employs several metrics specifically tailored for assessing face editing, that have not been used in previous benchmarks.

## 2. Related Works

**Image Editing Models.** Image editing models aim to generate modified images based on user-defined conditions such as text prompts, masks, or sketches. Early GAN-based [9] approaches like StyleGAN [10] focusing mainly on holistic style transfer, while models like MaskGan [11], BrushNet [12] performed inpainting in the masked region by fooling both local and global discriminators into classifying the composite output as real. With the advent of diffusion models [13, 14], fine-grained and semantically controllable editing became possible. Techniques like ControlNet [15] extended diffusion-based methods by enabling structured conditional inputs (e.g., pose, edge maps, or masks). Moreover, the emergence of CLIP [16] enabled alignment between image content and natural language, facilitating

(a) (b) (c) (d) (e)

Figure 1. **Example of editing tasks evaluated in our benchmark.** (a) Nasal bridge augmentation. (b) Chin reduction. (c) Jawline reduction. (d) Shadow removal. (e) Emotion change. Top row indicates original image and bottom row indicates edited images.

semantically rich, text-driven editing. Most recent editing methods like Prompt-to-Prompt [1] adopt an inversion-and-generation strategy, where an image is first mapped to a latent noise space before editing is applied through new prompts [17]. Few methods like instructPix2Pix [18] and MagicBrush [19] fine-tune generation models to edit by receiving instruction prompt as an input. FlowEdit [20] takes this further by eliminating the inversion step, allowing more direct and efficient edits.

**Facial Editing Models.** Facial image editing presents unique challenges compared to general image editing, due to the need for high precision and identity preservation. Even subtle modifications may compromise the perceived identity, which is critical in applications such as digital avatars, beauty retouching, or virtual try-ons. In the 2D domain, various model families have been proposed based on different forms of user guidance. ManiClip [21] and CA-Edit [4] are models specifically designed for 2D facial editing. However, other models, such as BrushNet [12], Prompt-to-Prompt [1], FlowEdit [20], and MagicBrush [19], were originally designed for general image editing but can also be effectively adapted for facial editing tasks. While these approaches have shown promise, the absence of standardized evaluation protocols makes it difficult to objectively compare their performance. In particular, facial editing demands specific evaluation criteria that account for semantic alignment, regional fidelity, and identity preservation to ensure the quality of edits while maintaining the integrity of the subject's appearance.

**Image Editing Benchmarks.** A number of benchmarks have been proposed to evaluate the performance of image editing models. EditBench [6] focuses solely on mask-guided inpainting tasks, offering a controlled setup for evaluating models based on spatially localized user input. I2EBench [7], in contrast, targets instruction-based editing scenarios, where models are guided by free-form natural language commands. While these benchmarks have advanced research in general-purpose image editing, their evaluation scope remains limited—each supports only a narrow subset of editing paradigms. Consequently, they are ill-suited for assessing the increasing diversity of editing models, particularly those that operate outside their predefined task formats. Recent efforts such as HATIE [8] have aimed to align evaluation more closely with human perception. By aggregating multiple metrics and automatically generating editing instructions using VQA annotations [22] and large language models [23], HATIE provides a more comprehensive and perception-aligned benchmark. However, it remains focused on general image content and does not explicitly account for the semantic constraints and evaluation challenges unique to facial image editing. These limitations point to the need for a dedicated benchmark that supports diverse facial editing paradigms—including mask-based, description-based, and instruction-based methods—while incorporating evaluation criteria that reflect both perceptual consistency and identity preservation.

## 3. FEBench

Our benchmark targets to evaluate everyday facial retouching tasks, including structural editing, shadow removal, emotion change. We provide a detailed definition of our evaluated editing tasks in Sec. 3.1, used metrics for evaluation in Sec. 3.2, and overall evaluation framework in Sec. 3.3.

### 3.1. Task Definition

**Facial Structural Edits.** To develop a benchmark specifically tailored for facial analysis, we focused on structural

changes in the face. These structural changes were categorized into three distinct types, as detailed below:

- **Nasal Bridge Augmentation:** As a centrally positioned facial feature, the nose plays a pivotal role in human facial recognition and perception. Due to its prominence, it is also among the most frequently targeted regions for facial modification, both in personal use and commercial applications. To reflect this demand, our benchmark includes a nasal bridge augmentation task, designed to assess a model's capability to perform fine-grained and semantically precise facial edits.
- **Chin Reduction:** Chin reduction is a facial editing task of decreasing the vertical length or anterior projection of the lower face. This modification is typically applied to elongated or protruding chins, with the aim of adjusting the overall vertical proportions of the face and alleviating the perceived visual imbalance associated with a long lower face. From an aesthetic perspective, an excessively long chin can influence perceived facial attractiveness and gender recognition. As such, chin reduction is frequently requested to enhance facial harmony and achieve a more balanced appearance.
- **Jawline Contouring:** Jawline contouring involves modifying the width or angularity of the mandibular angle to create a smoother and more refined jaw contour. This task is typically applied in cases where the lower jaw appears wide or angular, with the aim of producing a narrower, more tapered facial shape—commonly referred to as a "V-shaped facial contour". A prominent jawline can contribute to a wide facial impression, and preferences for jaw shape are often influenced by cultural and aesthetic standards.

**Emotion Change.** Emotion change refers to a facial image editing technique aimed at altering a subject's expression to convey a desired emotional state. For instance, this method may transform a squinting face—caused by strong sunlight—into a smiling face, or convert a neutral expression into a surprised look. Such techniques have gained growing attention in both personal and commercial domains, with applications ranging from social media content generation and advertisement portrait enhancement to the aesthetic refinement of family photograph.

**Shadow Removal.** Everyday facial photographs often contain shadows due to the position of the light source or occlusion by objects. Shadow removal refers to the editing process that adjusts the skin tone in shadowed regions to match that of the illuminated areas, effectively eliminating the appearance of shadows on the face. Shadow removal is also in high demand for producing ID photos, where the face is expected to appear evenly lit without any shadows.

We abbreviate each task for simplicity in the following sections, nasal bridge augmentation as NOSE, chin reduction as CHIN, jawline reduction as JAWLINE, emotion change as EMOTION and shadow removal as SHADOW.

### 3.2. Metrics

**CLIP Similarity.** CLIP computes the similarity between an image and a text prompt by projecting both into a shared semantic embedding space. This capability makes it a natural choice for evaluating how well an edited image reflects a given textual description. In our context, it serves as a measure of semantic alignment between the intended prompt and the visual outcome.

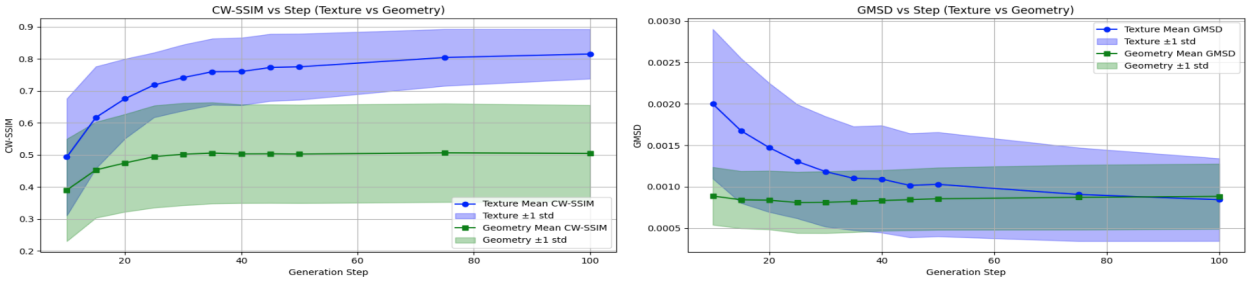
**Image Quality Measurement.** We use various metrics to assess the overall quality of the generated images, ensuring that they closely resemble the natural appearance of real-world images.

- **Fréchet Inception Distance (FID [24])**:The Fréchet Inception Distance (FID) evaluates the quality of generated images by comparing their feature distributions to those of real images. A lower FID indicates that the generated images are more similar to the real dataset in terms of both content and diversity.
- **Q-ALIGN [25]**: Q-ALIGN teaches Large Multimodality Models (LMMs) to assess image quality by mimicking how humans rate images using discrete text-defined levels (e.g., "good," "poor") instead of exact numerical scores. During inference, the LMM predicts probabilities for each level, and the final image quality score is obtained by computing a weighted average of these probabilities.

**Non-target Preservation.** To evaluate visual consistency across different regions of an image—excluding areas of intended modification—we employ several metrics:

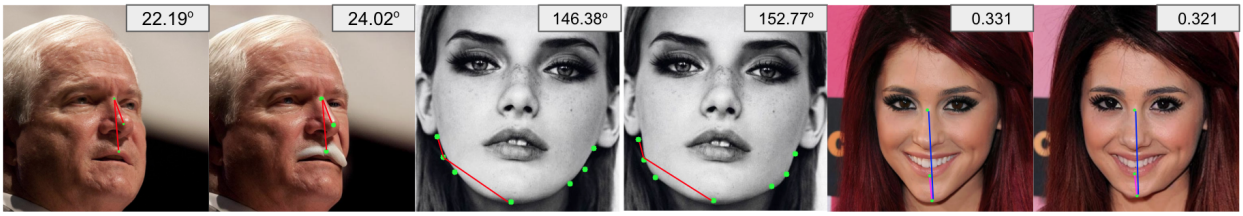
- **DINO [26], LPIPS [27]**: Both DINO and LPIPS assess image similarity using deep feature representations rather than raw pixels, enabling evaluation at a higher perceptual and semantic level. DINO, a self-supervised vision transformer, captures semantic and structural aspects of the image through its learned embeddings. LPIPS (Learned Perceptual Image Patch Similarity) measures perceptual similarity by comparing intermediate features from deep neural networks, aligning closely with human visual perception.
- **L2 Distance**: L2 distance measures the pixel-wise difference between two images.

These metrics together offer a comprehensive assessment of how well the edited image preserves non-target regions—capturing both high-level semantic/perceptual fi-




(a) CW-SSIM Verification (b) GMSD Verification

Figure 2. **Verification of our selected texture metrics.** We compare similarity to the original image under texture (blue) and structure (green) changes. CW-SSIM increases and GMSD decreases as texture becomes more realistic, while remaining stable under structural shifts, confirming their sensitivity to texture changes and robustness to geometry.




(a) (b) (c)

Figure 3. **Example of utilizing facial landmarks for structural edit evaluation.** (a) Evaluation of nasal bridge augmentation. (b) Evaluation of chin reduction. (c) Evaluation of Jawline reduction.

delity (via DINO and LPIPS) and low-level pixel differences (via L2 distance).

**Identity Preservation.** To ensure that the underlying identity of a person remains recognizable after editing—regardless of changes in facial structure, style, or lighting—we use FaceID models.

- **FaceNet [28], Keypoint Regression with Perceptual Embedding (KPRPE) [29]**: These models encode facial images into identity-preserving embeddings. The similarity between two embeddings indicates how likely the faces belong to the same individual. KPRPE, in particular, is designed to be robust to pose changes, making it well-suited for evaluating identity consistency under diverse editing conditions.

In our work, we use FaceID similarity to assess how well the core identity of the person is preserved after editing. This is crucial because a visually appealing facial edit is not meaningful if the resulting image is no longer recognizable as the same person.

**Texture Realism Preservation.** Facial editing often leads to subtle yet perceptible changes in skin or hair textures—such as unnaturally uniform smoothness or loss of fine surface variation—that can evoke a sense of visual dissonance or uncanniness, even when the overall structure appears intact. To assess whether these delicate textures are faithfully retained after editing, we employ texture-sensitive metrics.

- **GMSD, CW-SSIM**: We use Gradient Magnitude Similarity Deviation (GMSD) and Complex Wavelet Structural Similarity Index (CW-SSIM), both of which are sensitive to subtle texture changes while being robust to larger structural transformations which can also be seen in Fig. 2. GMSD focuses on local gradient patterns, capturing distortions in fine textures such as skin pores or hair strands. CW-SSIM, built on wavelet transforms, emphasizes localized phase consistency and is particularly suited for detecting structural inconsistencies in texture without being affected by global shape changes.

We conducted a verification experiment on these two metrics as in Fig. 2, by evaluating the reconstructed images of same person, degraded gradually by decreasing the number of generation steps in diffusion process. The results show our metrics successfully captures the preservation of texture realism. Example of measuring CW-SSIM can be also found in Fig. 4.


Figure 4. **Examples of texture realism preservation measurement.** We measure CW-SSIM similarity between the edited and original images. (a) shows texture changes by altering number of generating steps in reconstruction, using diffusion model.

**Head Pose Consistency.** To evaluate head pose consistency, we estimate the 3D head orientation (yaw, pitch, and roll) for both the original and edited images using facial landmarks and the RANSAC-based Perspective-n-Point (PnP) algorithm with a predefined 3D facial model [30]. The resulting rotation vectors are converted to Euler angles, and angular differences are computed along each axis. These differences are normalized based on typical human head rotation ranges and aggregated into a similarity score ranging from 0 (identical pose) to 1 (maximally different). Lower scores reflect higher pose consistency.

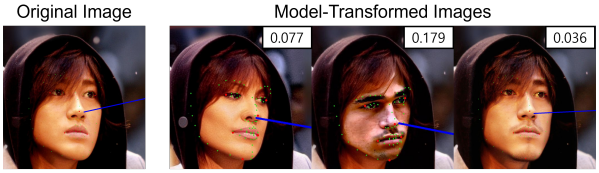



Figure 5. **Examples of head pose consistency measurement.** In all images, green dots indicate facial landmarks, red dots represent the subset used for facial pose consistency evaluation, and the blue line denotes the estimated facial orientation. Lower value indicates good head pose consistency.

**Facial Geometry Change Measurement.** We utilized the 68-point facial landmark model [30] to quantify changes in the nose, chin, and jawline regions. The corresponding evaluation metrics for each region are defined as follows:

- **NOSE:** To evaluate nasal bridge augmentation, we measure the angle formed at landmark 27 (the top of the nasal bridge) with respect to landmark 30 (the base of the nose) and landmark 51 (the top of the upper lip). An increase in this angle compared to the original image indicates a more elevated or prominent nasal bridge.
- **CHIN:** We assess changes in chin length by calculating the ratio formed by three facial landmarks: point 27 (the top of the nasal bridge), point 57 (the bottom of the lower lip), and point 8 (the lowest point on the chin). A reduction in this ratio compared to the original image indicates a decrease in perceived chin length.
- **JAWLINE:** To quantify jawline slimming, we measure angular changes between lateral jaw landmarks (4, 5 on the left and 11, 12 on the right) and reference points such as 3 and 13 (angles near the ears) and 8 (chin bottom). An increased average angle in the edited image indicates a more tapered jawline, reflecting a slimmer and sharper mandibular contour.

**Shadow Removal Measurement.** We use three metrics to evaluate shadow removal quality. Face symmetry difference measures the mean absolute difference between the luminance of a face and its mirrored version, after aligning the face using landmarks. Retinex illumination range computes the standard deviation of the log-reflectance, estimating local brightness uniformity—higher variation indicates shadow presence. Shadow edge contrast detects strong dark-to-bright transitions via Canny and Laplacian filters, then compares average gradient magnitudes across these regions in the original and edited images. Example of evaluation can be seen in Fig. 6.

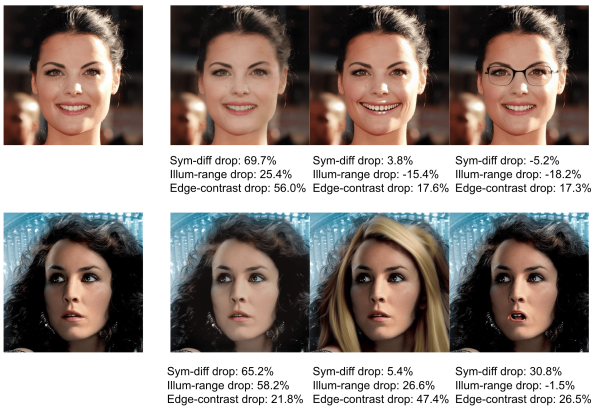



Figure 6. **Example of shadow removal measurement.** High percentage value indicates good shadow removal.

**Emotion Classification Confidence Measurement.** We assess the fidelity of the intended emotional expression in the edited image. The evaluation encompasses seven emotion categories: Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, and Contempt. Emotion classification is conducted using LibreFace [31], a deep learning-based Facial Expression Recognition (FER) model, which outputs confidence scores for each class. Based on these confidence values, a normalized score ranging from 0 to 1 is computed to quantify how accurately the target emotion is reflected in the edited output.

### 3.3. Evaluation Framework

**Dataset Selection.** Due to the presence of facial editing models requiring masks as an input, we select CelebAMask-HQ [11] as our base dataset. CelebAMask-HQ contains 30,000 human facial images with 512×512 sized masks, each annotated with 19 classes including facial attributes and accessories (e.g. skin, lips, hair, earrings, hat).

**Annotation Generation.** For each editing task, we produce an annotation bundle comprising: the original image, the target region mask, original emotion, a caption of the original image, a description of the desired edited image, and a editing instruction. We created caption of the original image with llama-3-vision [32], which is a VLM. Then we utilized GPT-4o-mini [23] to generate a description of desired edited image, by requesting to model to modify the original caption. By this process, we created 20 annotations for each editing task, resulting in total 100 annotations. Each model then edits the image following the created annotation.

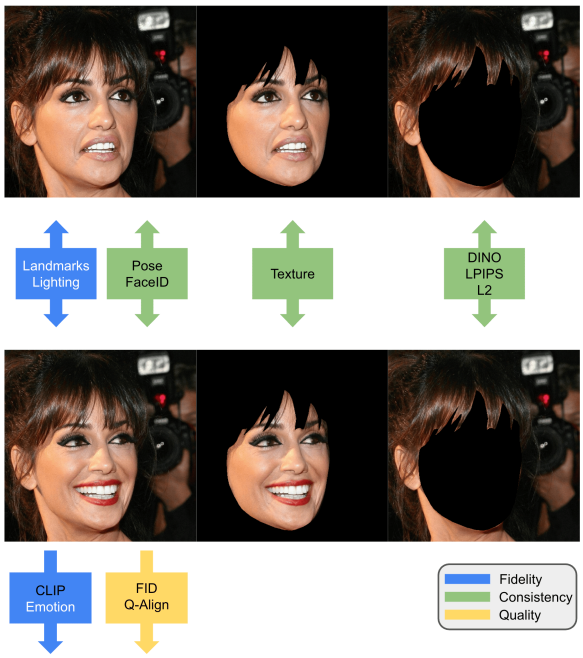



Figure 7. **Evaluation workflow of FEbench.** The top row presents the original image and its corresponding target and non-target segments, while the bottom row presents those of the edited image. Metrics shown in black are commonly computed across all editing tasks, while metrics shown in white are specific to the corresponding editing task. Bidirectional diagram calculates the metric by comparing the original and edited image, unidirectional diagram calculates the metric with only edited image.

**Evaluation.** After the edited images are obtained, various metrics are aggregated to evaluate the model's face editing capability by aggregating multiple diverse metrics explained in Sec. 3.2. Overall framework is described in Fig. 7. We categorize the metrics used for evaluation into following three groups:

- **Fidelity:** We categorize fidelity as how well the edited image reflect the editing request. We compute CLIP [16] alignment between edited image and the target description. Additionally, for each type of tasks, corresponding fidelity metric is computed together. Facial geometry change measurement and shadow removal measurement is computed between original image and edited image, emotion classification confidence is measured from edited image.
- **Consistency:** We categorize consistency as how well the edited image preserves the non-target properties. We invert the target segment mask to extract the non-target region from both the edited image and the original image, then measure their similarity using LPIPS [27], L2 distance, and DINO [26] features. Moreover, we measure the faceID similarity and face pose similarity between original and edited image. Then measure texture similarity between segmented skin.
- **Quality:** We measure overall quality of the edited image by assessing global realism with FID [24] and Q-ALIGN-Quality [25].

## 4. Experimental Results

We conduct facial editing with four models, covering all the aforementioned types of facial editing models. We select BrushNet [12] as a mask-based model, Prompt-to-Prompt [1] and FlowEdit [20] as a description-based model, and MagicBrush [19] as an instruction based model. CA-edit [4] and ManiCLIP [21] are facial editing-specific models, but due to CA-edit being unavailable and ManiCLIP's inability to edit real images, we opted to evaluate general-purpose editing models.

We show our quantitative benchmark evaluation results in Tab. 1 and Tab. 2 each aggregated by task types and score criteria. To avoid an excessively narrow range of scores, spreading is first applied. Then, the results were aggregated by averaging multiple evaluation metrics. As a result, Tab. 1 shows that BrushNet was the best in NOSE, and FlowEdit was the best in remaining tasks. In terms of score criteria, BrushNet was the best in satisfying high editing fidelity and overall image quality. FlowEdit was the best in preserving non-target consistency. Qualitative results of edited images is presented in Fig. 8. As seen in Fig. 8, most of the models fail in editing CHIN, JAWLINE, SHADOW. BrushNet attempts to edit the facial geometry, but largely alters the identity of the person. Note that due to the different editing paradigms of each model, we avoid which model is superior among the models.

| Models | NOSE | CHIN | JAWLINE | SHADOW | EMOTION |
|---|---|---|---|---|---|
| BrushNet | 0.878 | 0.673 | 0.688 | 0.585 | 0.627 |
| Prompt-to-Prompt | 0.751 | 0.780 | 0.792 | 0.670 | 0.757 |
| FlowEdit | 0.844 | 0.839 | 0.858 | 0.689 | 0.780 |
| MagicBrush | 0.786 | 0.790 | 0.725 | 0.658 | 0.751 |

Table 1. **Comparison of facial editing results across editing task types.**

| Models | Fidelity | Consistency | Quality |
|---|---|---|---|
| BrushNet | 0.441 | 0.709 | 0.871 |
| Prompt-to-Prompt | 0.422 | 0.841 | 0.768 |
| FlowEdit | 0.430 | 0.911 | 0.791 |
| MagicBrush | 0.381 | 0.838 | 0.778 |

Table 2. **Comparison of facial editing results across score criteria.**

## 5. Conclusion and Limitations

In this paper, we present FEBench, an unprecedented multi-dimensional facial editing benchmark that focuses on evaluating everyday facial edits. Evaluation was conducted by

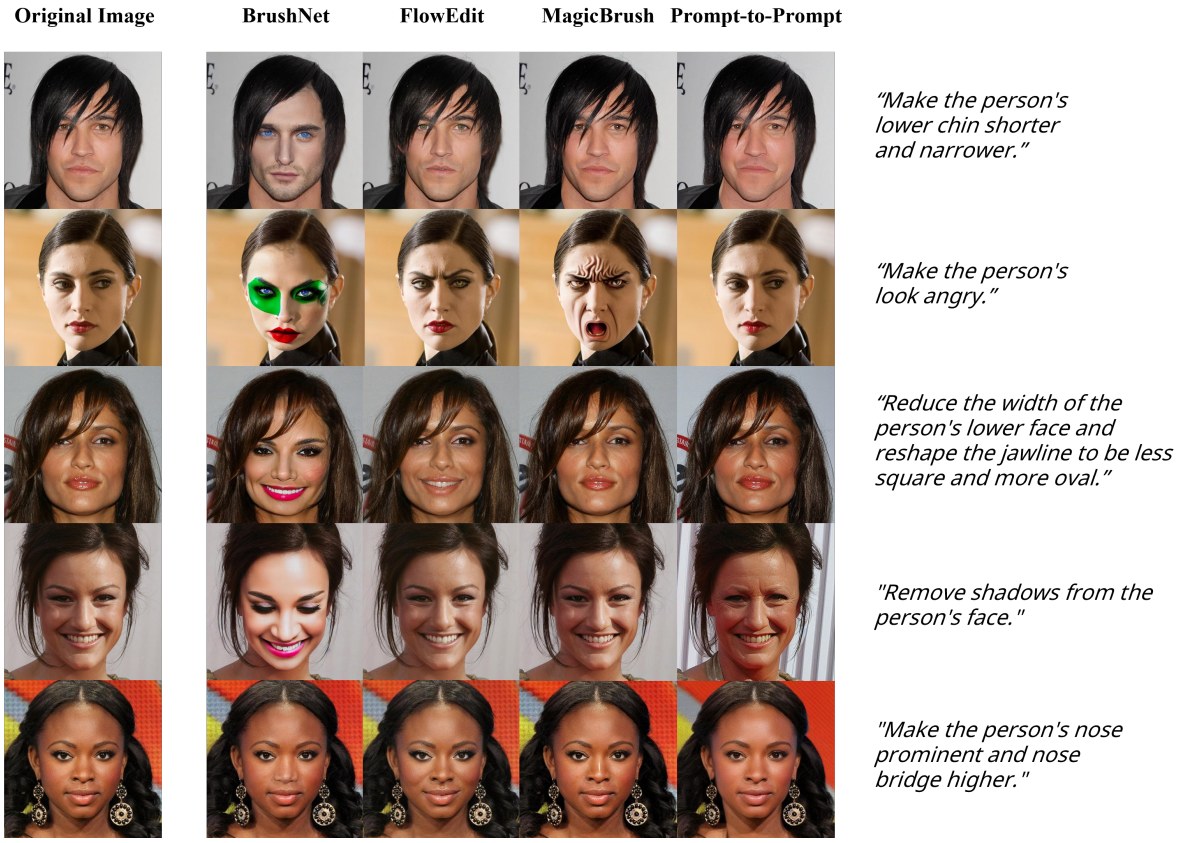



Figure 8. **Examples of edited outputs by four editing models.**

combining various metrics, which we categorized into fidelity, consistency and quality according to their purposes. FEBench can be easily scaled by utilizing VLM and LLM for generating prompts needed as an input for editing models.

There are few limitations of our work. First, our first aim was to follow the idea of HATIE [8], so we should also fit the coefficients of aggregating the multiple metric scores to better align with the human perception. Second, validation on some metrics we used is insufficient. The suitability of the shadow removal measurement, head pose consistency and facial geometry change measurement was empirically confirmed on a small number of samples. These limitations will be resolved in our future work.

## References

[1] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022. 1, 2, 6

[2] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations*, 2022.

[3] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," 2023. 1

[4] X. Xian, X. He, Z. Niu, J. Zhang, W. Xie, S. Song, Z. Yu, and L. Shen, "Ca-edit: Causality-aware condition adapter for high-fidelity local facial attribute editing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 8593–8601, 2025. 1, 2, 6

[5] S. Basu, M. Saberi, S. Bhardwaj, A. M. Chegini, D. Massiceti, M. Sanjabi, S. X. Hu, and S. Feizi, "Editval: Benchmarking diffusion based text-guided image editing methods," 2023. 1

[6] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut, *et al.*, "Imagen editor and editbench: Advancing and evaluating text-guided image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18359–18369, 2023. 2

[7] Y. Ma, J. Ji, K. Ye, W. Lin, Z. Wang, Y. Zheng, Q. Zhou, X. Sun, and R. Ji, "I2ebench: A comprehensive benchmark for instruction-based image editing," *arXiv preprint arXiv:2408.14180*, 2024. 2

[8] S. Ryu, K. Kim, E. Baek, D. Shin, and J. Lee, "Towards scalable human-aligned benchmark for text-guided image editing," *arXiv preprint arXiv:2505.00502*, 2025. 1, 2, 7

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,

"Generative adversarial networks," *Commun. ACM*, vol. 63, p. 139–144, Oct. 2020. 1

[10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019. 1

[11] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 5

[12] X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, and Q. Xu, "Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion," in *European Conference on Computer Vision*, pp. 150–168, Springer, 2024. 1, 2, 6

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020. 1

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 1

[15] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023. 1

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021. 1, 6

[17] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2022. 2

[18] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023. 2

[19] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "Magicbrush: A manually annotated dataset for instruction-guided image editing," 2024. 2, 6

[20] V. Kulikov, M. Kleiner, I. Huberman-Spiegelglas, and T. Michaeli, "Flowedit: Inversion-free text-based editing using pre-trained flow models," *arXiv preprint arXiv:2412.08629*, 2024. 2, 6

[21] H. Wang, G. Lin, A. G. del Molino, A. Wang, J. Feng, and Z. Shen, "Maniclip: Multi-attribute face manipulation from text," *International Journal of Computer Vision*, vol. 132, no. 10, pp. 4616–4632, 2024. 2, 6

[22] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019. 2

[23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. 2, 5

[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017. 3, 6

[25] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun, *et al.*, "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," *arXiv preprint arXiv:2312.17090*, 2023. 3, 6

[26] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021. 3, 6

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. 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. 3, 6

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015. 4

[29] M. Kim, Y. Su, F. Liu, A. Jain, and X. Liu, "Keypoint relative position encoding for face recognition," 2024. 4

[30] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017. 5

[31] D. Chang, Y. Yin, Z. Li, M. Tran, and M. Soleymani, "Libreface: An open-source toolkit for deep facial expression analysis," 2023. 5

[32] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, *et al.*, "The llama 3 herd of models," 2024. 5