FED-Bench: A Cross-Granular Benchmark for Disentangled Evaluation of Facial Expression Editing

Fengjian Xue[1⋆], Xuecheng Wu[2⋆], Heli Sun[2†], Yunyun Shi[2], Shi Chen[2], Liangyu Fu[4], Jinheng Xie[5], Dingkang Yang[6], Hao Wang[3†], Junxiao Xue[7], and Liang He[2]

[1] School of Software Engineering, Xi'an Jiaotong University, China
[2] School of Computer Science and Technology, Xi'an Jiaotong University, China
[3] School of Journalism and New Media, Xi'an Jiaotong University, China
[4] School of Software, Northwestern Polytechnical University, China
[5] Department of Electrical and Computer Engineering, NUS, Singapore
[6] College of Intelligent Robotics and Advanced Manufacturing, Fudan, China
[7] Research Center for Space Computing System, Zhejiang Lab, China
{hlsun, haowangx, lhe}@xjtu.edu.cn

Abstract. Facial expression image editing requires fine-grained control to strictly preserve human identity and background while precisely manipulating expression. However, existing editing benchmarks primarily focus on general scenarios, lacking high-quality facial images and corresponding editing instructions. Furthermore, current evaluation metrics exhibit systemic biases in this task, often favoring lazy editing or overfit editing. To bridge these gaps, we propose FED-Bench, a comprehensive benchmark featuring rigorous testing and an accurate evaluation suite. First, we carefully construct a benchmark of 747 triplets through a cascaded and scalable pipeline, each comprising an original image, an editing instruction, and a ground-truth image for precise evaluation. Second, we introduce FED-Score, a cross-granularity evaluation protocol that disentangles assessment into three dimensions: Alignment for verifying instruction following, Fidelity for testing image quality and identity preservation, and Relative Expression Gain for quantifying the magnitude of expression changes, effectively mitigating the aforementioned evaluation biases. Third, we benchmark 18 image editing models, revealing that current approaches struggle to simultaneously achieve high fidelity and accurate expression manipulation, with fine-grained instruction following identified as the primary bottleneck. Finally, leveraging the scalable characteristic of introduced benchmark engine, we provide a 20k+ in-the-wild facial training set and demonstrate its effectiveness by fine-tuning a baseline model that achieves significant performance gains. Our benchmark and code are available at https://github.com/hiixfj/FED-Bench.

Keywords: Benchmark · Data construction · Facial expression · Image editing

⋆ Equal contribution. † Corresponding authors.

Fig. 1: Overview of FED-Bench. The figure illustrates representative samples from our benchmark covering seven basic emotions.

1 Introduction

Recently, driven by the rapid advancement of generative technologies such as Diffusion Models [13], text-guided image editing has achieved remarkable success. Users can now perform rich modifications to image content by simply providing straightforward text prompts. Among its various applications, Facial Expression Image Editing has drawn significant attention due to its immense potential in areas like real-time interaction and visual effects for film and television. However, unlike general image editing tasks such as altering object colors or performing global style transfers, facial expression editing is highly challenging. It requires models to strictly preserve the subject's identity and background information while achieving fine-grained control over expressions. Due to the fine-grained nature of this task, most existing general-purpose editing models struggle to balance these requirements, often resulting in distorted facial features or unintended modifications to non-target regions.

Although facial expression editing technology is evolving rapidly, further development in this field is severely hindered by the lack of rigorous evaluation benchmarks. On the one hand, prevailing image editing benchmarks (e.g., Imgedit [41], EditBench [35], and I2ebench [27]) primarily focus on general scenarios. They severely lack high-quality facial samples and do not design corresponding editing instructions for the subtle movements of facial muscles. On the other hand, and more critically, most current benchmarks only provide a source image and a text instruction, lacking a Ground Truth (GT) reference image for rigorous comparison. Without pixel-level GT as an anchor, existing evaluations remain at a coarse-grained or subjective level, making it impossible to conduct objective and quantitative assessments of a model's true performance in fine-grained expression editing tasks.

In addition to the absence of benchmark data, currently widely-used image quality evaluation metrics exhibit severe systemic biases when applied to facial expression editing tasks. Specifically, the existing evaluation systems often face a dilemma: if heavily weighted towards image fidelity metrics (e.g., DINO Score [3]), the evaluation protocol favors Lazy Editing, where a model receives an exceptionally high score by outputting the source image directly without any

modification. Conversely, if heavily weighted towards semantic alignment metrics (e.g., CLIP Score [11]), the evaluation system easily falls into Overfit Editing, where the model generates overly exaggerated and distorted expressions to cater to the text prompt, even at the expense of the subject's original identity. Such persistent evaluation biases not only fail to reflect the true editing capabilities of the models but may also mislead the design direction of future models.

To break this dilemma and compensate for the deficiencies in existing benchmarks and evaluation systems, we propose FED-Bench: a comprehensive benchmark framework specifically tailored for fine-grained facial expression editing. We collect and clean 747 high-quality data triplets. Each triplet consists of a source image, an editing instruction, and an accurately paired Ground Truth image. Addressing the specific nature of this task, we design a rigorous data filtering pipeline to obtain the GT images, thereby providing a reliable reference for objective and accurate performance evaluation.

Furthermore, based on this dataset, we innovatively introduce FED-Score, a novel cross-granular evaluation protocol. Discarding the ambiguity of single metrics, FED-Score explicitly decouples the evaluation process into three independent yet complementary dimensions: Alignment for verifying instruction execution, Fidelity for ensuring overall image quality and identity preservation, and Relative Expression Gain for precisely quantifying the magnitude of expression changes. Through the mutual constraint and joint assessment of these three dimensions, our protocol establishes a reliable and robust evaluation framework that significantly mitigates evaluation biases such as lazy editing and overfit editing.

Extensive experiments demonstrate that, compared to existing image evaluation metrics, FED-Score exhibits higher consistency with human subjective judgment. Meanwhile, benchmarking based on FED-Bench profoundly reveals the performance bottlenecks of current mainstream image editing models regarding fine-grained control. To further facilitate continuous research in this field, we leverage the scalability of our automated pipeline to construct a large-scale training set of 20k+ in-the-wild facial expression editing pairs. Fine-tuning an existing model on this data yields substantial improvements in both expression accuracy and fidelity, demonstrating the practical value of the proposed data construction paradigm. In summary, the main contributions of this paper are as follows:

1. **Construction of FED-Bench:** We propose the first fine-grained test benchmark specifically designed for facial expression editing. It contains 747 high-quality triplets (source image, editing instruction, Ground Truth), providing the necessary data foundation for precise evaluation.
2. **Proposal of the FED-Score Evaluation Protocol:** We design a cross-granular, decoupled evaluation system that comprehensively assesses alignment, fidelity, and relative expression gain. This effectively overcomes the systemic biases (e.g., favoring lazy editing and overfit editing) prevalent in existing metrics.

3. **In-depth Evaluation and Analysis:** We conduct comprehensive evaluations of state-of-the-art image editing models, revealing their performance bottlenecks in balancing identity preservation with precise editing.
4. **Open-sourcing a Large-scale Training Set:** We provide a high-quality, in-the-wild facial image training set containing 20k+ samples to facilitate the development of finer and more robust facial editing models in the future.

2 Related Work

2.1 Image Editing and Facial Expression Manipulation

General image editing has undergone a profound evolution from early GAN/VAE-based attribute manipulation to instruction-driven editing powered by large-scale diffusion models and autoregressive backbones. InstructPix2Pix [2] established the mainstream framework of training conditional diffusion models on synthetic paired data. AnyEdit [43] introduced task-aware routing and mixture-of-experts modules to support more diverse editing requirements. OmniGen [40] adopted a single Transformer backbone for joint text-image encoding, eliminating the need for external encoders entirely. On the multimodal language-vision collaboration front, MGIE [10] leverage MLLMs to provide precise instruction grounding signals. Step1X-Edit [26] significantly improves complex instruction following and background consistency preservation through a dual-stream bridging mechanism. ReasonEdit [42] and related works further incorporate reasoning loops into the editing pipeline. Overall, the above methods are continuously advancing toward open-ended, text-guided image editing with stronger semantic controllability and fidelity.

Facial expression manipulation presents greater challenges than general scene editing, as it demands high-precision control over local geometric structures and subtle appearance cues while strictly preserving identity and background consistency. Early text-driven methods such as StyleCLIP [28] demonstrated promising qualitative results, but tend to cause unintended identity alterations when the magnitude of instruction-driven changes is large. To address this, subsequent works introduced structured geometric conditions: MagicFace [36] conditions on action unit (AU) deltas paired with a dedicated identity encoder to achieve fine-grained expression editing; LaTo [46] encodes facial landmarks as semantic tokens to enable precise pose and expression control while preserving identity; DiffFERV [4] specializes diffusion priors to the face domain and incorporates temporal modeling to achieve high-fidelity expression editing on video sequences.

Despite achieving promising results in specific settings, these methods still rely primarily on supervised labels, action unit annotations, or geometric priors, and are not designed end-to-end for natural language expression instructions, making it difficult to capture the nuanced emotional descriptions present in various user prompts.

2.2 Image Editing Datasets and Benchmarks

The evaluation of instruction-based image editing has rapidly evolved to address the growing complexity of user intents and the demand for higher visual fidelity. To capture realistic human instructions, datasets like MagicBrush [44] provide manually annotated instruction-image pairs for real-world scenarios, ensuring high alignment with human intent. Meanwhile, HQ-Edit [15] focuses on high-resolution data collection to guarantee the visual quality of the edited outcomes. To encompass a wider range of editing semantics, AnyEdit [43] constructs a comprehensive dataset that integrates various operations, such as object addition, removal, and replacement, into a unified framework.

Beyond dataset construction, recent evaluation paradigms have shifted towards assessing holistic model capabilities on harder, real-image distributions. Modern benchmarks, such as RISEBench [47] and SmartEdit [14], have been proposed to systematically evaluate models on complex reasoning, spatial understanding, and multi-dimensional instruction following. Together, these protocols provide standardized comparisons for multi-turn edits and compositional instructions, reflecting the field's growing need for robust evaluation in open-ended editing tasks.

2.3 Evaluation Metrics for Image Editing

Evaluating AI-generated and edited images requires assessing both visual fidelity and text-image alignment. Traditional measures like the Inception Score (IS) [30] and Fréchet Inception Distance (FID) [12] are widely employed to evaluate overall realism and image quality. For measuring the semantic alignment between the edited image and the instruction or text prompt, metrics such as the CLIP Score [11] and BLIP Score [20] have become the established standards to ensure the generated content accurately reflects the textual guidance.

Unlike standard text-to-image generation, subject-driven image editing, especially facial expression manipulation, strictly demands the preservation of the source image's structural integrity and the subject's identity. To measure this perceptual and structural consistency, Learned Perceptual Image Patch Similarity (LPIPS) [45] is frequently used, while deep self-supervised feature extractors like DINO [3] and specialized face recognition models [6] evaluate semantic consistency and robust identity preservation. Recently, Large Multimodal Models (LMMs) such as GPT-4o have also emerged as powerful automatic evaluators [17], leveraging their advanced reasoning and instruction-following capabilities to comprehensively assess text-image alignment and evaluate the nuanced visual changes inherent in complex facial manipulation tasks.

3 Construction of FED-Bench

To obtain high-quality Ground Truth images for refined evaluation, we design a rigorous multi-stage screening pipeline. As shown in Fig. 2, the pipeline consists

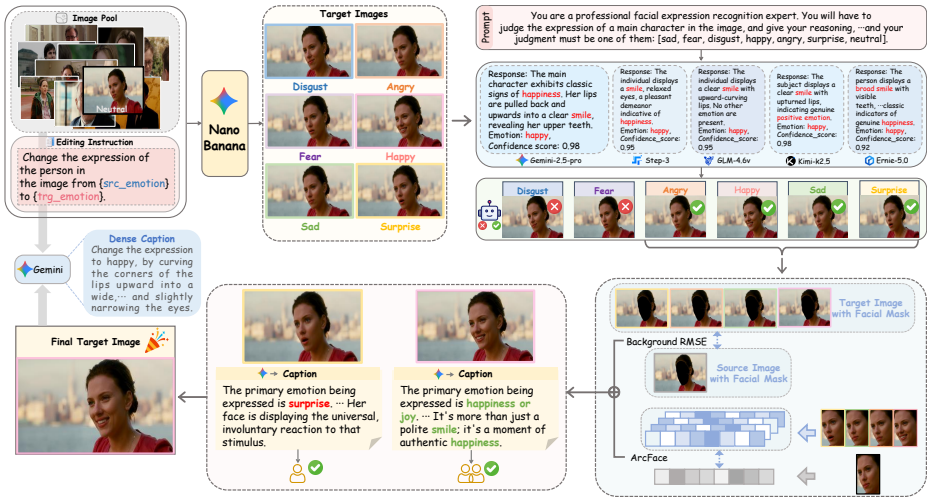

Fig. 2: The overall illustrations of FED-Bench construction pipeline.

of five stages: source data screening, candidate target image generation, coarse-grained expression recognition, fidelity ranking, and human verification.

Source Data. Facial expression editing demands source images with rich in-the-wild backgrounds and high-quality facial details, so that both background preservation and identity preservation can be rigorously tested. We select SFEW 2.0 [7–9] and DFEW [16] as raw data sources, as their movie-clip origins naturally provide diverse backgrounds and sufficient clarity. After filtering out images with wrong expression label, low resolution, severe occlusion, or extreme lighting, we curate 747 high-quality source images for FED-Bench.

Candidate target image generation. Due to the scarcity of perfectly paired in-the-wild facial expression images, we use a state-of-the-art editing model $\mathcal{G}$ (Nano Banana) to generate candidate targets. Our benchmark covers seven basic expressions: angry, disgust, fear, happy, neutral, sad, and surprise. For each source image I_{src}, we generate candidates for the remaining six expressions using the template "change the expression from {src_emotion} to {trg_emotion}":

$$I_{\text{cand}}^{(i)} = \mathcal{G}(I_{\text{src}}, T(src\_emotion, trg\_emotion(i))), \quad i = 1, 2, \ldots, 6. \tag{1}$$

Coarse-grained facial expression recognition. We first verify whether candidates successfully manifest the target expressions. Expert FER models generalize poorly in the wild (<40% accuracy, Tab. 1), so we adopt MLLMs instead. However, even the best MLLM (Gemini-2.5-Pro) achieves only 64.93% on fine-grained 7-class classification, and strict filtering at this accuracy would discard many valid samples. To address this, we propose a **Coarse-grained Expression Recognition Strategy** that regroups the seven expressions into three polarities: Positive (happy), Neutral, and Negative (remaining five), boosting peak accuracy to 81.23% (Tab. 1). We further introduce a **Multi-model Vot-**

Table 1: Comparison of facial expression recognition accuracy.

Model	Fine	Coarse
Expert Models		
BEiT [1,33]	35.7%	–
Luxand	39.8%	–
DeepFace [32]	36.5%	–
MLLMs		
Gemini-2.5-Pro	**64.93%**	81.23%
Doubao-Seed-1.6-Vision	62.52%	80.59%
ERNIE-5.0-Thinking	59.17%	80.76%
GLM-4.6V	58.77%	77.48%
Kimi-K2.5	57.83%	**81.79%**
Step-3	57.56%	79.76%
Claude-Sonnet-4.5	53.55%	76.31%
Qwen3-VL-Plus	48.59%	76.68%
Claude-Haiku-4.5	48.19%	71.75%
Grok-4.1	43.51%	66.53%

Table 2: Accuracy comparison of different voting ensemble sizes. The optimal 5-model in Fine-grained FER ensemble consists of Gemini-2.5-Pro, GLM-4.6V, Kimi-K2.5, ERNIE-5.0, and Step-3. The optimal in Coarse-grained FER ensemble consists of Gemini-2.5-Pro, GLM-4.6V, Kimi-K2.5, ERNIE-5.0, and Doubao-seed-1.6-vision.

Setting	Fine-grained (7-cls)		Coarse-grained (3-cls)	
	Best Acc	Avg Acc	Best Acc	Avg Acc
3-model	66.13	59.94	84.47	81.04
5-model	**66.67**	61.75	**84.87**	82.27
7-model	66.00	62.78	84.61	82.88
9-model	63.99	63.25	83.40	82.90

ing Mechanism using a 5-model ensemble, which achieves optimal performance (Tab. 2). This combined strategy effectively eliminates invalid samples that fail to exhibit the desired expression changes. Comprehensive test results are presented in supplementary materials.

Fidelity Ranking. In this stage, our evaluation encompasses two complementary dimensions. First, to measure identity preservation (ID), we utilize the pre-trained face recognition network ArcFace [6] to extract identity feature vectors from both the source and candidate images, subsequently calculating their cosine similarity $S_{\text{id}}^{(i)}$. Second, to assess pixel-level background consistency, we compute the Root Mean Squared Error (RMSE) across all pixels outside the facial expression region, denoted as $S_{\text{bg}}^{(i)}$. After normalizing both distance metrics, we combine them into an overall preservation score through a weighted fusion:

$$S_{\text{total}}^{(i)} = \text{Norm}(S_{\text{id}}^{(i)}) + \text{Norm}(S_{\text{bg}}^{(i)}). \tag{2}$$

Based on this score, we sort all expression-verified samples in descending order and strictly retain only the top two candidate images as potential Ground Truth references for the final stage.

Human Verification. Automated filtering cannot capture subtle facial nuances and microscopic artifacts, so we incorporate a human-in-the-loop verification as the final quality assurance step. We first use an MLLM to generate a detailed expression caption for each candidate image as an objective reference. Three human evaluators then independently inspect the candidate pairs, considering expression naturalness, instruction alignment, and identity/background preservation, with the final Ground Truth determined by majority voting.

Dense Instruction Generation. We introduce a mechanism to generate fine-grained editing instructions. This serves two primary purposes: to further refine the control dimensions of facial expression editing, and to comprehensively evaluate the instruction-following capabilities of image editing models when presented with detailed textual guidance. Specifically, we employ MLLM as a visual difference analyzer. We input the paired source image and its corresponding GT

image into the MLLM, prompting it to meticulously compare the specific facial expression alterations between the two images (e.g., the movement of facial muscles, the curvature of the eyes, and the opening or closing of the mouth). Based on this cross-image difference analysis, the MLLM generates a highly descriptive and fine-grained editing instruction for each data triplet. Compared to the simple instruction used in earlier stages, these dense instructions provide more precise guidance for generative models and offer a more challenging linguistic input for subsequent performance evaluation.

Scalability and Training Set. Our automated pipeline also exhibits excellent scalability. We further selected 20k+ in-the-wild facial images from RAF [21, 22] and DFEW, using Qwen-Image-Edit-2511 as the generator, to efficiently construct a large-scale training set. Fine-tuning FLUX-Kontext-Dev on this data yields significant improvements in both expression precision and fidelity (Sec. 5.3), validating the practical value of our data construction paradigm.

4 The FED-Score Evaluation Suite

To comprehensively and fairly evaluate facial expression editing models, it is imperative to break away from the traditional reliance on single-dimensional metrics such as standalone CLIP-Scores [11] or LPIPS [45]. Sole reliance on these isolated measurements often allows models to exploit evaluation loopholes, resulting in either lazy editing where the model fails to make necessary changes, or overfit editing where it generates exaggerated distortions. To address this, we propose the FED-Score, a decoupled evaluation protocol that deeply integrates rule-based (ID, BG, REG) computations with the model-based (PQ, SC, GTA) perceptual capabilities of MLLMs. Inspired by VIEScore [17], we establish a robust MLLM-based scoring framework for our perceptual metrics. The sub-score aggregation and normalization details, and prompt templates are provided in supplementary material. Specifically, we explicitly decompose the comprehensive evaluation into three distinct dimensions: Fidelity, Alignment, and Relative Expression Gain.

4.1 Fidelity Metrics

The Fidelity dimension is designed to assess whether the model faithfully preserves elements that should remain unchanged while maintaining high overall visual quality. Our fidelity score ($\mathcal{S}_{\text{fid}}$) is formulated by combining three distinct components: Identity Preservation (ID), Background Consistency (BG), and Perceptual Quality (PQ).

Identity Preservation (ID). The fundamental premise of facial editing is that the subject's identity must be preserved. We extract identity feature vectors from the source image I_{src} and the target image I_{trg} using a pre-trained ArcFace network, denoted as ArcFace$(\cdot)$. We then calculate their cosine similarity to serve as the ID score:

$$\text{ID}(I_{\text{src}}, I_{\text{trg}}) = \cos(\text{ArcFace}(I_{\text{src}}),\ \text{ArcFace}(I_{\text{trg}}))\,. \tag{3}$$

Background Consistency (BG). Since facial expression modifications should be strictly confined to the facial region, we first calculate the Root Mean Square Error (RMSE) of pixels exclusively in the non-facial areas, defined by a background mask M_{bg}, between I_{src} and I_{trg}. To align with the evaluation logic where a higher score indicates better performance, we formulate the BG score by applying a normalization function $\text{Norm}(\cdot)$:

$$\text{BG}(I_{\text{src}}, I_{\text{trg}}) = \text{Norm}\left(\sqrt{\frac{1}{|M_{\text{bg}}|}\sum_{(x,y)\in M_{\text{bg}}}\left(I_{\text{src}}(x,y) - I_{\text{trg}}(x,y)\right)^2}\right), \tag{4}$$

the detailed implementation of the $\text{Norm}(\cdot)$ function is provided in the supplementary material.

Perceptual Quality (PQ). Beyond physical pixel preservation, we employ an MLLM as an advanced visual evaluator to score the overall perceptual quality of I_{trg}, checking for unnatural artifacts or degradation.

By integrating these three complementary metrics, we formulate the comprehensive fidelity score $\mathcal{S}_{\text{fid}}$ as their arithmetic mean:

$$\mathcal{S}_{\text{fid}} = \text{Mean}\Big(\text{ID}(I_{\text{src}}, I_{\text{trg}}),\ \text{BG}(I_{\text{src}}, I_{\text{trg}}),\ \text{PQ}(I_{\text{trg}})\Big). \tag{5}$$

4.2 Alignment Metrics

The Alignment dimension evaluates how accurately the generated image executes the editing instructions and manifests the desired expression. It comprises two core components: Semantic Consistency (SC) and GT-based Expression Alignment (GTA).

Semantic Consistency (SC). We prompt the MLLM to act as a rigorous judge to evaluate the semantic matching degree between the generated target image I_{trg} and the input text instruction T.

GT-based Expression Alignment (GTA). Benefiting from the precise reference images provided by our FED-Bench, we transcend text-only evaluation. We leverage the MLLM to directly compare I_{trg} with the reference image I_{gt}, specifically scoring the similarity of their facial expressions. This image-to-image semantic supervision significantly enhances the reliability of the alignment assessment.

Similarly, we integrate these two dimensions to formulate the comprehensive alignment score $\mathcal{S}_{\text{align}}$, defined as their arithmetic mean:

$$\mathcal{S}_{\text{align}} = \text{Mean}\Big(\text{SC}(I_{\text{trg}}, T),\ \text{GTA}(I_{\text{trg}}, I_{\text{gt}})\Big). \tag{6}$$

4.3 Relative Expression Gain

The primary vulnerability of existing image editing evaluation protocol is their susceptibility to lazy editing. If a model simply outputs the source image without

any modification, it achieves perfect fidelity scores (e.g., ID and BG metrics). To fundamentally resolve this systemic flaw, we introduce the Relative Expression Gain (REG), a novel quantitative metric designed to measure the actual magnitude of expression change relative to the expected change provided by our benchmark.

We first compute the perceptual distance of the facial region between the generated target image and the source image using LPIPS [45], which represents the model's actual expression alteration. We then normalize this distance by the optimal perceptual change derived from our precise Ground Truth. To penalize both lazy editing and overfit editing, we formulate the final REG score $\mathcal{S}_{\text{reg}}$ as a Gaussian-like penalty function centered at 1.0:

$$\text{REG} = \frac{\text{LPIPS}_{\text{face}}(I_{\text{src}},\ I_{\text{trg}})}{\text{LPIPS}_{\text{face}}(I_{\text{src}},\ I_{\text{gt}})}, \quad \mathcal{S}_{\text{reg}} = \exp\left(-\frac{(\text{REG}-1)^2}{2\sigma^2}\right), \tag{7}$$

where σ is a hyperparameter controlling the tolerance width of the penalty, which is empirically set to 0.5 in our evaluation. In this formulation, the model achieves the maximum score of 1.0 *only* when its expression alteration magnitude perfectly matches that of the Ground Truth. Consequently, this metric naturally penalizes both lazy editing and overfit editing. We emphasize that REG measures the *perceptual* magnitude of facial-region change as a proxy for expression-change intensity, rather than a semantic action-unit measure; the LPIPS ratio may therefore also absorb texture and lighting variations. We validate this proxy against AU–based measures in the supplementary material.

4.4 The Comprehensive FED-Score

Having established three orthogonal evaluation dimensions, we integrate them into a unified metric—the **FED-Score**—by computing their product:

$$\text{FED-Score} = \mathcal{S}_{\text{fid}} \times \mathcal{S}_{\text{align}} \times \mathcal{S}_{\text{reg}}. \tag{8}$$

This multiplicative formulation ensures that a near-zero value in any single dimension is sufficient to suppress the entire FED-Score, so that all three dimensions must be simultaneously satisfied to achieve a high overall score. We empirically validate that the FED-Score aligns significantly better with human perceptual preferences than conventional metrics in Sec. 5.2.

5 Experiments

5.1 Experimental Setup

Human Study Protocol. To validate metric reliability, we conduct a human correlation study using a Two-Alternative Forced Choice (2AFC) protocol. We randomly sample 2,760 image pairs, where each pair consists of editing results from two different models given the same source image and instruction. Three

independent annotators are asked to select the better image within each pair from three perspectives: identity preservation, expression change magnitude, and overall quality. The final human preference is determined by majority voting, and we report the accuracy between each metric's algorithmic preference and the human consensus.

Evaluated Models. To comprehensively assess the current landscape of facial expression editing, we benchmark 18 image editing models, including our fine-tuned **FLUX-Kontext-FED**, which is obtained by fine-tuning FLUX-Kontext-Dev on our proposed 20k+ training set to validate the effectiveness of the curated training data. All models are evaluated using their officially recommended default inference configurations to ensure a fair and reproducible comparison.

Evaluation Metrics. We evaluate all models using our proposed FED-Score and its constituent sub-metrics. To demonstrate the superiority of FED-Score over conventional approaches, we also compare against widely-used baseline metrics, including L2 distance, LPIPS [45], CLIP-I, DINO Score, CLIP Score [11], and ArcFace cosine similarity [6].

Evaluation Protocol. All experiments are conducted on the full FED-Bench test set without subset splitting. Each model is evaluated under both *simple instructions* (*e.g.*, "change the expression from neutral to happy") and *dense instructions* that describe fine-grained facial muscle movements, enabling a thorough assessment of instruction-following capabilities at different granularities.

Implementation Details. For the MLLM-based scoring components (PQ, SC, and GTA), we uniformly adopt Gemini-2.5-Pro as the evaluation backbone to ensure consistency across all perceptual assessments. Identity Preservation is computed using ArcFace-R100 for face embedding extraction. The facial region mask M_{bg} required for Background Consistency is obtained via Grounded SAM 2 [25, 29], which provides precise segmentation of the face region. For the REG metric, we employ LPIPS with a VGG-16 backbone, where the facial crop is directly obtained from the bounding box predicted by Grounding DINO.

5.2 Human Alignment Test

Identity Preservation. As shown in Tab. 3a, $\text{ArcFace}_{\text{F}}$ achieves the highest agreement with human judgment among all face-region baselines, followed closely by DINO_{F} and LPIPS_{F}. While all metrics surpass the random baseline of 0.5, their moderate absolute values (0.54–0.66) indicate that assessing identity preservation from face crops alone remains inherently challenging. Notably, CLIP-T_{F} barely exceeds chance level, confirming that text-image similarity metrics are fundamentally unsuitable for identity evaluation. These results validate our choice of ArcFace cosine similarity as the ID component in FED-Score.

Relative Expression Gain. As shown in Tab. 3b, our proposed $\text{LPIPS-Ratio}_{\text{F}}$ achieves the highest accuracy (0.7386) for evaluating expression change magnitude, outperforming all GT-based baselines. Perceptual similarity measures ($\text{LPIPS}_{\text{F,GT}}$, $\text{DINO}_{\text{F,GT}}$) perform reasonably well, as they are inherently sensitive to fine-grained visual differences. $\text{RPM-Modify}_{\text{F}}$ [23], another ratio-based metric, achieves moderate accuracy (0.6154), falling between the perceptual

Table 3: Human alignment test. We report the accuracy between each metric's preference and the human consensus under a 2AFC protocol. $_{F}$: computed on the face region; $_{GT}$: compared with the ground truth image.

(a) ID Fidelity

Metrics	Acc.
$L2_{F}$	0.6446
$DINO_{F}$	0.6596
$LPIPS_{F}$	0.6528
CLIP-I_{F}	0.6384
CLIP-T_{F}	0.5420
$ArcFace_{F}$	**0.6606**

(b) REG

Metrics	Acc.
$L2_{F,GT}$	0.6047
$DINO_{F,GT}$	0.6818
$LPIPS_{F,GT}$	0.7045
CLIP-$I_{F,GT}$	0.5909
$ArcFace_{F,GT}$	0.5455
$RPM\text{-}Modify_{F}$	0.6154
LPIPS-$Ratio_{F}$	**0.7386**

(c) Overall Score

Metrics	Acc.
L2	0.5711
$DINO_{GT}$	0.6216
DINO	0.5816
$LPIPS_{GT}$	0.6099
LPIPS	0.5384
CLIP-I_{GT}	0.6850
CLIP-I	0.6325
CLIP-T	0.4480
FED-Score	**0.7700**

(d) FED-Score Abl.

Metrics	Acc.
w/o Rule	0.7422
w/o REG	0.7577
w/o Fidelity	0.7379
w/o Alignment	0.7279
w/o Model	0.7202
FED-Score	**0.7700**

baselines and identity-oriented metrics. In contrast, identity-oriented metrics like $ArcFace_{F,GT}$ and CLIP-$I_{F,GT}$ drop near random chance, suggesting that their feature spaces are not designed to capture expression change magnitude. The superiority of LPIPS-$Ratio_{F}$ stems from its ratio-based formulation: by normalizing the actual expression change against the expected GT change, it directly quantifies relative editing effort and naturally penalizes both lazy editing (ratio $\ll 1$) and overfit editing (ratio $\gg 1$).

Overall Score. When evaluating holistic editing quality (Tab. 3c), FED-Score achieves 0.77 agreement with human consensus, substantially outperforming the best single baseline CLIP-I_{GT}. Across all baselines, GT-based variants consistently surpass their source-target counterparts, reinforcing the value of reference-guided evaluation. Strikingly, CLIP-T falls below random chance, indicating that standalone text-image similarity is fundamentally unreliable for this task. These results confirm that the multi-dimensional design of FED-Score captures complementary evaluation aspects that no single metric can address alone.

Ablation Study. We conduct an ablation study on FED-Score (Tab. 3d). Removing any single dimension consistently degrades performance, confirming that all components are essential. The model-based MLLM metrics prove most critical: removing them (w/o Model) causes the largest drop, as perceptual quality and semantic consistency cannot be captured by rule-based computation alone. Removing alignment and fidelity also leads to substantial degradation. Rule-based metrics likewise complement the MLLM components, showing that objective measurements of identity and background preservation remain valuable even alongside powerful vision-language models. REG exhibits the smallest individual impact, which is expected since it specifically targets expression change magnitude—a dimension less relevant in identity-focused and quality-focused evaluation pairs. Overall, this ablation validates that the synergy between rule-based computation and model-based perception is the key to FED-Score's superior human alignment.

Table 4: Benchmarking results on FED-Bench. Left: Dense instructions (ranked by Dense FED-Score). Right: Simple instructions (ranked by Simple FED-Score). We report raw mean values for sub-metrics and normalized FED-Score. BG↓: lower is better. REG: optimal at 1.0. **Bold**: best; underline: second best.

Dense Instructions								Simple Instructions							
Method	ID	BG	PQ	SC	GTA	REG	Score	Method	ID	BG	PQ	SC	GTA	REG	Score
Qwen-image-edit-plus	.58	17.5	9.7	8.8	**5.7**	1.18	**.469**	Qwen-image-edit-plus	.63	16.8	9.8	8.8	**5.8**	1.13	**.492**
SeedDream 4.0 [31]	.62	15.5	9.5	9.1	4.3	1.37	.379	SeedDream 4.0	.69	16.4	9.7	8.0	5.0	1.26	.413
FLUX.2 Pro [18]	.58	13.9	9.8	9.0	4.0	1.37	.377	FLUX.2 Pro	.58	14.0	9.6	8.8	5.2	1.37	.400
Qwen-image-edit	.45	19.6	9.3	8.9	4.2	1.37	.337	Qwen-image-edit-2511	.44	15.3	9.8	**9.4**	4.0	1.42	.361
FLUX-Kontext-FED	.68	7.3	9.7	6.2	3.4	0.95	.332	Qwen-image-edit	.43	19.9	9.6	8.7	4.4	1.37	.343
FLUX-Kontext-Pro	.52	7.3	9.8	7.6	3.5	**0.99**	.327	Step1X v1p2	.52	17.8	9.7	9.3	3.1	1.41	.333
FLUX-Kontext-Max	.50	9.7	9.7	7.7	3.5	1.07	.320	FLUX-Kontext-FED	.68	7.8	9.7	6.5	2.9	0.96	.325
Qwen-image-edit-2511 [37]	.46	15.8	9.5	**9.3**	3.4	1.43	.317	FLUX-Kontext-Max	.49	8.1	**9.9**	5.2	3.7	**1.03**	.259
Step1X v1p2	.56	17.1	9.4	7.7	3.0	1.31	.303	SeedEdit 3.0	.48	12.8	9.1	7.6	2.7	1.56	.239
FLUX-Kontext-Dev [19]	.72	23.8	9.6	6.7	3.0	0.67	.243	FLUX-Kontext-Pro	.55	8.5	**9.9**	4.8	3.4	**0.97**	.227
SeedEdit 3.0 [34]	.49	11.6	7.8	7.4	1.9	1.55	.203	Bagel	.53	13.1	5.4	4.1	2.4	1.29	.163
UniWorld-v2 [24]	.37	31.6	9.1	7.6	2.5	1.45	.201	Step1X	.32	17.3	9.7	5.6	2.3	1.67	.149
DreamOmni2 [39]	**.81**	24.1	9.6	5.0	2.6	0.45	.168	OmniGen2	.44	77.6	8.3	4.1	3.6	1.61	.122
OmniGen2 [38]	.56	63.7	7.3	6.3	2.6	1.52	.155	FLUX-Kontext-Dev	**.86**	**4.8**	9.8	2.3	3.1	0.35	.120
FLUX-Kontext-Fill	.11	**5.1**	**9.9**	5.5	1.7	1.56	.155	UniWorld-v2	.30	34.9	8.6	3.4	2.7	1.60	.110
Step1X [26]	.33	16.3	7.3	6.9	1.3	1.72	.127	FLUX-Kontext-Fill	.11	5.1	9.7	3.3	1.1	1.52	.096
Bagel [5]	.42	47.2	6.5	6.8	1.8	1.57	.115	DreamOmni2	.45	36.9	8.9	1.2	2.0	1.40	.049
InstructPix2Pix [2]	.08	47.0	0.0	0.0	0.0	2.56	.001	InstructPix2Pix	.08	41.2	0.2	0.1	0.1	2.42	.004

5.3 Benchmarking SOTA Models

Overall Rankings. Tab. 4 presents the comprehensive benchmarking results across 18 models. Qwen-Image-Edit-Plus consistently ranks first under both Dense and Simple instructions, achieving the highest GTA by a significant margin while maintaining strong performance across all other dimensions. SeedDream 4.0 and FLUX.2 Pro secure the second and third positions in both settings, with SeedDream 4.0 exhibiting particularly strong ID preservation and FLUX.2 Pro delivering well-balanced sub-metric scores. Notably, the top-3 ranking remains stable across instruction types, suggesting that these models possess genuine comprehensive editing proficiency rather than excelling at a particular instruction granularity. It is also worth highlighting that FLUX-Kontext-FED, fine-tuned on our proposed training set, ranks 5th under Dense instructions and 7th under Simple instructions, substantially outperforming the original FLUX-Kontext-Dev, which validates the practical value of our released training data.

Fidelity vs. Alignment Trade-off. A key insight revealed by FED-Score is that no single model excels across all dimensions simultaneously. FLUX-Kontext-Dev and DreamOmni2 achieve the highest ID scores, yet their extremely low REG exposes a classic *lazy editing* pattern: preserving identity by barely modifying the expression. On the opposite extreme, InstructPix2Pix exhibits severe *overfit editing* with REG>2.5 while catastrophically degrading visual quality (PQ≈0). FLUX-Kontext-Fill presents yet another failure mode—best BG and PQ but worst ID, generating high-quality images that completely lose the subject's identity. Under conventional single-metric evaluation, lazy editors would be incorrectly ranked at the top based on ID or BG alone. FED-Score effectively penalizes such dimensional imbalances, ensuring that only models with genuinely balanced performance receive high overall scores.

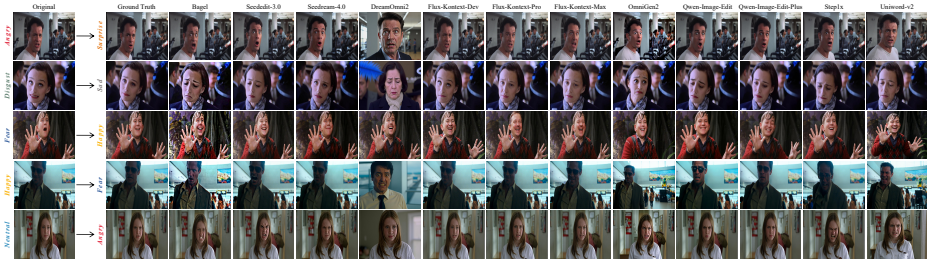

Fig. 3: The qualitative comparison of facial expression editing on our FED-Bench. Each row shows a different editing task with the original image, ground truth target, and results from part of evaluated models. From left to right: original image, ground truth, followed by the evaluation results.

Dense vs. Simple Instructions. The ranking shifts between the two settings expose notable differences in instruction-following capability. FLUX-Kontext-Pro drops significantly from Dense to Simple, with its SC nearly halving, suggesting it benefits disproportionately from the richer guidance in dense instructions. Conversely, FLUX-Kontext-Dev's lazy editing tendency intensifies under simple instructions, as its REG further deteriorates. The Alignment metrics (SC, GTA) exhibit the largest cross-setting variance, confirming that fine-grained instruction following remains the primary bottleneck for current models.

Qualitative Results. Fig. 3 presents qualitative comparisons that corroborate the quantitative findings. The FLUX-Kontext series produces high-fidelity outputs with well-preserved backgrounds and identity, yet their expression transformations remain subtle—a visual manifestation of the *lazy editing* pattern identified in our quantitative analysis. In contrast, Qwen-Image-Edit-Plus generates clear and natural target emotions closely aligned with the ground truth, consistent with its top-ranked FED-Score. Models such as DreamOmni [39] and OmniGen2 [38] introduce significant non-expression modifications including background alterations and lighting changes, visually confirming the fidelity-alignment trade-off revealed by the benchmarking results. Furthermore, based on qualitative analysis and REG experimental data, most models exhibit overfit editing. Detailed qualitative analysis results are presented in the supplementary materials.

6 Conclusion and Discussions

We present FED-Bench, a comprehensive benchmark for facial expression image editing. We construct 747 high-quality triplets via a scalable multi-stage pipeline, and propose FED-Score, a cross-granularity protocol that disentangles evaluation into Fidelity, Alignment, and Relative Expression Gain, effectively penalizing lazy editing and overfit editing. Human alignment experiments confirm that FED-Score substantially outperforms all traditional metrics. Benchmarking 18 image editing models and revealing that balancing fidelity with accurate

expression manipulation remains challenging, with fine-grained instruction following as the primary bottleneck. We additionally release a 20k+ in-the-wild training set and validate its utility by fine-tuning FLUX-Kontext-Dev, yielding notable gains in both expression accuracy and fidelity. One limitation is that the reference images anchoring our evaluation are *pseudo* ground truth—synthesized rather than captured from the same subject—so each reflects one plausible target rather than a unique answer. Nonetheless, our robustness analysis in the supplementary material shows that model rankings stay stable when the reference generator and the MLLM judge are swapped. Collecting truly paired expression data remains future work.

Acknowledgement

This work was supported in part by the CAAI-Tencent Rhino-Bird Open Research Fund, the Xianyang Major Scientific and Technological Achievements Transformation Special Project (Grant No. L2024-ZDKJ-ZDCGZH-0012), the Key Research and Development Program of Shaanxi (Grant No. 2024GX-YBXM-533), the Central Government Guiding Local Science and Technology Development Fund (Grant No. 2026ZY04001), and Sichuan Science and Technology Program (Grant No. 2025ZNSFSC0468).

Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies. Besides, the authors would like to thank Kunchang Li (ByteDance Seed) for his constructive insights and assistance.

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. CoRR **abs/2106.08254** (2021), `https://arxiv.org/abs/2106.08254`
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chen, X., Xue, H., Song, L.: Diffferv: Diffusion-based facial editing of real videos. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 819–827 (2025)
5. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
7. Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In: Proceedings of the 16th international conference on multimodal interaction. pp. 461–466 (2014)

8. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops). pp. 2106–2112. IEEE (2011)
9. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. IEEE MultiMedia **19**(3), 34–41 (2012). https://doi.org/10.1109/MMUL.2012.26
10. Fu, T.J., Hu, W., Du, X., Wang, W.Y., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. arXiv preprint arXiv:2309.17102 (2023)
11. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7514–7528. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.595, https://aclanthology.org/2021.emnlp-main.595/
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8362–8371 (2024)
15. Hui, M., Yang, S., Zhao, B., Shi, Y., Wang, H., Wang, P., Zhou, Y., Xie, C.: Hq-edit: A high-quality dataset for instruction-based image editing. arXiv preprint arXiv:2404.09990 (2024)
16. Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J.: Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2881–2889 (2020)
17. Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: Viescore: Towards explainable metrics for conditional image synthesis evaluation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12268–12290 (2024)
18. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025)
19. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)

21. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing **28**(1), 356–370 (2019)
22. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2584–2593. IEEE (2017)
23. Li, Z., Xu, Z., Peng, Y., Liu, Y.: Balancing preservation and modification: A region and semantic aware metric for instruction-based image editing (2025), https://arxiv.org/abs/2506.13827
24. Li, Z., Liu, Z., Zhang, Q., Lin, B., Wu, F., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., Wang, S., Cheng, X., Yuan, L.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback (2025), https://arxiv.org/abs/2510.16888
25. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
26. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
27. Ma, Y., Ji, J., Ye, K., Lin, W., Wang, Z., Zheng, Y., Zhou, Q., Sun, X., Ji, R.: I2ebench: A comprehensive benchmark for instruction-based image editing. Advances in Neural Information Processing Systems **37**, 41494–41516 (2024)
28. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2085–2094 (2021)
29. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024), https://arxiv.org/abs/2408.00714
30. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 2234–2242. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
31. Seedream, T., :, Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., Jian, X., Kuang, H., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., Liu, W., Lu, Y., Luo, Z., Ou, T., Shi, G., Shi, Y., Sun, S., Tian, Y., Tian, Z., Wang, P., Wang, R., Wang, X., Wang, Y., Wu, G., Wu, J., Wu, W., Wu, Y., Xia, X., Xiao, X., Xu, S., Yan, X., Yang, C., Yang, J., Zhai, Z., Zhang, C., Zhang, H., Zhang, Q., Zhang, X., Zhang, Y., Zhao, S., Zhao, W., Zhu, W.: Seedream 4.0: Toward next-generation multimodal image generation (2025), https://arxiv.org/abs/2509.20427
32. Serengil, S.I., Ozpinar, A.: Boosted lightface: A hybrid dnn and gbm model for boosted facial recognition. Gazi University Journal of Science **39**(1), 452–466 (2026). https://doi.org/10.35378/gujs.1794891, https://dergipark.org.tr/en/pub/gujs/article/1794891
33. Tanneru: Beit-large fine-tuned on affectnet for emotion detection. https://hf-mirror.com/Tanneru/Facial-Emotion-Detection-FER-RAFDB-AffectNet-BEIT-Large (2025)

34. Wang, P., Shi, Y., Lian, X., Zhai, Z., Xia, X., Xiao, X., Huang, W., Yang, J.: Seededit 3.0: Fast and high-quality generative image editing (2025), https://arxiv.org/abs/2506.05083
35. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., 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)
36. Wang, Y., Zhang, W., Jin, C.: Magicface: Training-free universal-style human image customized synthesis. arXiv preprint arXiv:2408.07433 (2024)
37. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
38. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
39. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., Tan, H., Wu, S., Wang, C., Wang, Y., et al.: Dreamomni2: Multimodal instruction-based editing and generation. arXiv preprint arXiv:2510.06679 (2025)
40. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13294–13304 (2025)
41. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
42. Yin, F., Liu, S., Han, Y., Wang, Z., Xing, P., Wang, R., Cheng, W., Wang, Y., Li, A., Yin, Z., et al.: Reasonedit: Towards reasoning-enhanced image editing models. arXiv preprint arXiv:2511.22625 (2025)
43. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
44. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)
46. Zhang, Z., Zhang, Z., Liao, J., Meng, X., Hu, Q., Zhu, S., Zhang, X., Qin, L., Wang, W.: Lato: Landmark-tokenized diffusion transformer for fine-grained human face editing (2026), https://arxiv.org/abs/2509.25731
47. Zhao, X., Zhang, P., Tang, K., Zhu, X., Li, H., Chai, W., Zhang, Z., Xia, R., Zhai, G., Yan, J., et al.: Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. arXiv preprint arXiv:2504.02826 (2025)