This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

F-Bench: Rethinking Human Preference Evaluation Metrics for Benchmarking Face Generation, Customization, and Restoration

Lu Liu[1,*], Huiyu Duan[1,*], Qiang Hu[1,†], Liu Yang[1], Chunlei Cai[2],
Tianxiao Ye[2], Huayu Liu[1], Xiaoyun Zhang[1], Guangtao Zhai[1]
[1]Shanghai Jiao Tong University, Shanghai, China [2]Bilibili Inc.,Shanghai, China

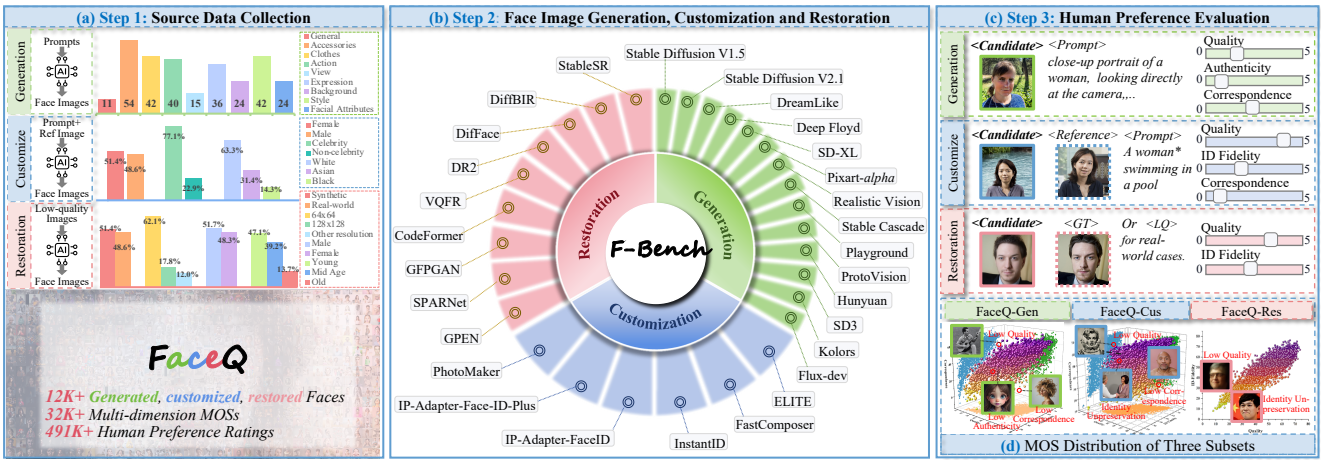


Figure 1. **Data construction pipeline of *F-bench* and content overview of *FaceQ* database**. From *left* to *right*, **(a)** Diverse input source data for face generation, customization, and restoration. **(b)** ***F-Bench*** benchmarks 29 face generative models, including 14 face generation models, 6 face customization models, and 9 face restoration models. **(c)** Multi-dimensional subjective preference evaluation in ***F-Bench***. **(d)** 3D scatter plots of MOS distributions across FaceQ three subsets.

Abstract

*Recent artificial intelligence (AI) generative models have demonstrated remarkable capabilities in image production, and have been widely applied to face image generation, customization, and restoration. However, many AI-generated faces (AIGFs) still suffer from issues such as unique distortions, unrealistic details, and unexpected identity shifts, underscoring the need for a comprehensive quality evaluation method for AIGFs. To this end, we introduce **FaceQ**, the first comprehensive AI-generated **Face** image database with fine-grained **Q**uality annotations aligned with human preferences, which consists of 12K images and 491K ratings across multiple dimensions. Using the FaceQ database, we establish **F-Bench**, a benchmark for comparing and evaluating face generation, customization, and restoration models, highlighting strengths and weaknesses across various prompts and evaluation dimensions. Additionally, we assess the performance of existing image quality assessment (IQA) methods on FaceQ, and further propose a large multimodal model (LMM) based **F**ace quality **Eval**uator (**F-Eval**) to accurately assess the multi-dimensional quality of generated faces in a one-for-all manner. Extensive experimental results demonstrate the state-of-the-art performance of our F-Eval. The project page is: https://mediax-sjtu.github.io/F-Bench/.*

1. Introduction

Generative models, including VAE [47], GAN [26], and diffusion models [33, 74, 80], have made remarkable progress in recent years, and have been extended to the fields of text-to-image generation [3], text-driven image editing [44], image restoration [99], *etc*, leading to high-quality artificial intelligence generated images (AIGIs). As an important aspect of AIGIs, many studies have also investigated face generation [100], face customization [57, 108], and face restoration [62, 87], producing creative, high-quality face images. Despite these advances, many AI-generated faces (AIGFs) still suffer from perceptual deficiencies.

Many natural face image quality assessment databases have been established in the literature [48, 51, 59, 86]. However, natural images are mainly degraded by distor-

* Equal contribution. † Corresponding author.

Table 1. Comparison between **FaceQ** and existing AIGC quality assessment databases *(top)* as well as natural face quality assessment databases *(bottom)*. FaceQ is the first large-scale AI-generated, customized and restored face quality assessment database.

Database	Domain	Source	Images	Scores	Dimensions
AGIQA-3K (*TCSVT2024*) [51]	General	AI-generated	2,982	MOS	Perception, Alignment
AIGCIQA (*CVPR2024*) [86]	General	AI-generated	2,400	MOS	Quality, Authenticity, Correspondence
RichHF-18K (*CVPR2024*) [59]	General	AI-generated	17,760	MOS	Plausibility, Aesthetics, Text-image Alignment, Overall
PIQ23 (*CVPR2023*) [13]	Face	Real	5,116	Pair	Overall, Details, Exposure
FaceQ	**Face**	**AI-generated, customized and restored**	**12,255**	**MOS**	**Quality, Authenticity, ID Fidelity, Correspondence**

tions such as noise, blur, compression [23, 66, 110], *etc.*, while AIGIs generally suffer from unique distortions including unrealistic structures and text-image inconsistencies [59, 86], *etc*. Recently, many studies have investigated human preferences for AIGIs [39, 50, 59, 86, 103]. However, these general AIGI quality assessment works may not be well aligned with human preferences on AIGFs, since the perceptual quality of AIGFs can be significantly affected by small deviations in facial attributes,(*e.g.*, eye size, skin texture, identity shifts), while the number of AIGFs in these AIGI databases is limited. Moreover, current AIGI quality assessment works mainly focus on the image generation, while human preferences for face customization have rarely been considered, where identity fidelity is significant.

The above limitation highlights the necessity of establishing a comprehensive quality evaluator for AIGFs. To this end, we propose **FaceQ**, the first AI-generated **Face** image **Q**uality assessment database with three popular tasks. Specifically, FaceQ contains 12K+ images obtained from 29 generative models across three tasks: 1) face generation, 2) face customization, and 3) face restoration, as shown in Fig.1. Based on the generated images, we collect 491K+ quality ratings from multiple dimensions including quality, authenticity, identity (ID) fidelity, and text-image correspondence, respectively, finally forming 32K+ mean opinion scores (MOSs). Using FaceQ, we establish a human preference benchmark, *i.e.*, **F-Bench**, for face generation, customization, and restoration models, and thoroughly analyze the generation performance across different evaluation perspectives, models, and categories. In addition, we also compare the correlation of existing image quality assessment (IQA), face quality assessment (FQA), AIGI quality assessment, and preference evaluation metrics with human preferences.

Furthermore, we present **F-Eval**, the first generated **F**ace image quality **Eval**uation method designed to perform comprehensive multi-dimensional quality evaluation for generated, customized, and restorted faces in a **one-for-all** manner. Specifically, F-Eval leverages large multimodal models (LMMs) [11] and adopts instruction tuning [63] for preference-related vision-language adaption. We also apply low-rank adaptation (LoRA) [35] techniques for more efficient feature refining. Extensive experimental results demonstrate that F-Eval outperforms existing quality assessment methods in terms of multiple dimensions relevant to human preference for all three tasks. The main contributions of our work are highlighted as follows:

- We construct **FaceQ**, the first generated face quality assessment database consisting 491K+ human preference ratings from quality, authenticity, ID fidelity, correspondence dimensions, on 12K+ generative face images across three tasks.
- We establish **F-Bench**, a comprehensive human preference benchmark for 29 face generation, customization, and restoration methods. It also extensively evaluates the alignment of 26 existing quality assessment methods with human preferences.
- We present **F-Eval**, a LMM-based quality assessment model to predict multi-dimensional human preference on generated faces. Extensive experiments demonstrate the state-of-art QA performance of the proposed F-Eval.

2. Related Work

Face Generation & Customization & Restoration. **(1)** For face generation models, earlier works such as GANs [26, 43, 85] have demonstrated significant influences. Recently, diffusion models [73, 74, 77] have achieved rapid developments in text-to-image generation. These models can be categorized into pixel-space diffusion models [1, 72, 73, 76], and latent space diffusion models [17, 58, 71, 74]. **(2)** For face customization, this is a recently defined task that aims to create images of a specific person with the input of an identity reference image and prompts [98]. It can be divided into two categories including conventional test-time optimization [24, 49, 75] and tuning-free customization methods utilizing pre-trained diffusion models [20, 30, 57, 91, 101, 108]. **(3)** For face restoration, which aims to restore high-quality face images from their low-quality counterparts. Some traditional restoration methods utilize geometric priors [16, 19, 46] or reference priors [54, 55] while in recent years, exploring generative prior such as GAN priors [15, 28, 94, 107, 113] and diffusion priors [36, 37, 62, 84, 87, 97, 109] has become a prevalent trend. Despite these improvements, the generated, customized, and restored images may suffer from various quality issues, highlighting the necessity of evaluators.
Face Image Quality Assessment Database. Existing face image quality assessment (FIQA) can be divided into Biometric FIQA [12, 31, 60, 69] and Generic FIQA [13, 18, 82, 92]. BFIQA methods are most developed to evaluate the biometric utility for face recognization. GFIQA meth-

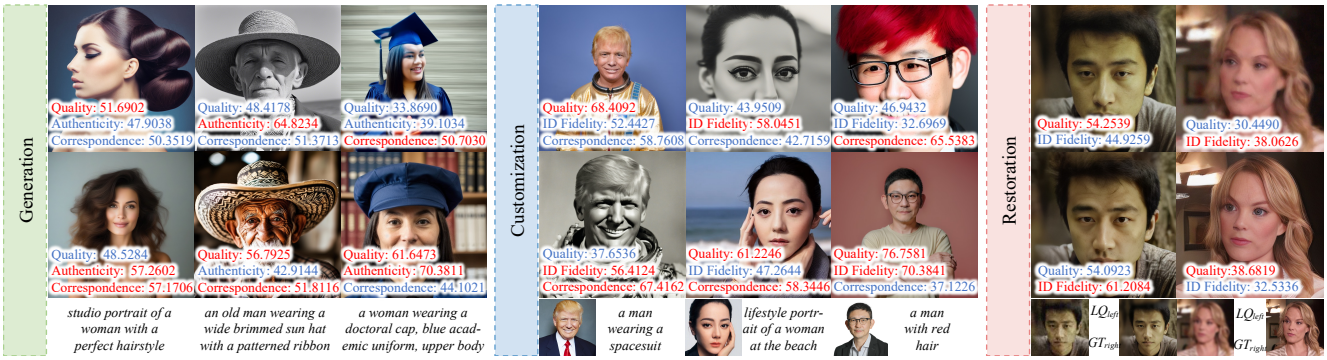


Figure 2. **Rating comparisons of eight dimensions.** Each column presents a pair of intuitive examples of each dimension, with red indicating the better rating and blue indicating the worse one. From *left* to *right*, the subsets are face generation, face customization, and face restoration subsets. The last row displays the corresponding prompts, reference image-prompt pairs, and the GT-LQ image pairs.

ods [82] focus on perceptual degradation in face restoration task. Both of these databases consist of natural face images, which has distinctly different nature with generated faces.

AIGC Quality Assessment Database. In recent years, plenty of researchers have raised interest in assessing AI-generated images(AIGIs) quality assessment [23, 40, 48, 51, 86, 104] by building various databases as shown in Tab 1. All of these existing database contain various categories of general AIGI images, leaving a notable gap in datasets dedicated to AI-generated face images.

3. *FaceQ* Database Construction

In this section, we introduce the proposed generative face quality assessment database **FaceQ**. The database includes 12,255 images and 491,130 human opinion ratings. Our FaceQ demonstrates the merits of comprehensive generated visual content and multi-dimensional labeling, which can significantly facilitate the advancement of AIGC research.

3.1. Data Collection

3.1.1. *FaceQ-Gen* Subset Collection

Prompt Sources. The prompts in *FaceQ-Gen* subset are sourced from MS-COCO [61] [57] and GPT-4o [10]. These face-centric prompts can be categorized into nine types: *General, Accessories, Clothes, Action, View, Expression, Background, Style*, and *Facial Attributes*.

Models Collection. We collect 14 representative face generation models, including Deep Floyd [1], Stable Diffusion V1.5 [74], Stable Diffusion V2.1 [74], SD-XL [71], Realistic Vision [8], Stable Cascade [70], ProtoVision V6.6 [7], SD3 [9], Playground V2.5 [52], Kolors [6], PixArt-alpha [17], Hunyuan [58], and Flux-dev [3]. Negative prompts and safe sensors are enabled. In total, *FaceQ-Gen* includes 4,032 generated face images.

3.1.2. *FaceQ-Cus* Subset Collection

Identity and Prompt Sources. Face customization relies on a reference identity image and a guiding prompt. The reference images in *FaceQ-Cus* are partially self-collected, with the rest sourced from [57]. The prompts are selected and filtered from the *FaceQ-Gen* prompt list.

Models Collection. We collect six representative face customization methods, including ELITE [98], FastComposer [101], IP-Adapter-FaceID [4], InstantID [91], IP-Adapter-Face-ID-Plus [5], and PhotoMaker V2 [57]. Here, we use single reference image as input. In total, *FaceQ-Cus* includes 4,200 generated face images.

3.1.3. *FaceQ-Res* Subset Collection

Low Quality Images Source. **(1) Synthetic**: We construct two degraded pipelines following previous works [15, 93, 95, 97] to simulate real-world degradations on both in-domain data from VFHQ [90], CelebRef [56] and out-of-domain self-collected high-quality Asian face images **(2) Real-world**: Real-world low-quality images are sourced from various public datasets, including AgeAB [68], CelebChild [28], Wider [105], LFW [38] and MegaFace [45], ensuring the degradation diversity.

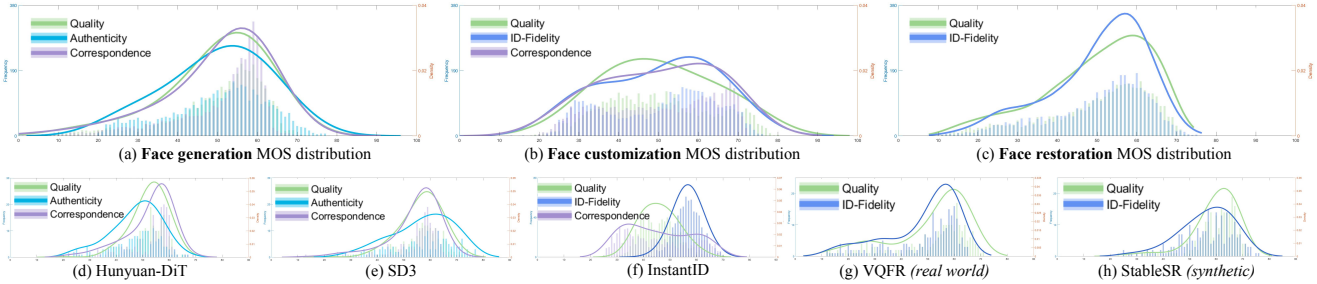
Models Collection and Face Image Restoration. We evaluate both diffusion-prior-based methods such as DiffBIR [62], DifFace [109], StableSR [87], and GAN or Transformer-based methods GFPGAN [94], GPEN [107], CodeFormer [113], VQFR [28] and Face-SPARNet [15]. Each method is benchmarked on both the *F-Res* synthetic dataset and the real-world dataset. In total,*FaceQ-Res* include **4023** restored images, with 2007 images in the real-world case and 2016 images in the synthetic case.

3.2. Human Preference Dimension Design

The evaluation dimensions of generated faces need specific design to adequately reflect human preference, given the complexity of perceptual factors involved. To the end, we decompose "quality" into four core dimensions: *Quality*, *Authenticity*, *ID Fidelity*, and *Correspondence*.

***Quality*.** This dimension evaluates the overall perceptual quality of the image, considering factors such as color accuracy, blur, noise, and artifacts. It is used in all tasks.

***Authenticity*.** This dimension evaluates how closely the



(a) **Face generation** MOS distribution (b) **Face customization** MOS distribution (c) **Face restoration** MOS distribution

(d) Hunyuan-DiT (e) SD3 (f) InstantID (g) VQFR *(real world)* (h) StableSR *(synthetic)*

Figure 3. **Mean Opinion Score (MOS) distribution histograms and kernel density curves.** (a)-(c) MOS distributions for three subsets. (d)-(h) Model-wise MOS distributions. MOS Distributions of other methods are provided in the *supplementary material*.

generated image resembles a natural, real-life photograph, with particular emphasis on realistic skin texture, facial details, and wrinkles. It is used in face generation task.

***ID Fidelity*.** This dimension evaluates how well the identity of the reference image is preserved in the customized or restored image in face customization and restoration tasks.

***Correspondence*.** This dimension evaluates text-to-image alignment in face generation and customization tasks.

3.3. Subjective Experiment

Main Assessment. To ensure a comprehensive, fair, and reliable subjective assessment, 180 participants with normal or corrected-to-normal eyesight are recruited for the experiment. They first take expert-led tutorials before the main assessment. In the main assessment, participants are asked to rate the image by dragging a sliding window from 0 to 5 with a two-digit decimal from different dimensions. All images are displayed on 27-inch 4K Dell monitors, randomly presented, at their original resolution under standard lighting conditions. The experiment follows the ITU-R BT.500-14 guidelines [78] for subjective evaluations.

Subjective Score Processing. To eliminate the inter-annotator variability, outlier detection based on Kurtosis is conducted with a rejection rate of 3% [78]. As a result, each image is rated by at least 15 subjects. These valid ratings are converted and scaled into Z-scores ranging from $[0, 100]$ with the following formulas:

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \quad z'_{ij} = \frac{100(z_{ij} + 3)}{6}, \tag{1}$$

$$\mu_i = \frac{1}{M_i}\sum_{j=1}^{M_i} r_{ij},\ \sigma_i = \sqrt{\frac{1}{M_i - 1}\sum_{j=1}^{M_i}(r_{ij} - \mu_i)^2}, \tag{2}$$

Here, r_{ij} represents the original ratings provided by the i-th evaluator for the j-th image,M_i denotes the total number of images that were assessed by the i-th participant in this group. Finally, the mean opinion score (MOS) of the image j (MOS_j) is computed by averaging the rescaled z-scores as follows:

$$MOS_j = \frac{1}{N_j}\sum_{i=1}^{N_j} z'_{ij} \tag{3}$$

where N_j denotes the number of valid subjects for image j, and z'_{ij} denotes rescaled z-scores.

4. F-Bench: Benchmarking Current Generative Face Models

In this section, we establish a human preference benchmark **F-Bench** for today's state-of-the-art face generation, customization, and restoration methods, from *quality, authenticity, ID fidelity, and correspondence* dimensions. This benchmark analyzes the strengths and weaknesses, providing new insights and potential directions for future work.

4.1. Perspective Analysis

Fig.3 demonstrate the MOS distribution across eight dimensions in the three subsets, highlighting distinct variations between dimensions. (1) In the face generation task, the *Correspondence* score is the highest among the three dimensions, indicating that existing generative models perform well in aligning with prompts. The distribution of *Authenticity* is skewed towards lower scores, suggesting that current face generation models still fall short in generating photorealistic face images that satisfy human preferences. (2) In face customization, models perform the worst in the *Quality* dimension overall, demonstrating the difficulty of injecting identity information without compromising image quality. *Identity Fidelity* shows a bimodal distribution, reflecting the dichotomous nature of human judgment regarding whether the generated face accurately represents the same identity. (3) In face restoration, models perform better in *Quality* than in *Identity Fidelity*.

4.2. Model-wise Comparison

Face Generation. Fig.4a presents the average MOS scores across three dimensions: *Quality*, *Authenticity*, and *Correspondence*. *Correspondence* scores show a steady increase

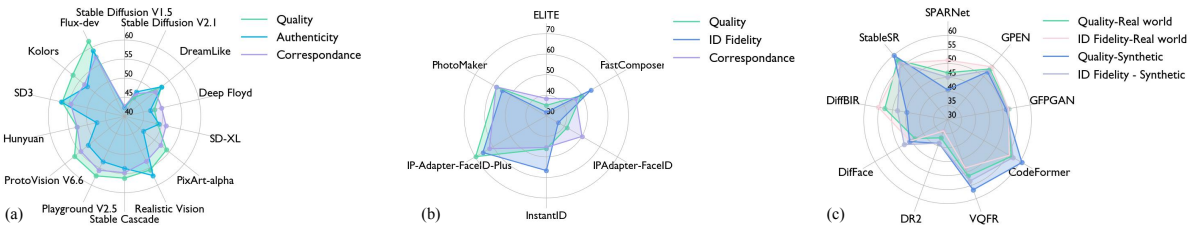


Figure 4. **Average MOS score comparison across all models and dimensions.** (a) Face generation. (b) Face customization. (c) Face restoration. The models are arranged in a clockwise order by release date.

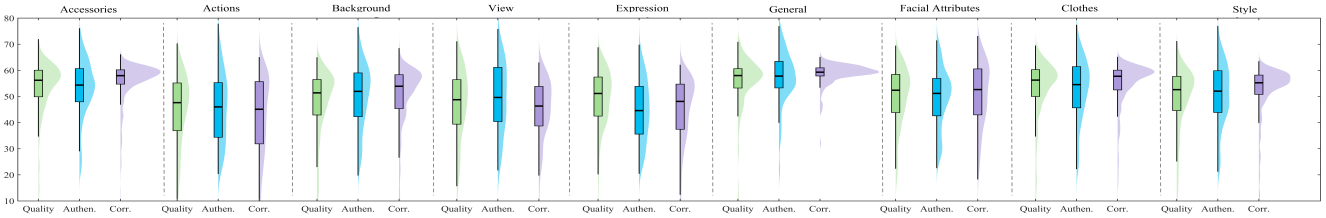


Figure 5. **Violin plots of quality, authenticity and correspondance scores in *FaceQ-Gen* subset across nine prompt categories.**

over time, indicating consistent improvement in text comprehension. *Quality* scores generally rise with the release year, although Hunyuan [58] and Dreamlike [2] exhibit deviations. Hunyuan [58] oversmoothes the facial details and performs worse compared to its contemporaries, while Dreamlike [2] performs the opposite. *Authenticity* scores exhibit the greatest variation among models, suggesting that current face generation models place insufficient emphasis on authenticity. Except for Realistic Vision [8], SD3 [9], Flux-dev [3], and Dreamlike [2], most models lag in authenticity relative to quality, highlighting a common phenomenon of high-quality but less realistic outputs in generative models. Fig.3d,e give some examples of MOS distributions of different methods, revealing a reasonable fine-grained performance difference.

Face Customization. Fig.4b is the performance radar chart for face customization models. IP-Adapter-FaceID-Plus[5] achieves the best performance across all dimensions, followed by PhotoMaker [57]. This highlights their great robustness in identity encoders. ELITE [98] scores the lowest in all dimensions. IP-Adapter-FaceID [4] produces high-quality images but lacks accuracy in *Identity Fidelity*. This may be because the base model [71] introduces many AI artifacts, which affect both *Identity Fidelity* and *Quality*. InstantID [91] and FastComposer [101] score well in *Identity Fidelity* but underperform significantly in *Quality* and *Correspondence*, indicating strong identity preservation but limited text-driven editing capabilities. Fig.3h illustrate the model-wise MOS distributions, revealing similar trends to those observed in the radar chart.

Face Restoration. Fig.4c displays the average performance comparison of nine face restoration models. DR2 [97] exhibits the weakest performance in both *Quality* and *Identity fidelity*. This may be due to the degradation removal process of DR2 [97], which results in the loss of fine details and excessive smoothing. In contrast, StableSR [89] and CodeFormer [113] achieve the best overall performance among the nine face restoration models, successfully preserving both intricate details and accurate identity representations. When comparing the performance on real-world and synthetic data, most methods struggle to achieve satisfactory results on complex real-world degradations, except for SAPRNet [15] and DiffBIR [62]. StableSR [89] and GFPGAN [94] show strong robustness, performing consistently well on both real-world and synthetic data.

4.3. Class-wise Comparison

For the face generation task, we analyze the Mean Opinion Score (MOS) distributions of *Quality*, *Authenticity*, and *Correspondence* across nine prompt categories, as shown in Fig.5. The violin plots highlight distinct performance patterns across these categories. The *Action*, *Expression*, and *Facial Attributes* categories exhibit lower MOS values in all three dimensions, with notable variability. In the *Action* category, limited visibility of the face due to body movements affects the generation of fine facial details, resulting in reduced scores, especially in *Quality* and *Correspondence*. The *Expression* category poses challenges in maintaining spatial relationships and physical constraints, leading to distortion and artifacts in expressions with exaggerated facial movements, such as laughing or crying, thereby impacting both *Quality* and *Authenticity*. For *Facial Attributes*, generating images with precise attributes (e.g., specific eye color, hair style) is challenging, particularly when multiple attributes are specified, which lowers the scores. In contrast, the *General* and *Background* categories achieve the highest MOS scores across all dimensions, as simpler or less specific prompts are easier for the model to interpret accurately. The *Clothes* and *Accessories* categories also perform well, indicating that the model handles clothes and accessory prompts effectively. Overall, prompts with specific or

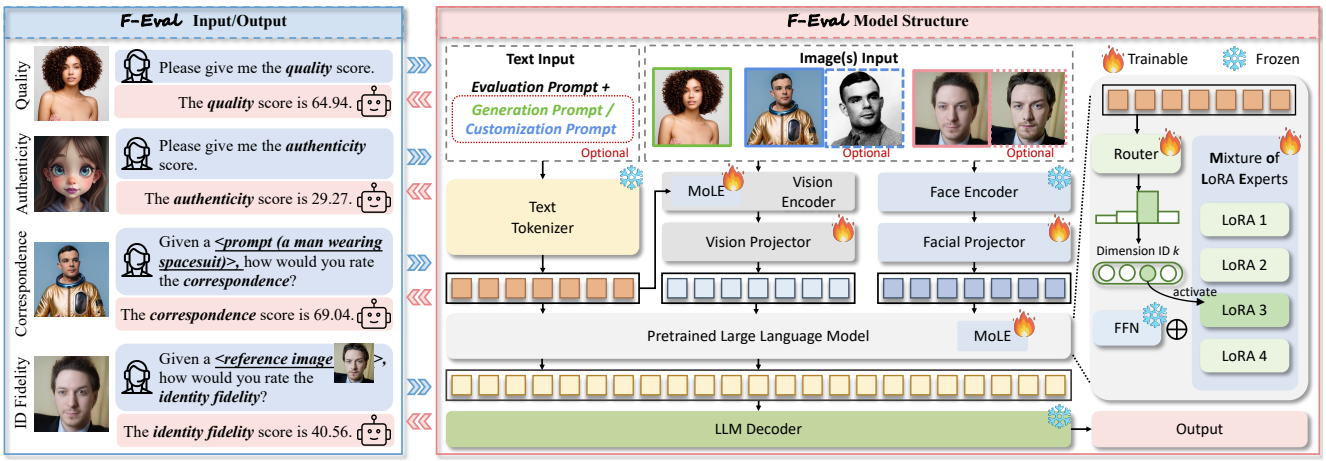


Figure 6. **The overall framework of F-Eval**. F-Eval can evaluate quality, authenticity, correspondence, and identity fidelity in a **one-for-all** framework. It can process both single and paired images, along with prompts, to produce quality scores. It consists of three encoders, including a vision encoder, a face encoder, and a text tokenizer to process multi-modal inputs. These features are projected into the same space by trained projectors. A pre-trained large language model is utilized to fuse the features while fine-tuned with four LoRA experts. Specific LoRA will be activated by dimension ID, which is classified by a trainable router.

complex attribute requirements tend to result in lower MOS scores, while general or simple prompts yield consistently high scores across dimensions. This can be attributed to the absence of high-quality fine-grained text-face pair datasets during the training of generative models.

5. F-Eval Method

In this section, we introduce our **one-for-all** generated face quality assessment method, **F-Eval**, towards predicting quality scores from multiple dimensions for three tasks using one model.

5.1. Overall Architecture

The overall architecture of F-Eval is depicted in Fig. 6. As shown on the left, F-Eval can assess four dimensions: quality, authenticity, text-image correspondence, and identity fidelity, with multimodal inputs, *i.e.*text prompts, image and optional reference image, to output quality-related answers. On the right, the input image(s) can originate from all three tasks. F-Eval employs a vision encoder and a face encoder to extract visual features, with projectors projecting both into a unified language space. User prompts are encoded into textual embeddings through a tokenizer. These multi-modal features are concatenated and fed into a pre-trained large language model for the final output.

5.2. Model Design

Vision Encoding. For input image(s), F-Eval adopt a vision encoder to extract global features and a face encoder captures face features related to facial structure and identity. Specifically, the vision encoder, Qwen-VL-2.5-ViT [11], is tuned on FaceQ to adjust different dimensions. Face encoder is Arcface [22], which is frozen. Two two-layer multi-layer perception projectors then map the features .

Feature Fusion via the LLM. For input prompts, it is tokenized into text tokens and then concatenated with visual tokens. Here, Qwen-VL2.5-7B [11] is employed as the pre-trained LLM backbone to combine visual tokens and text tokens to perform multi-modal learning.. Finally, the output features of the LLM are decoded with a text decoder and then projected to the prompt-corresponded quality space.

5.3. Multi-Expert LoRA Tuning

Instruction Tuning. Previous IQA methods are often designed for specific dimensions [106, 112], addressing only one aspect of the problem or requiring retraining of model weights for different dimensions. Leveraging the generalization ability of LLMs, instruction tuning allows for the integration of different dimensions within a unified framework. Therefore, we construct diverse instruction data for the four facial quality evaluation dimensions in the FaceQ dataset and fine-tune F-Eval by adjusting the visual encoder, projector, and LLM parameters to adapt it to the specific task. As a result, F-Eval is capable of evaluating facial quality using unified weights.

Mixture of LoRA Experts (MoLE). Tuning a holistic quality model for different dimensions using a single LoRA expert introduces inherent interference between dimensions. Hence, F-Eval integrates four LoRA experts for four dimensions when tuning vision encoder and LLM, constituting MoLE modules. A router is trained to classify input prompts into the four dimensions and output the corresponding dimension ID k. Specifically, for a frozen linear layer $h = Wx$ with input $x \in \mathbb{R}^{d_i}$ and weight matrix $W \in \mathbb{R}^{d_o \times d_i}$, a LoRA expert is trained to learn a low-rank decomposed update:

$$h = Wx + \Delta Wx = Wx + \frac{\alpha}{r} BAx, \tag{4}$$

Table 2. **Performance comparison on *FaceQ-Gen* and *FaceQ-Cus* subsets**. ♠, ♣, ♢, and ♡ denote traditional IQA models, face IQA models, classical deep learning-based IQA models, and AIGC IQA models respectively. * indicates fine-tuned results.

Task	Face Generation									Face Customization								
Dimension	Quality			Authenticity			Correspondence			Quality			ID Fidelity			Correspondence		
Model	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑
♠NIQE [67]	0.1726	0.1142	0.2553	0.0869	0.0571	0.1346	0.0752	0.0507	0.1098	0.1365	0.0899	0.1436	0.1490	0.0991	0.1310	0.1186	0.0784	0.1245
♠ILNIQE [111]	0.1981	0.1319	0.3159	0.0675	0.0445	0.1505	0.0889	0.0591	0.1665	0.2032	0.1327	0.2479	0.2192	0.1452	0.2778	0.1878	0.1251	0.2275
♠HOSA [102]	0.2537	0.1708	0.2972	0.1103	0.0728	0.1446	0.122	0.0814	0.1495	0.1019	0.0688	0.0597	0.0753	0.0506	0.0552	0.1036	0.0696	0.0748
♠BPRI-PSS [64]	0.2362	0.1595	0.2413	0.1700	0.1142	0.1772	0.1673	0.1121	0.1647	0.0238	0.0166	0.0314	0.1425	0.0972	0.1758	0.1100	0.0729	0.1067
♠FISBLIM [27]	0.0901	0.0595	0.0433	0.0028	0.0003	0.0175	0.1000	0.0665	0.0196	0.1614	0.1024	0.1690	0.0256	0.0173	0.0126	0.0852	0.0559	0.0854
♣ArcFace[22]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.3099	0.2094	0.3605	0.5062	0.3439	0.5572	0.1837	0.1243	0.2244
♣SER-FIQ[83]	0.1552	0.1034	0.1561	0.0470	0.0314	0.0575	0.1130	0.075	0.1110	-0.1241	-0.080	-0.1241	-0.0109	-0.0078	0.0042	-0.1680	-0.1121	-0.1911
♣DSL-FIQA[18]	0.5606	0.3927	0.5945	0.3156	0.2096	0.3960	0.3706	0.2514	0.4355	0.3615	0.2450	0.3651	0.2429	0.1652	0.2519	0.1040	0.0703	0.0974
♣IFQA[41]	0.2313	0.1549	0.3261	0.1597	0.1062	0.2097	0.1597	0.1077	0.2260	0.2401	0.1614	0.2220	0.1503	0.0973	0.1739	0.3731	0.2546	0.3941
♢CNNIQA* [42]	0.4219	0.2895	0.3510	0.3037	0.2054	0.2973	0.2824	0.1931	0.2595	0.6244	0.4376	0.6620	0.6511	0.4441	0.6381	0.3765	0.2545	0.3264
♢VGG16* [79]	0.5846	0.4141	0.6006	0.5010	0.3508	0.5181	0.4349	0.2994	0.4723	0.7217	0.5324	0.7857	0.6735	0.4717	0.6636	0.5635	0.3932	0.5589
♢VGG19* [79]	0.5728	0.4043	0.5644	0.4681	0.3248	0.4479	0.4171	0.2873	0.4754	0.7952	0.6031	0.8096	0.7544	0.5417	0.7555	0.6585	0.4742	0.6465
♢ResNet18* [29]	0.6150	0.4417	0.6660	0.4600	0.3160	0.4604	0.4829	0.3360	0.4628	0.7966	0.6013	0.8006	0.7473	0.5335	0.7656	0.6308	0.4488	0.6191
♢ResNet34* [29]	0.6092	0.4354	0.6537	0.5564	0.3949	0.5401	0.4410	0.3065	0.5207	0.8141	0.6211	0.8151	0.7969	0.5878	0.8157	0.5708	0.4000	0.5810
♢HyperIQA* [81]	0.6333	0.4545	0.6591	0.6106	0.4406	0.6042	0.4235	0.2912	0.4240	0.8524	0.6631	0.8205	0.8419	0.6463	0.8588	0.6628	0.4757	0.6454
♢TReS* [25]	0.7766	0.5867	0.8047	0.6458	0.4703	0.6579	0.5708	0.4038	0.6270	0.8883	0.7083	0.8951	0.8330	0.6409	0.8658	0.7393	0.5497	0.7400
♢MANIQA* [106]	0.7871	0.5951	0.8150	0.7308	0.5436	0.7794	0.6278	0.4443	0.7066	0.8970	0.7185	0.8952	0.8590	0.6654	0.8796	0.7849	0.5941	0.7822
♢PromptIQA* [14]	0.7823	0.5907	0.8245	0.7232	0.5327	0.7558	0.6514	0.4698	0.7065	0.9071	0.7315	0.8821	0.8693	0.6792	0.8876	0.8348	0.6342	0.8281
♡CLIPScore [32]	0.0967	0.0642	0.1315	0.0640	0.0408	0.0817	0.2396	0.1591	0.2733	0.3126	0.2106	0.3185	0.1816	0.1218	0.1811	0.7008	0.5098	0.6986
♡BLIPScore [53]	0.1306	0.0872	0.1929	0.0935	0.0613	0.1569	0.2051	0.1364	0.2963	0.3073	0.2063	0.3136	0.1213	0.0796	0.1266	0.6707	0.4797	0.6123
♡ImageReward [103]	0.3849	0.2631	0.4358	0.2606	0.1736	0.3181	0.5155	0.3542	0.5871	0.3232	0.2166	0.3196	0.1905	0.1270	0.1839	0.7714	0.5673	0.7740
♡MINTIQA[88]	0.7002	0.5160	0.8047	0.6217	0.4450	0.6826	0.5491	0.3891	0.6584	0.5378	0.3775	0.5574	0.4088	0.2816	0.4041	0.2886	0.1932	0.2959
♡MINTIQA* [88]	0.8312	0.6474	0.8974	0.8177	0.6306	0.8511	0.7908	0.5991	0.8667	0.8904	0.7092	0.8816	0.8524	0.6697	0.8755	0.8391	0.6413	0.8318
F-Eval (Ours)	0.8486	0.6670	0.9085	0.8312	0.6585	0.8578	0.8471	0.6637	0.9106	0.9462	0.7961	0.9461	0.9188	0.7640	0.9322	0.9460	0.7959	0.9457

where $A \in \mathbb{R}^{r \times d_i}$ and $B \in \mathbb{R}^{d_o \times r}$ are the low-rank matrices, $r \ll \min(d_o, d_i)$ is the rank, and α is scaling factor. For prompt of certain dimension, the corresponding expert is activated based on the dimension ID k, which can be formulated as:

$$E_k(x) = \frac{\alpha_k}{r_k} B_k A_k x, \tag{5}$$

where $E_k(\cdot)$ is the k-th LoRA expert.

6. Experiments

6.1. Implementation Details

The LLM used in F-Eval is Qwen-2.5VL [11], with an input token channel dimension of 3584. The vision encoder, face encoder, and LLM are all frozen during training. The ranks of LoRA experts are 16,32,16,16, respectively. The LoRA alphas are set equal to LoRA ranks. LoRA dropout is set as 0.05. The model is trained on 2 NVIDIA GeForce RTX 3090 GPUs with flash-attn [21] enabled. We use AdamW Optimizer to train the network for 10 epochs with a batch size of 4 and learning rate $1e^{-4}$.

6.2. Comparison with existing QA Models

We conduct comprehensive experiments on the proposed FaceQ database to evaluate the effectiveness of both current quality assessment methods and our proposed methods, as shown in Tab.2, Tab.3, and Tab.4. The evaluation metrics are Spearman rank-order correlation coefficient (SRCC), Kendall rank order correlation coefficient (KRCC), and Pearson linear correlation coefficient (PLCC), which assess the correlation between predicted scores by the QA methods and the MOS scores provided by humans. Higher metrics indicated better alignment with human preferences. The comparison QA models falls into four categories: 1) Traditional IQA Models. NR-IQA models include NIQE [67], ILNIQE [111], HOSA [102], BPRI-PSS [64], FISBLIM [27]. FR-IQA models include LPIPS [112], SSIM [96] and PSNR [34]. 2) Classical Deep learning-based IQA Models. This category includes CNNIQA [42], VGG16 [79], VGG19 [79], ResNet18 [29], ResNet34 [29], HyperIQA [81], TReS [25], MANIQA [106],PromptIQA[14]. All models are retrained on FaceQ with an 80:20 training/testing split. 3) AIGC IQA Models. This category includes pre-trained vision-language models, including CLIPScore [32], BLIPScore [53], zero-shot metric ImageReward [103], and the AIGC-specific QA model MINTIQA [88]. 4) Face IQA Models. This category includes SER-FIQ [83], DSL-FIQA [18], ArcFace [22], IFQA[41].

6.3. Performance Analysis

Traditional IQA Models. Traditional IQA methods, including both NR-IQA and FR-IQA models, generally perform poorly across all tasks and dimensions, with low SRCC, KRCC, and PLCC scores.

Classical Deep Learning-based IQA Models. After being trained on our dataset, deep learning-based IQA models, such as MANIQA [106] and HyperIQA [81], significantly outperform traditional models in terms of quality and authenticity. However, these methods face challenges in text understanding, leading to unsatisfactory correlations.

AIGC IQA Models. CLIPScore [32] BLIPScore [53], and

Table 3. **Performance comparison on *FaceQ-Res* synthetic subset.** ♠, ♡, ♣, and ♢ denote traditional NR-IQA models, FR-IQA models, face IQA models, and classical deep learning-based IQA models respectively. * indicates fine-tuned results.

Dimension	Quality			ID Fidelity		
Model	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑
♠BMPRI[65]	0.2161	0.1428	0.1862	0.1113	0.0739	0.1069
♠NIQE[67]	0.1798	0.1213	0.1642	0.2193	0.1480	0.2118
♠ILNIQE[111]	0.2505	0.1727	0.3180	0.2696	0.1857	0.3321
♠HOSA[102]	0.3758	0.2522	0.3748	0.3418	0.3455	0.2301
♠BPRI-LSSn [64]	0.1398	0.0957	0.0999	0.1244	0.0837	0.0973
♠BPRI [64]	0.1637	0.1065	0.1480	0.0195	0.0121	0.0281
♠FISBLIM [27]	0.1589	0.1089	0.2038	0.1054	0.0709	0.1715
♡LPIPS[112]	0.3909	0.2658	0.4643	0.4372	0.3018	0.4870
♡SSIM[96]	0.1017	0.0675	0.0488	0.0933	0.0601	0.0504
♡PSNR [34]	0.0198	0.0113	0.0536	0.0986	0.0662	0.1212
♣ArcFace[22]	0.3466	0.2327	0.3774	0.5728	0.4030	0.5762
♣SER-FIQ [25]	0.3813	0.2584	0.3746	0.3951	0.2641	0.3707
♣DSL-FIQA [18]	0.6387	0.4508	0.6528	0.4882	0.3360	0.5379
♣IFQA [41]	0.2273	0.1527	0.3348	0.3214	0.2185	0.3874
♢CNNIQA* [42]	0.4661	0.3197	0.3836	0.4678	0.3216	0.4288
♢ResNet18* [29]	0.7254	0.5295	0.6982	0.6662	0.4802	0.6527
♢ResNet34*[29]	0.6693	0.4916	0.6550	0.5905	0.4170	0.6084
♢VGG16* [79]	0.7487	0.5513	0.7101	0.6177	0.4413	0.6381
♢VGG19* [79]	0.6994	0.5126	0.6722	0.6038	0.4278	0.5999
♢HyperIQA* [81]	0.8269	0.6323	0.8228	0.7615	0.5585	0.7471
♢MANIQA* [106]	0.8287	0.6403	0.8778	0.7704	0.5660	0.7619
♢TReS* [25]	0.8656	0.6830	0.8622	0.7951	0.5879	0.7829
♢PromptIQA* [14]	0.8519	0.6661	0.8733	0.8220	0.6275	0.8254
F-Eval (Ours)	0.8692	0.6855	0.9009	0.8507	0.6731	0.8726

Table 4. **Performance comparison on *FaceQ-Res* real-world subset.** ♠, ♣, and ♢ denote traditional IQA models, face IQA models, and classical deep learning-based IQA models respectively. * indicates fine-tuned results.

Dimension	Quality			ID Fidelity		
Model	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑
♠BMPRI[65]	0.1582	0.1093	0.2799	0.0149	0.0108	0.1386
♠NIQE[67]	0.1783	0.1186	0.1962	0.1245	0.0810	0.1701
♠ILNIQE[111]	0.2230	0.1610	0.2291	0.1964	0.1362	0.2340
♠HOSA[102]	0.3927	0.4277	0.2654	0.3328	0.2234	0.3813
♠BPRI-LSSn [64]	0.2715	0.1840	0.3022	0.2059	0.1380	0.2475
♠BPRI [64]	0.1920	0.1313	0.2200	0.0846	0.0570	0.1136
♠FISBLIM [27]	0.2512	0.1705	0.2725	0.2341	0.1581	0.2699
♣ArcFace[22]	0.2371	0.1565	0.4183	0.4956	0.3470	0.6330
♣SER-FIQ [25]	-0.0482	-0.0299	0.1023	0.0421	0.0299	0.1584
♣DSL-FIQA [18]	0.6509	0.4641	0.6494	0.4840	0.3362	0.5118
♣IFQA [41]	0.0251	0.0210	0.1221	0.0893	0.0603	0.2086
♢CNNIQA* [42]	0.3650	0.2506	0.4705	0.3252	0.2197	0.3112
♢VGG16*[79]	0.4996	0.3439	0.4269	0.4117	0.2859	0.4622
♢VGG19* [79]	0.5871	0.4181	0.5133	0.3054	0.2044	0.3594
♢ResNet18* [29]	0.5977	0.4221	0.5722	0.5554	0.3909	0.6161
♢ResNet34* [29]	0.6259	0.4475	0.6168	0.5036	0.3539	0.6136
♢HyperIQA* [81]	0.7926	0.5965	0.7926	0.7783	0.5845	0.8163
♢MANIQA* [106]	0.8356	0.6482	0.8467	0.7597	0.5655	0.8081
♢TReS* [25]	0.8287	0.6391	0.8501	0.7287	0.5441	0.8009
♢PromptIQA* [14]	0.8115	0.6186	0.8645	0.7398	0.5526	0.8297
F-Eval(Ours)	0.8448	0.6577	0.8705	0.7957	0.6057	0.8366

ImageReward [103] exhibit weak correlations with human preference, likely due to their focus on high-level semantics rather than low-level details crucial for image quality assessment. MINTIQA [88] achieves best performance among current AIGC IQA methods, but its performance is still unsatisfactory, particularly on the customization task. Fine-tuning on our dataset leads to significant improvements, highlighting the great potential of FaceQ.

Face IQA Models. ArcFace similarity [22] demonstrates its ability to extract identity-related features. However, it lacks perceptual sensitivity for image quality, leading to poor performance in quality assessment. DSL-FIQA [18], as the latest method, achieves the best performance in predicting quality scores among face IQA methods, but it is still far from satisfactory. SER-FIQ fails to align with human perception across most dimensions and tasks. Experiments confirm that these natural face evaluation algorithms struggle to assess the complex characteristics of AIGFs, highlighting the need for an effective AIGF evaluator.

F-Eval (Ours). Our method achieves optimal performance across all tasks and dimensions, with the most significant gain in the customization task, as existing methods rarely focus human preferences in face customization. Among the four dimensions, the improvements in correspondence and ID fidelity stand out, thanks to dual-stream vision encoding and the powerful LLM backbone for text understanding and modality fusion. The MoLE technique allows the model to learn distinct features for each dimension.

6.4. Ablation Study

We conducted the ablation study to validate the utility of the core components of F-Eval in Tab. 5.

Effectiveness of Vision Encoding. 1) The effectiveness of the face encoder is demonstrated by comparing the second and last rows. Without the face encoder, a significant performance drop is observed, indicating that the extracted features help the LLM interpret face images. 2) Removing the projector leads to a dramatic performance drop.

Effectiveness of MoLE. We conducted three sets of LoRA adaptation experiments: fine-tuning a single LoRA for all dimensions and tasks, fine-tuning three LoRA experts for three tasks, and fine-tuning four LoRA experts for four dimensions. The results show that fine-tuning multiple dimensions with a single LoRA leads to performance deterioration, which further confirms the independence of the characteristics across different dimensions.

Table 5. Ablation study of the proposed F-Eval method. Here only present SRCC due to space limitations.

Feature & Strategy						Face Generation			Face Customization			Face Restoration syn		Face Restoration rw	
Vision	Projector	Face	Single LoRA	Task LoRA	Dim. LoRA	Quality	Authen.	Corres.	Quality	ID.	Corres.	Quality	ID.	Quality	ID.
✔		✔			✔	0.808	0.750	0.846	0.924	0.843	0.927	0.777	0.776	0.707	0.478
✔	✔				✔	0.815	0.752	0.835	0.924	0.855	0.921	0.812	0.783	0.716	0.603
✔	✔	✔	✔			0.793	0.776	0.819	0.941	0.885	0.942	0.838	0.823	0.802	0.671
✔	✔	✔		✔		0.823	0.780	0.838	0.942	0.873	0.942	0.824	0.826	0.754	0.638
✔	✔	✔			✔	**0.849**	**0.831**	**0.847**	**0.946**	**0.919**	**0.946**	**0.869**	**0.851**	**0.845**	**0.796**

7. Conclusion

In this work, we conduct a pioneering generative face image quality assessment study. Specifically, we introduce FaceQ, the first generated face quality assessment database comprising 12,255 face images and 491,130 human preference ratings. Then, we establish F-Bench, a benchmark for evaluating the capabilities of existing face generation, customization, and restoration models, while also reveals the limitations of current Image Quality Assessment (IQA) and Face Quality Assessment (FQA) methods. Furthermore, we propose an evaluation model F-Eval to accurately predict human preferences on generated face images, which demonstrates state-of-the-art performance. We believe this work will contribute to advancing generative models and help overcome current limitations in AIGC community.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grants 62271308, 62401365, 62225112, 62132006, U24A20220, STCSM under Grants 24ZR1432000, 24511106902, 24511106900, 22511105700, 22DZ2229005, 111 plan under Grant Number BP0719010, China Postdoctoral Science Foundation under Grants BX20250411, 2025M773473, and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

[1] Deep Floyd. https://github.com/deep-floyd/IF, Accessed: 2024-10-03. 2, 3

[2] Dreamlike V2. https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0, Accessed: 2024-10-03. 5

[3] Flux-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, Accessed: 2024-10-03. 1, 3, 5

[4] IP-Adapter-FaceID-SDXL. https://huggingface.co/h94/IP-Adapter, Accessed: 2024-10-03. 3, 5

[5] IP-Adapter-FaceID-Plus. https://huggingface.co/h94/IP-Adapter-FaceID, Accessed: 2024-10-03. 3, 5

[6] Kolors. https://huggingface.co/Kwai-Kolors/Kolors, Accessed: 2024-10-03. 3

[7] ProtoVision V6.6. https://huggingface.co/stablediffusionapi/protovision-xl-v6.6, Accessed: 2024-10-03. 3

[8] Realistic Vision V5.1. https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE, Accessed: 2024-10-03. 3, 5

[9] SD3. https://huggingface.co/stabilityai/stable-diffusion-3-medium, Accessed: 2024-10-03. 3, 5

[10] GPT-4o. https://chatgpt.com/, Accessed: 2024-10-10. 3

[11] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6, 7

[12] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. Cr-fiqa: face image quality assessment by learning sample relative classifiability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5836–5845, 2023. 2

[13] Nicolas Chahine, Stefania Calarasanu, Davide Garcia-Civiero, Théo Cayla, Sira Ferradans, and Jean Ponce. An image quality assessment dataset for portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9968–9978, 2023. 2

[14] Chen et al. Promptiqa: Boosting the performance and generalization for no-reference image quality assessment via prompts. In *ECCV*, 2024. 7, 8

[15] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K Wong. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing (TIP)*, 30:1219–1231, 2020. 2, 3, 5

[16] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11896–11905, 2021. 2

[17] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arxiv:2310.00426*, 2023. 2, 3

[18] Wei-Ting Chen, Gurunandan Krishnan, Qiang Gao, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Dsl-fiqa: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2931–2941, 2024. 2, 7, 8

[19] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2018. 2

[20] Siying Cui, Jiankang Deng, Jia Guo, Xiang An, Yongle Zhao, Xinyu Wei, and Ziyong Feng. Idadapter: Learning mixed features for tuning-free personalization of text-to-image models. *arXiv preprint arXiv:2403.13535*, 2024. 2

[21] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 7

[22] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 6, 7, 8

[23] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, et al. Finevq: Fine-grained user generated content video quality assessment. In *m*, 2025. 2, 3

[24] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[25] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1220–1230, 2022. 7, 8

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Pro-*

ceedings of the Advances in Neural Information Processing Systems, 2014. 1, 2

[27] Ke Gu, Guangtao Zhai, Min Liu, Xiaokang Yang, Wenjun Zhang, Xianghui Sun, Wanhong Chen, and Ying Zuo. Fisblim: A five-step blind metric for quality assessment of multiply distorted images. In *Proceedings of the SiPS*, pages 241–246, 2013. 7, 8

[28] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. *arXiv preprint arXiv:2205.06803*, 2022. 2, 3

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 7, 8

[30] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 2

[31] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 2

[32] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 33: 6840–6851, 2020. 1

[34] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *Proceedings of the IEEE International Conference on Pattern Recoginition (ICPR)*, pages 2366–2369. IEEE, 2010. 7, 8

[35] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2

[36] Qiang Hu, Qihan He, Houqiang Zhong, Guo Lu, Xiaoyun Zhang, Guangtao Zhai, and Yanfeng Wang. Varfvv: View-adaptive real-time interactive free-view video streaming with edge computing. *IEEE Journal on Selected Areas in Communications*, 2025. 2

[37] Qiang Hu, Zihan Zheng, Houqiang Zhong, Sihua Fu, Li Song, Xiaoyun Zhang, Guangtao Zhai, and Yanfeng Wang. 4dgc: Rate-aware 4d gaussian compression for efficient streamable free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 875–885, 2025. 2

[38] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 3

[39] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 36:78723–78747, 2023. 2

[40] Haoyun Jiang, Fei Huang, Qiang Hu, Minmin Sun, Shuai Xiao, Yong Li, Junyang Lin, Jiangchao Yao, et al. Catekv: On sequential consistency for long-context llm inference acceleration. In *Proceedings of International Conference on Machine Learning (ICML)*, 2025. 3

[41] Byungho Jo, Donghyeon Cho, In Kyu Park, and Sungeun Hong. Ifqa: Interpretable face quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3444–3453, 2023. 7, 8

[42] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7, 8

[43] 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 (CVPR)*, pages 4401–4410, 2019. 2

[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, 2023. 1

[45] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, 2016. 3

[46] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*, 2019. 2

[47] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes. 2013. 1

[48] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Proceedings of Advances in neural information processing systems (Neurips)*, 36:36652–36663, 2023. 1, 3

[49] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, 2023. 2

[50] Li et al. Agiqa-3k: An open database for ai-generated image quality assessment. *arXiv preprint arXiv:2306.04717*, 2023. 2

[51] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *arXiv prepreint arXiv: 2306.04717*, 2023. 1, 2, 3

[52] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. 3

[53] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Preceedings of International conference on machine learning (ICML)*, pages 12888–12900, 2022. 7

[54] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *The European Conference on Computer Vision (ECCV)*, 2018. 2

[55] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2706–2715, 2020. 2

[56] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 3

[57] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023. 1, 2, 3, 5

[58] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuandit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 2, 3, 5

[59] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19401–19411, 2024. 1, 2

[60] Zhang Lijun, Shao Xiaohu, Yang Fei, Deng Pingling, Zhou Xiangdong, and Shi Yu. Multi-branch face quality assessment for face recognition. In *2019 IEEE 19th International Conference on Communication Technology (ICCT)*, pages 1659–1664. IEEE, 2019. 2

[61] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European conference on computer vision (ECCV)*, pages 740–755. Springer, 2014. 3

[62] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 1, 2, 3, 5

[63] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 34892–34916, 2023. 2

[64] Xiongkuo Min, Ke Gu, Guangtao Zhai, Jing Liu, Xiaokang Yang, and Chang Wen Chen. Blind quality assessment based on pseudo-reference image. *IEEE Transactions on Multimedia (TMM)*, 20(8):2049–2062, 2018. 7, 8

[65] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting*, 64(2): 508–517, 2018. 8

[66] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual video quality assessment: A survey. *Science China Information Sciences*, 67(11): 211301, 2024. 2

[67] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 7, 8

[68] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 3

[69] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, 2021. 2

[70] Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023. 3

[71] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 5

[72] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[73] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[74] 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 (CVPR)*, pages 10684–10695, 2022. 1, 2, 3

[75] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 2

[76] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed

Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[77] 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. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 2

[78] B Series. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, 500(13), 2012. 4

[79] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, art. arXiv:1409.1556, 2014. 7, 8

[80] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 1

[81] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3664–3673, 2020. 7, 8

[82] Shaolin Su, Hanhe Lin, Vlad Hosu, Oliver Wiedemann, Jinqiu Sun, Yu Zhu, Hantao Liu, Yanning Zhang, and Dietmar Saupe. Going the extra mile in face image quality assessment: A novel database and model. *IEEE Transactions on Multimedia*, 2023. 2, 3

[83] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5651–5660, 2020. 7

[84] Yuan Tian, Kaiyuan Ji, Rongzhao Zhang, Yankai Jiang, Chunyi Li, Xiaosong Wang, and Guangtao Zhai. Towards all-in-one medical image re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30774–30786, 2025. 2

[85] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. *Proceedings of the European conference on computer vision (ECCV)*, pages 170–186, 2020. 2

[86] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. 2023. 1, 2, 3

[87] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 1, 2, 3

[88] Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Understanding and evaluating human preferences for ai generated images with instruction tuning. *arXiv preprint arXiv:2405.07346*, 2024. 7, 8

[89] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024. 5

[90] Liangbin Xie. Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution, 2022. 3

[91] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3, 5

[92] Tao Wang, Kaihao Zhang, Xuanxi Chen, Wenhan Luo, Jiankang Deng, Tong Lu, Xiaochun Cao, Wei Liu, Hongdong Li, and Stefanos Zafeiriou. A survey of deep face restoration: Denoise, super-resolution, deblur, artifact removal. *arXiv preprint arXiv:2211.02831*, 2022. 2

[93] 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)*. 3

[94] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5

[95] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[96] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 7, 8

[97] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1704–1713, 2023. 2, 3, 5

[98] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15943–15953, 2023. 2, 3, 5

[99] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13095–13105, 2023. 1

[100] 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 (CVPR)*, pages 2256–2265, 2021. 1

[101] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *Inter-*

national Journal of Computer Vision (IJCV), 133(3):1175–1194, 2025. 2, 3, 5

[102] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David S. Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing (TIP)*, 25:4444–4457, 2016. 7, 8

[103] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 2, 7, 8

[104] Liu Yang, Huiyu Duan, Long Teng, Yucheng Zhu, Xiaohong Liu, Menghan Hu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Aigcoiqa2024: Perceptual quality assessment of ai generated omnidirectional images. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1239–1245. IEEE, 2024. 3

[105] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016. 3

[106] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. *arXiv e-prints*, art. arXiv:2204.08958, 2022. 6, 7, 8

[107] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 672–681, 2021. 2, 3

[108] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2

[109] Zongsheng Yue and Chen Change Loy. Diface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022. 2, 3

[110] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020. 2

[111] Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing (TIP)*, 24(8):2579–2591, 2015. 7, 8

[112] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 6, 7, 8

[113] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 35:30599–30611, 2022. 2, 3, 5