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

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

Limitations and Social Impact. First, we are mindful of privacy concerns related to face datasets, and all data collection and sharing adhere to relevant privacy policies. Second, while the maximum resolution in our dataset is limited to 1024 pixels, we recognize that future work may incorporate higher-resolution images as generative models continue to advance.

1. *FaceQ*: Dataset Construction

1.1. Face Generation, Customization and Restoration Models

Model Implementation Details. Tab. 1 provides a comprehensive summary of models evaluated for face generation, editing, and restoration, including the model links, released dates, the resolution, and the backbone architectures. (1) Face generation. All the generation models are inference by pre-trained checkpoints in their default resolutions and hyper-parameters. Specifically, Stable Diffusion V1.5 [19], DreamLike [2], and RealisticVision [8] support high-resolution generation, such as 1024× 1024, but we utilize their default training resolution due to severe subject repetition phenomenon. Deep Floyd [1] is a 3-stage pixel space diffusion model, here we only consider the third-stage results. For the dynamic step sampling, the number of steps per stage for StableCascade [17] and Deep Floyd [1] is reduced to the quarter. Each model's negative prompt is configured to exclude "anime" and "semi-realistic" outputs by using terms like "worst quality", "low quality", "illustration", "3D", "2D", "painting", "cartoons", "sketch", "anime", "animation", "cartoon", and "semi-realistic". Safe sensors are enabled to filter out NSFW content. (2) Face customization. All the customization models are inference by pre-trained checkpoints in their default resolutions and hyper-parameters. FastComposer [31], originally designed for multi-subject customization, is assessed here with a single reference image as input. We use the IP-Adapter release version, including IP-Adapter-FaceID-SDXL [4] and IP-Adapter-FaceID-PlusV2 [5] (referred to as IP-Adapter-FaceID and IP-Adapter-FaceID-Plus). Their backbones are SDXL and SD-v1.5, respectively. (3) Face restoration. For DR2 [29], we follow the hyperparameter settings recommended in the original paper: $N = 4, T = 35$ for real-world inputs and $N = 8, T = 35$ for synthetic inputs.

Table 1. **Summary of 29 face generation, customization and restoration models.**

Category	Model	Year	Resol.	Backbone
Face Generation	Stable Diffusion V1.5 [19]	2022.04	512^2	Latent Diffusion
	Stable Diffusion V2.1 [19]	2022.12	1024^2	Latent Diffusion
	DreamLike V2.0 [2]	2023.01	768^2	Latent Diffusion
	Deep Floyd [1]	2023.04	1024^2	Pixel Diffusion
	SD-XL [18]	2023.06	1024^2	Latent Diffusion
	PixArt-alpha [11]	2023.11	1024^2	Latent Diffusion (DiT)
	Realistic Vision V5.1 [8]	2023.12	512^2	Latent Diffusion
	Stable Cascade [17]	2024.02	1024^2	Latent Diffusion
	Playground V2.5 [13]	2024.02	1024^2	Latent Diffusion
	ProtoVision V6.6 [7]	2024.03	1024^2	Latent Diffusion
	Hunyuan [15]	2024.05	1024^2	Latent Diffusion (DiT)
	SD3 [9]	2024.07	1024^2	Latent Diffusion (DiT)
	Kolors [6]	2024.07	1024^2	Latent Diffusion
	Flux-dev [3]	2024.08	1024^2	Latent Diffusion (DiT)
Face Customization	ELITE [30]	2023.02	512^2	Latent Diffusion
	FastComposer [31]	2023.05	512^2	Latent Diffusion
	IP-Adapter-FaceID [4]	2023.12	512^2	Latent Diffusion
	InstantID [25]	2023.12	512^2	Latent Diffusion
	IP-Adapter-FaceID-Plus [5]	2023.12	512^2	Latent Diffusion
	PhotoMaker [14]	2023.12	512^2	Latent Diffusion
Face Restoration	SPARNet [10]	2020.12	512^2	GAN
	GPEN [32]	2021.05	512^2	GAN
	GFPGAN [27]	2021.06	512^2	GAN
	CodeFormer [34]	2022.08	512^2	VQ
	VQFR [12]	2022.07	512^2	VQ
	DifFace [33]	2022.12	512^2	Pixel Diffusion
	DR2 [29]	2023.05	512^2	Pixel Diffusion
	DiffBIR [16]	2023.08	512^2	Latent Diffusion
	StableSR [24]	2024.06	512^2	Latent Diffusion

Prompt Examples. The face-centric prompts used for face generation can be categorized into nine classes. Tab. 2 presents two example prompts for each category due to space constraints. We ensured equal numbers of prompts for male and female subjects. The prompts for face customization are a subset of those used for face generation.
Degradation Scheme. We construct two synthetic degradation pipelines to mimic real-world degradation. The first is first order pipeline following previous works [10, 28, 29] which can be expressed as

$$I_d = [(I \circledast k_\sigma) \downarrow_r + n_\delta]_{\text{JPEG}_q} \quad (1)$$

Table 2. **Several examples of prompts for nine categories.**

Category	**Prompt Examples**
General	a photo of a woman a photo of a middle-eastern man
Clothing	a woman wearing a purple wizard outfit a man wearing a hoodie with green stripes
Accessory	a man with black hair styled in a top bun an old woman with a vintage hairpin
Action	a woman coding in front of a computer a man playing the violin
Expression	a man crying disappointedly, with tears flowing a woman looking shocked, mouth wide open
Background	a woman laughing on the lawn a young woman with a colorful umbrella stands near a crowd
View	a man wearing a doctoral cap, upper body, with the left side of the face facing the camera a man playing the guitar in the view of left side
Style	instagram photo, portrait photo of a man, perfect face, natural skin, hard shadows, film grain editorial portrait of a man posing dramatically, sharp lighting, fashion magazine style
Facial Attributes	a young girl with large round blue eyes, a flat nose bridge, and purple lipstick A man with narrow black eyes, a high nose bridge, a thick beard, and fair skin

High-quality images are degraded through a series of operations, including blurring, downsampling, additive Gaussian noise, and JPEG compression, with respective probabilities of 70%, 100%, 20%, and 70%. The blur kernel is randomly selected from Gaussian, Average, Median, and Motion blur. The interpolation method is randomly selected from Nearest, Linear, Area, and Cubic interpolation. The downsampling scale factor is randomly chosen from 4, 8, or 16. The second is a second-order degradation pipeline from previous work [26].

$$x = \mathcal{D}^n(y) = (\mathcal{D}_n \circ \cdots \circ \mathcal{D}_2 \circ \mathcal{D}_1)(y). \tag{2}$$

Blur, resizing, noise, and JPEG compression are conducted in several orders, along with a sinc filter to simulate common ringing and overshoot artifacts. We used these two pipelines to generate 50% and 50% of the synthetic low-quality images, respectively.

1.2. Additional Examples of *FaceQ* Database

Figure 1, Figure 2, and Figure 3 present additional examples from the *FaceQ-Gen*, *FaceQ-Cus*, and *FaceQ-Res* subsets, respectively. Each row corresponds to a specific generative model, showcasing the extensive diversity of content covered by the FaceQ dataset.

1.3. Quantitative Analysis of *FaceQ* Database

We selected four low-level features—brightness, contrast, colorfulness, and sharpness—to quantitatively assess the content diversity of the *FaceQ* database. Fig. 4 illustrates the kernel distribution curves for each selected feature across the three subsets. The results indicate that the images in each subset exhibit a wide range of contrast, colorfulness, and sharpness. However, the *FaceQ-Cus* subset demonstrates a narrower distribution in terms of brightness compared to the other subsets. We further calculate the relative range R_i^k and coverage uniformity U_i^k of the three subsets across these selected features. The relative range R_i^k is defined as:

$$R_i^k = \frac{\max(C_i^k) - \min(C_i^k)}{\max_k(C_i^k)}, \tag{3}$$

where C_i^k denotes the distribution of k_{th} dataset on i_{th} feature. $\max_k(C_i^k)$ refers to the maximum value of i_{th} feature across all datasets. The coverage uniformity U_i^k is calculated as the entropy of the B-bin histogram of C_i^k for each subset, using the following formula:

$$U_i^k = -\sum_{b=1}^{B} p_b \log_B p_b, \tag{4}$$

where p_b denotes the normalized number in bin b at i_{th} feature for k_{th} dataset. Fig. 5 presents a quantitative comparison of uniformity and relative range. A higher coverage uniformity indicates a more uniform feature distribution within the database, while a higher relative range reflects greater intra- and inter-dataset differences. It can be observed that all three subsets exhibit a diverse range and a uniform distribution across the four low-level features.

2. *FaceQ*: Subjective Experiments

2.1. Implementation Details

Fig. 6 presents screenshots of the user rating interfaces for the four tasks. In the generation task, as shown in Figure 6 (a), participants are asked to rate images on a scale of 0 to 5 based on *quality*, *authenticity*, and *correspondence*. Prompts are displayed beneath the candidate images, accompanied by translations into the participants' native languages. In the customization task, as shown in Fig. 6 (b), the reference image is displayed on the left, with prompts and translation shown below. In Fig. 6 (c), the candidate image appears on the left, while the corresponding low-quality reference image is on the right. In Figure 6 (d), both the low-quality image and the ground truth are displayed in synthetic scenarios. Each subset in *FaceQ* was randomly divided into four groups, each containing approximately 1,000 images. Participants were compensated $14 for completing each group of experiments according to [21]. At

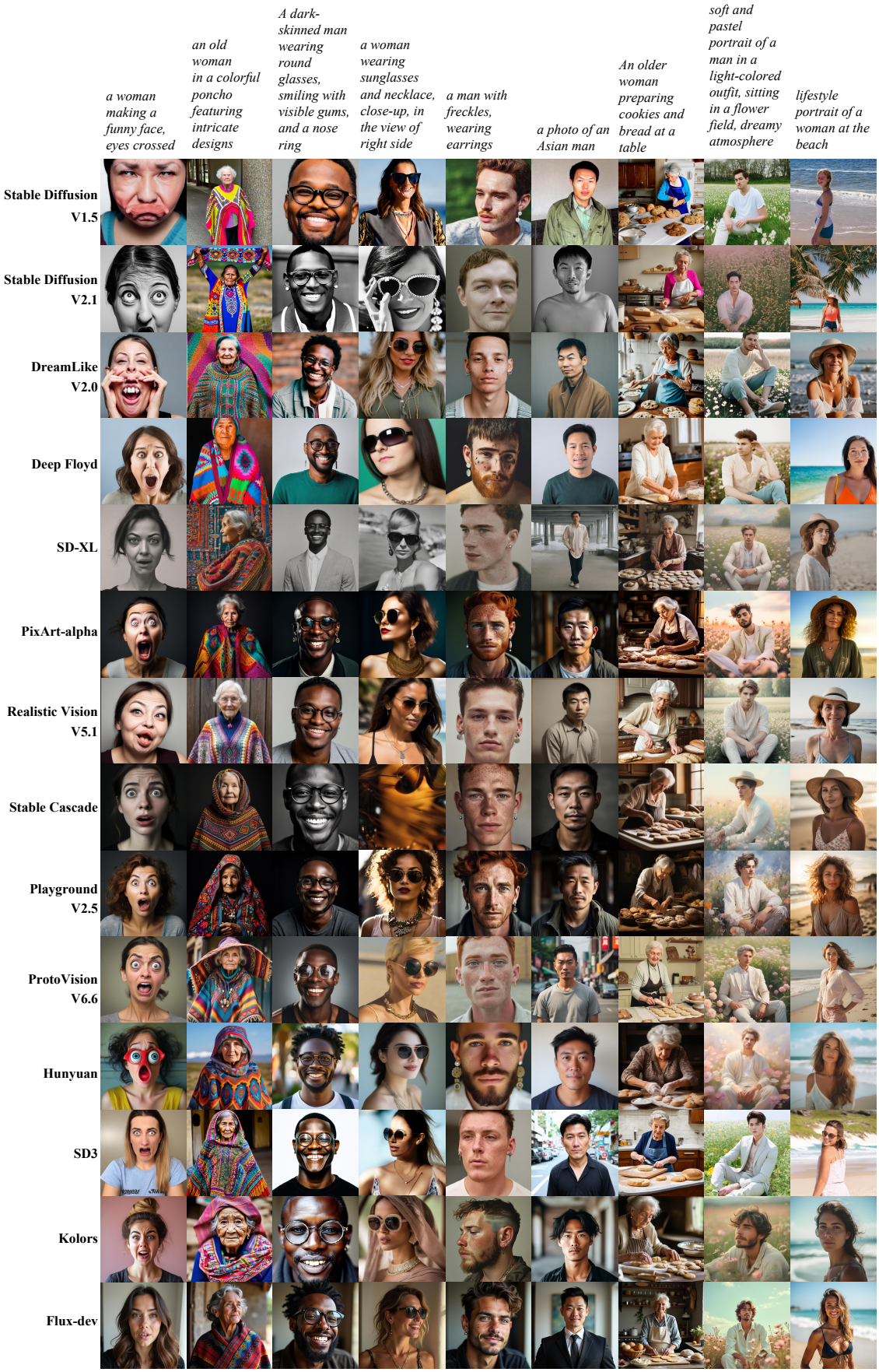

Figure 1. ***FaceQ-Gen*** **Examples.**

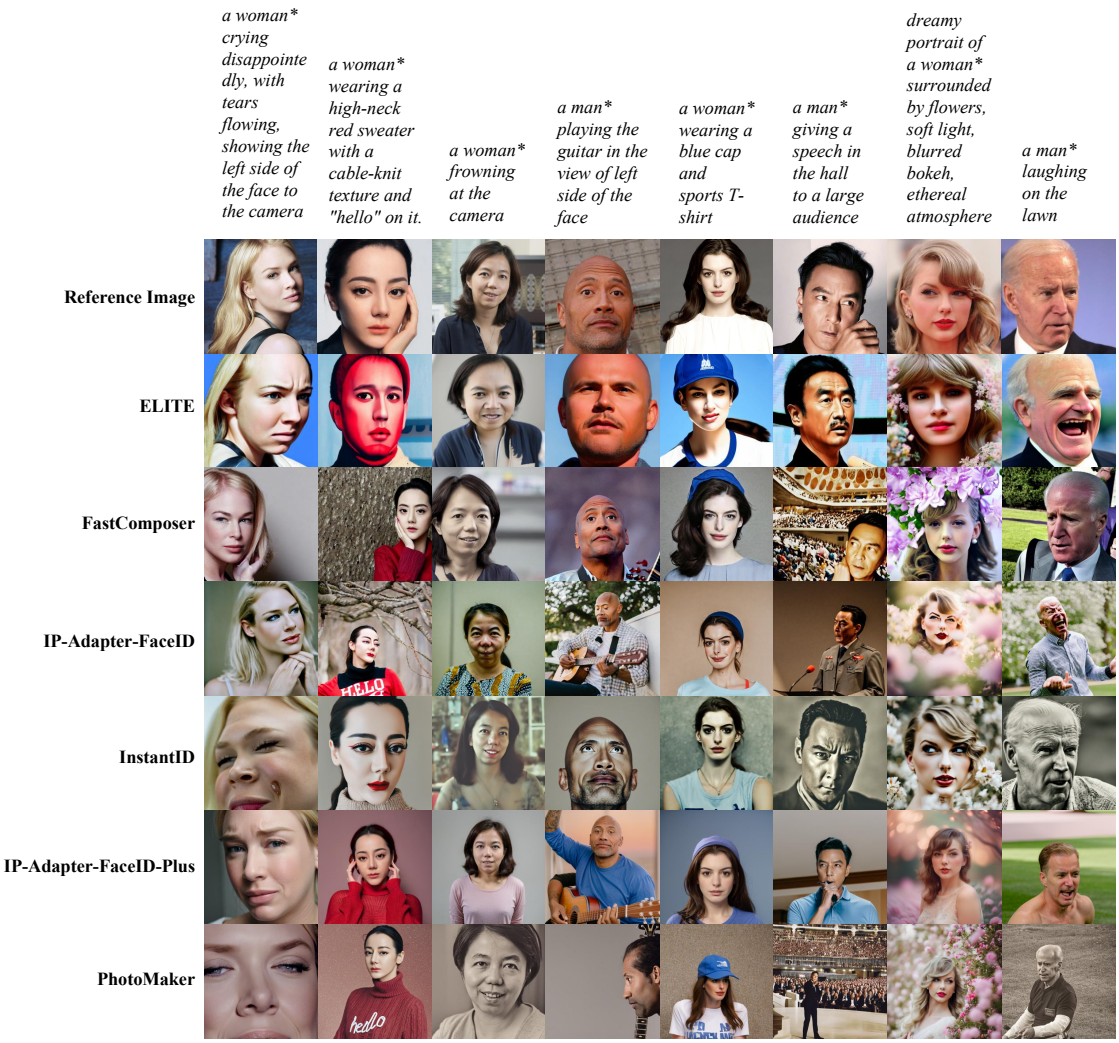

Figure 2. ***FaceQ-Cus* Examples.**

last, 3% invalid data are removed and no subject is removed.

2.2. Subjective Evaluation Examples

Fig. 7 provides a visual supplement to the 3D scatter plots described in the main submission. Fig. 7 (a) presents the 3D scatter plot for the *FaceQ-Gen* subset, showcasing five representative edge points. These images, ranked from top to bottom, correspond to overall good, low correspondence, low authenticity, low quality, and overall bad. As illustrated, the MOS scores effectively and intuitively capture the strengths and weaknesses of the images, accurately reflecting human preferences across different dimensions. Similarly, Fig. 7 (b) depicts the 3D scatter plot for the *FaceQ-Cus* subset, highlighting another set of five representative points. These images, ranked from top to bottom, correspond to overall good, low correspondence, low identity correspondence, low quality, and overall bad. The MOS score demonstrates a significant decline in dimensions where the image exhibits poor performance. This observation further substantiates the reliability and validity of human scoring in reflecting image quality across multiple dimensions.

3. *F-Bench*: More Analysis

3.1. MOS Distribution

Fig. 8 illustrates the MOS distributions of all fourteen face generation models across the dimensions of quality, authenticity, and correspondence, for both full-step and 1/4-step performances. The full-step distribution plots provide a comprehensive view of the performance distribution for different methods across the three dimensions, enabling a detailed evaluation of each method's effectiveness. In the 1/4-step distribution plots, it can be observed that models such as Stable Cascade [17], SDXL [18], and Pixart-alpha [11] exhibit high sensitivity to the reduction in step size. In contrast, models such as Flux[3] and RealisticVision[8]

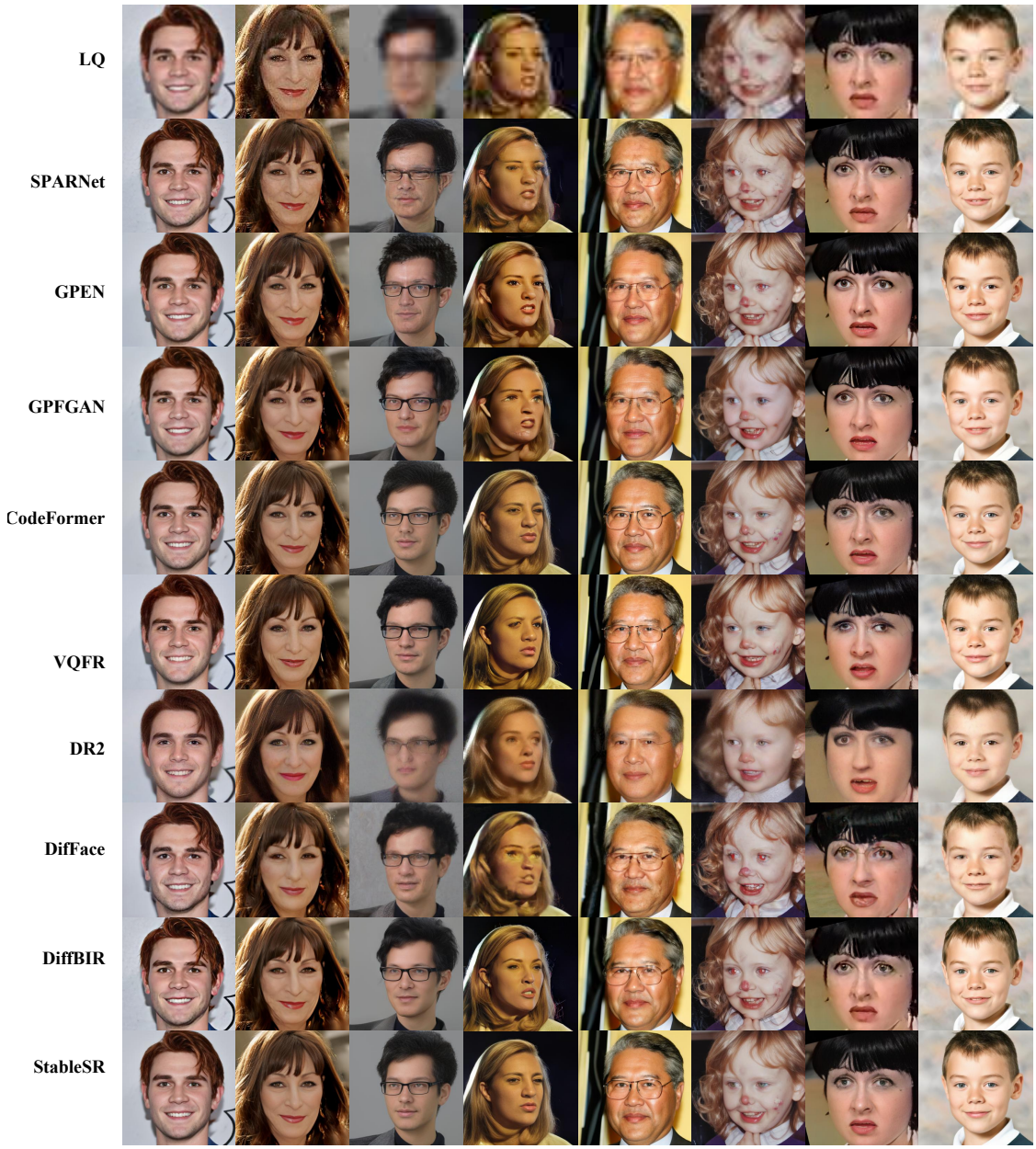

Figure 3. ***FaceQ-Res* Examples.**

demonstrate relatively stable performance with minimal degradation when reducing the steps. Fig. 9 illustrates the MOS distributions for six face customization models across three dimensions: quality, identity fidelity, and correspondence. Significant variations can be observed among different models and dimensions, highlighting distinct performance characteristics. Fig. 10 illustrates the MOS distributions for all nine face restoration models across the dimensions of quality and identity fidelity, evaluated for both real-world and synthetic cases. Most models exhibit varying performance between real-world and synthetic inputs, resulting in noticeable differences in their distributions, as exemplified by SPARNet [10].

3.2. Perspectiive Analysis

To provide a clearer comparison of the strengths and weaknesses of different methods, we present the average MOS scores across various dimensions in Figure 11. For the three dimensions of face generation, *authenticity* exhibits the largest disparity between methods, while *correspondence* and *quality* tend to cluster around higher scores. In the face customization task, the methods show inconsistent

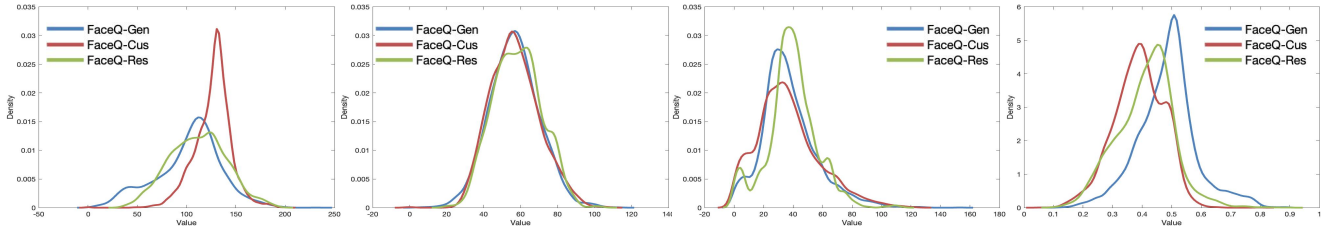

Figure 4. **Comparisons of the selected four low-level feature distributions calculated on proposed *FaceQ* dataset.**

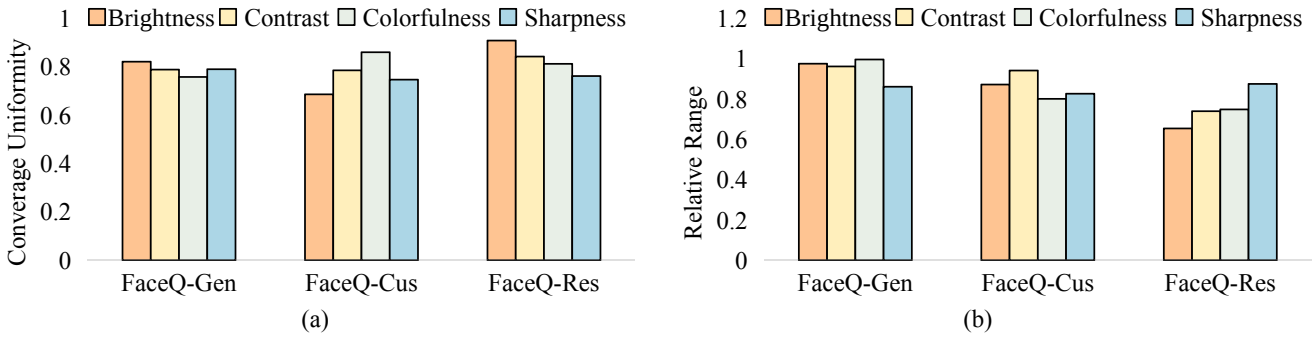

Figure 5. **Comparisons of the selected four low-level features calculated on the proposed *FaceQ* dataset.** (a) Coverage uniformity. (b) Relative range.

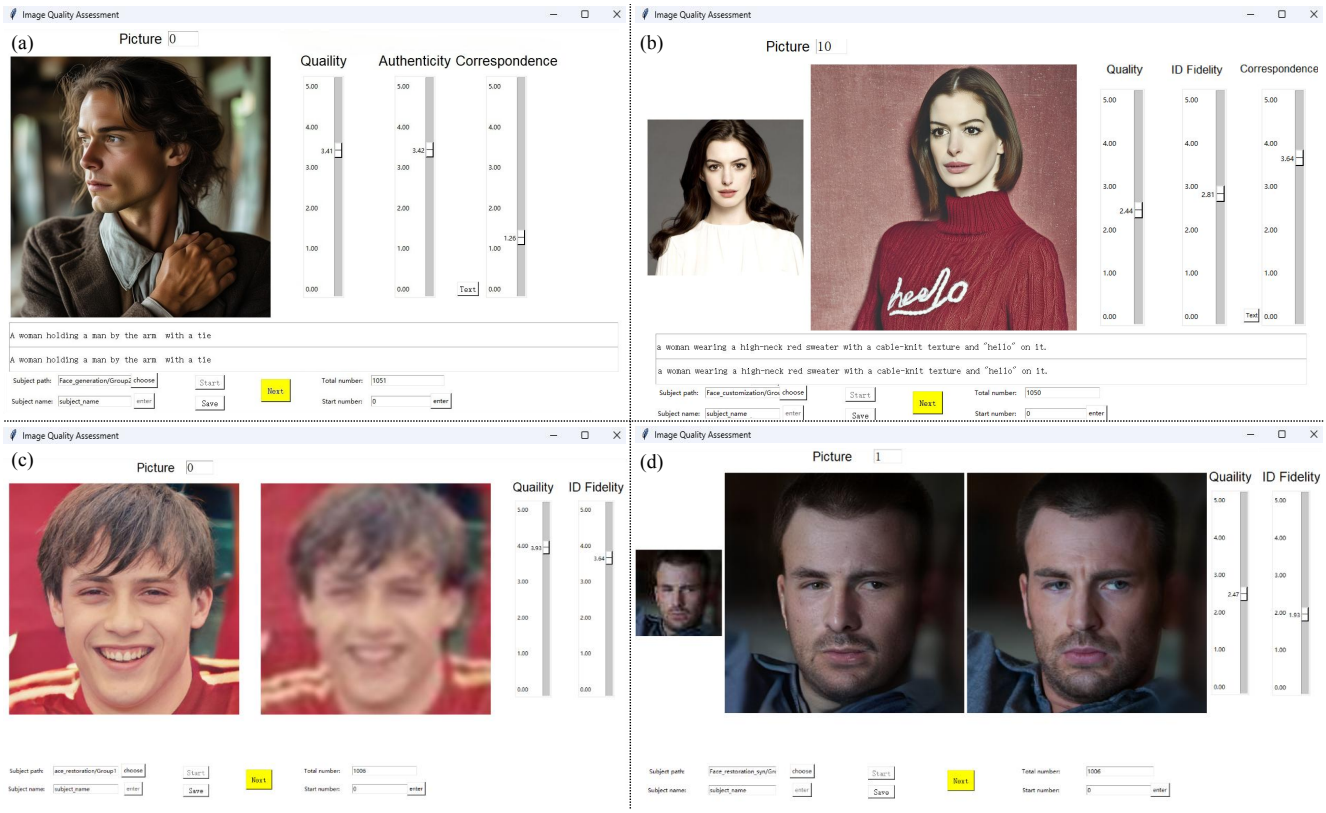

Figure 6. **Screenshots of the rating interface for human evaluation.** (a) Face generation evaluation interface. (b) Face customization evaluation interface. (c) Face restoration (real world) evaluation interface. (d) Face restoration (synthetic) evaluation interface.

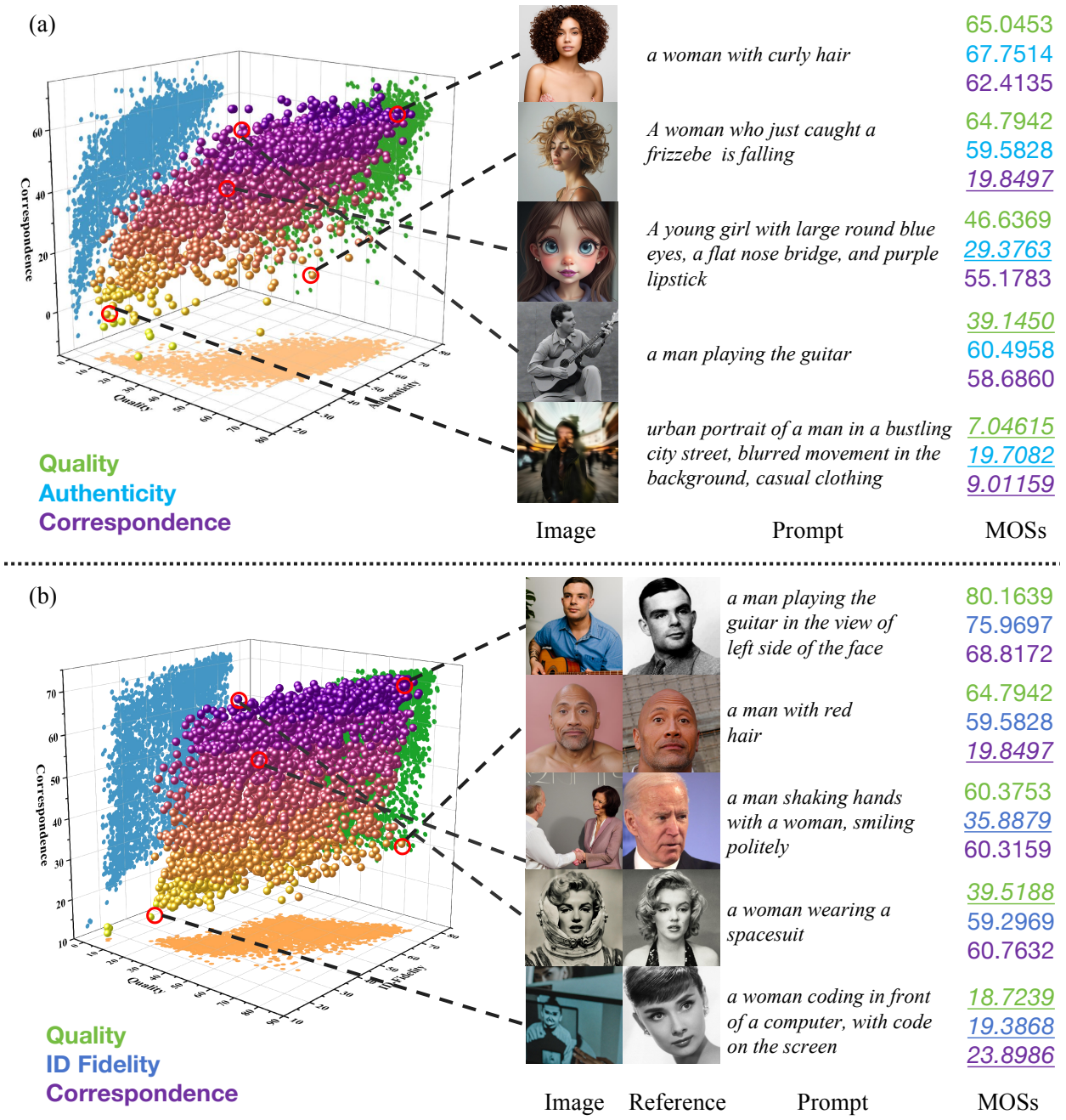

Figure 7. **Additional visualizations of the 3D scatter of MOSs.** We sample five representative points from the scatter and visualize their MOS scores across three dimensions. Each dimension is represented by a different color. The relatively low dimensions are *underlined*. (a) Face generation. (b) Face customization.

performance in *correspondence*, whereas *quality* remains relatively balanced. For face restoration, *quality-synthetic* emerges as the easiest metric to achieve high scores, followed by *quality-real-world*. Figure 12 displays the rankings of the various methods. For face generation, Flux [3] achieves the highest performance across all three dimensions. When considering *authenticity*, RealisticVision [8] and SD3 [9] outperform other methods. Playground [13] ranks second only to Flux [3] in terms of *correspondence*, while Kolors [6] and SD3 [9] follow Flux [3] in *quality*. On the other hand, SDv1.5 [**?**] performs the worst across all dimensions. For face customization, IP-Adapter-FaceID-

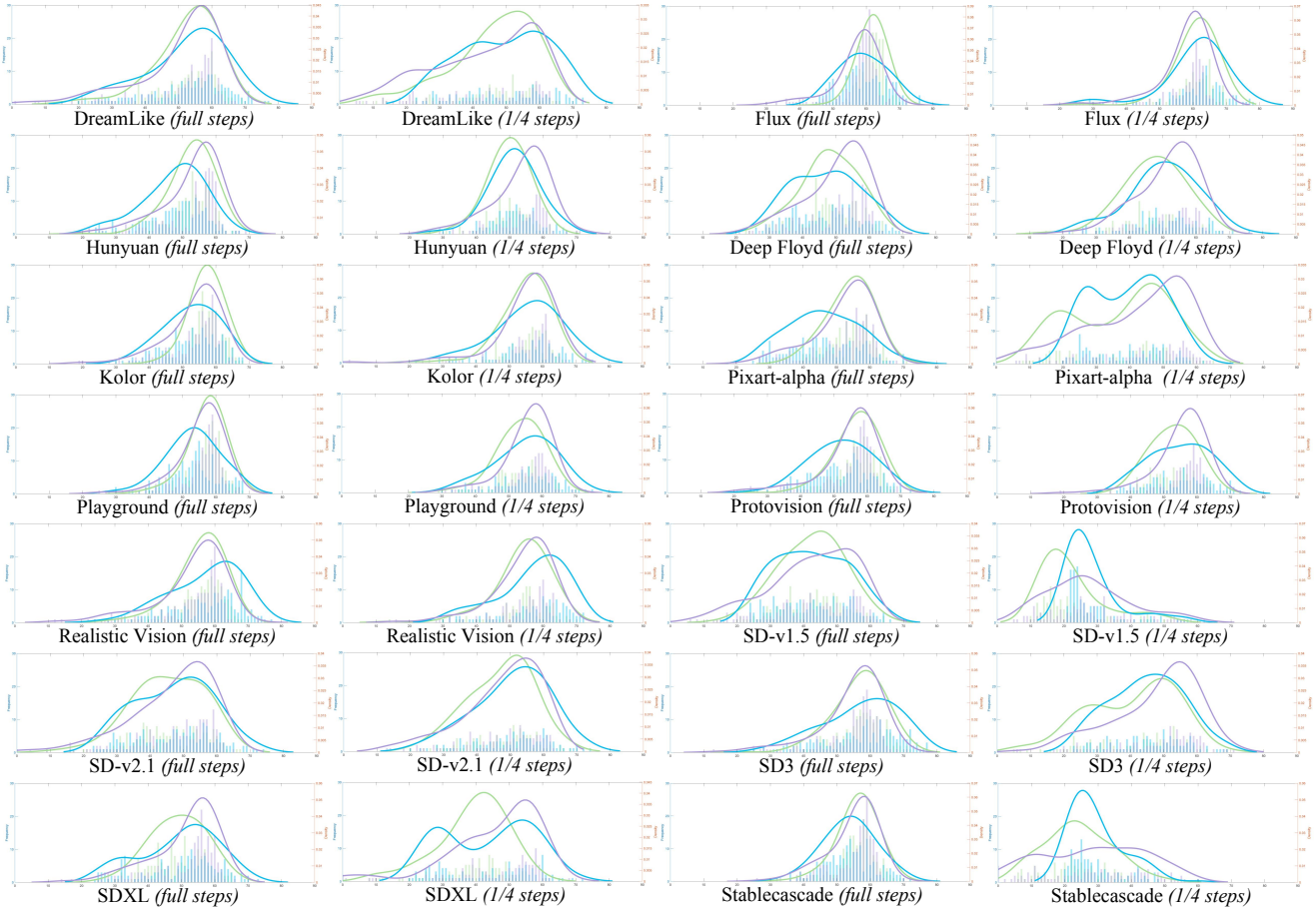

Figure 8. **MOS distribution histograms and kernel density curves across different *face generation* models.** "full steps"contains images generated in default sampling steps and "1/4 steps"contains the images generated by one-quarter of the default steps.

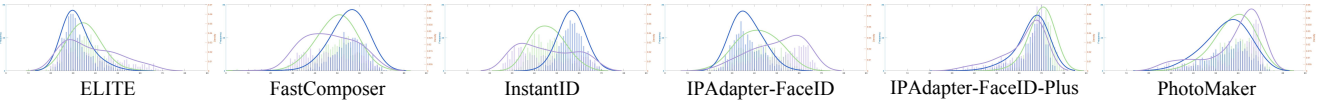

Figure 9. **MOS distribution histograms and kernel density curves across different *face customization* models.**

Plus [5], InstantID [25], and PhotoMaker [20] excel at preserving identity information. For face restoration, CodeFormer [34] demonstrates the best performance in synthetic scenarios, while StableSR achieves the highest scores in real-world scenarios.

3.3. Class-wise Comparison

Age. Figure 13 presents the multi-dimensional MOS distributions across three age groups (Young, Middle-aged, and Old) for face generation, face customization, and face restoration tasks. In the face generation task, the performance across age groups is relatively consistent across all dimensions. For face customization, more pronounced differences are observed, particularly in the *quality* scores, where older individuals exhibit larger variability. In the face restoration task, *quality* scores for old individuals are notably higher compared to middle-aged and young groups, while *identity fidelity* remains relatively consistent. These results highlight that face generation models are less sensitive to age-related factors, whereas face customization and restoration models demonstrate noticeable performance disparities among age groups, especially in dimensions such as ID Fidelity and Quality. The age and gender of the images are labeled InsightFace.

Gender. We visualize the distribution of MOS scores in three dimensions for men and women in each dimension in Fig. 14. It can be found that the male and female categories in face generation and face customization perform consistently across all evaluation dimensions, with minimal vari-

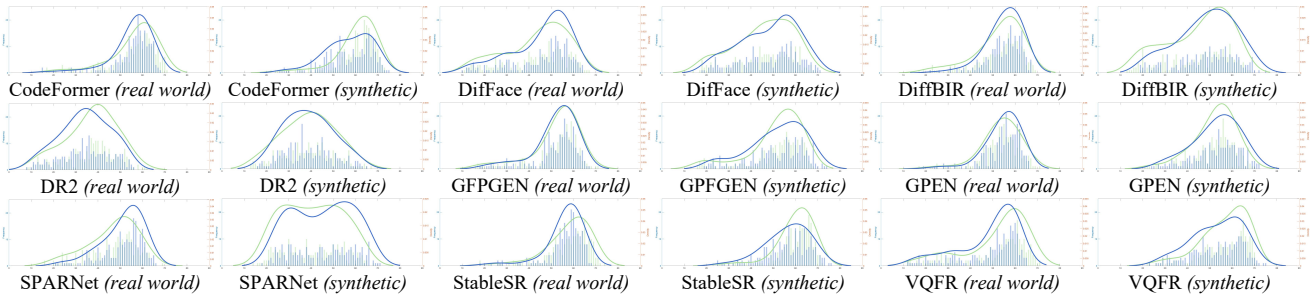

Figure 10. **MOS distribution histograms and kernel density curves across different *face restoration* models.** "synthetic"refers to images restored from the synthetic low-quality inputs while "real world"refers to the images restored from real-world low-quality inputs.

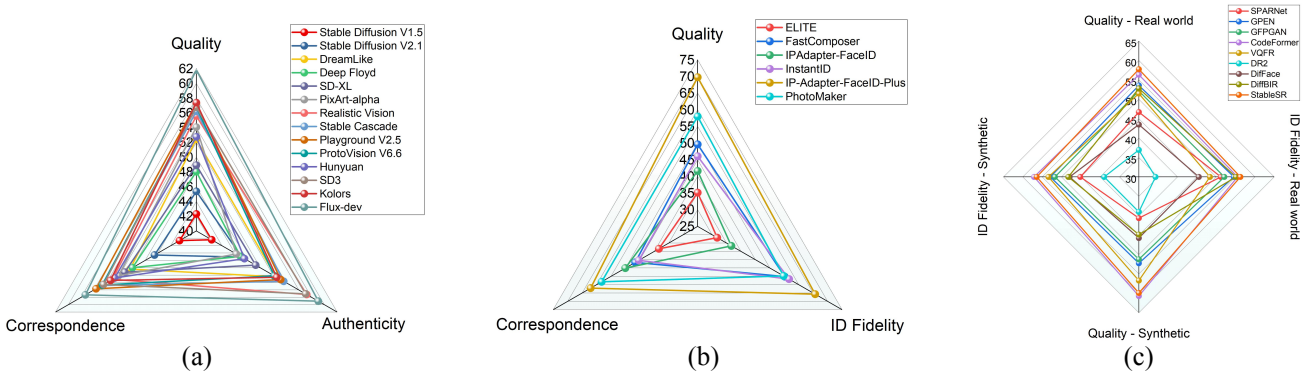

Figure 11. **Comparison of averaged MOS of different models across *Quality*,*Authenticity*, *ID Fidelity*, and *Correspondence*.** (a) Face generation models. (b) Face customization models. (c) Face restoration models.

ability observed. However, when it comes to face restoration tasks, the *quality* and *identity fidelity* of the male class are better. This suggests that generation and customization models trained on extensive datasets exhibit less gender bias than restoration models trained on smaller datasets.

4. QA Methods Implementation Details

4.1. Evaluation Metrics

We adopt three widely used metrics in IQA [22, 23]: Spearman rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), and Kendall rank correlation coefficient (KLCC) to evaluate the performance of quality assessment methods.

SRCC., which ranges from -1 to 1, evaluates the monotonic relationship between two variables. For N images, it is computed as:

$$SRCC = 1 - \frac{6\sum_{n=1}^{N}(v_n - p_n)^2}{N(N^2-1)}, \tag{5}$$

Here, v_n represents the rank of the ground truth value y_n, while p_n corresponds to the rank of the predicted value $\hat{y}_n$. When the SRCC value is higher, it signifies a stronger monotonic agreement between the ground truth and the predicted scores.

PLCC. quantifies the linear correlation between predicted scores and ground truth scores and is formulated as:

$$PLCC = \frac{\sum_{n=1}^{N}(y_n - \bar{y})(\hat{y}_n - \bar{\hat{y}})}{\sqrt{\sum_{n=1}^{N}(y_n - \bar{y})^2}\sqrt{\sum_{n=1}^{N}(\hat{y}_n - \bar{\hat{y}})^2}}, \tag{6}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the mean values of the ground truth scores and the predicted scores, respectively.

KLCC. measures the ordinal association between two measured quantities and is defined as:

$$KLCC = \frac{2(C-D)}{N(N-1)}, \tag{7}$$

where C is the number of concordant pairs and D is the number of discordant pairs among all possible pairs of N data points. A higher KLCC indicates a stronger rank correlation between the two variables. Together, these metrics provide a comprehensive evaluation of the relationship between predicted preference scores and ground truth MOS values across different aspects of correlation.

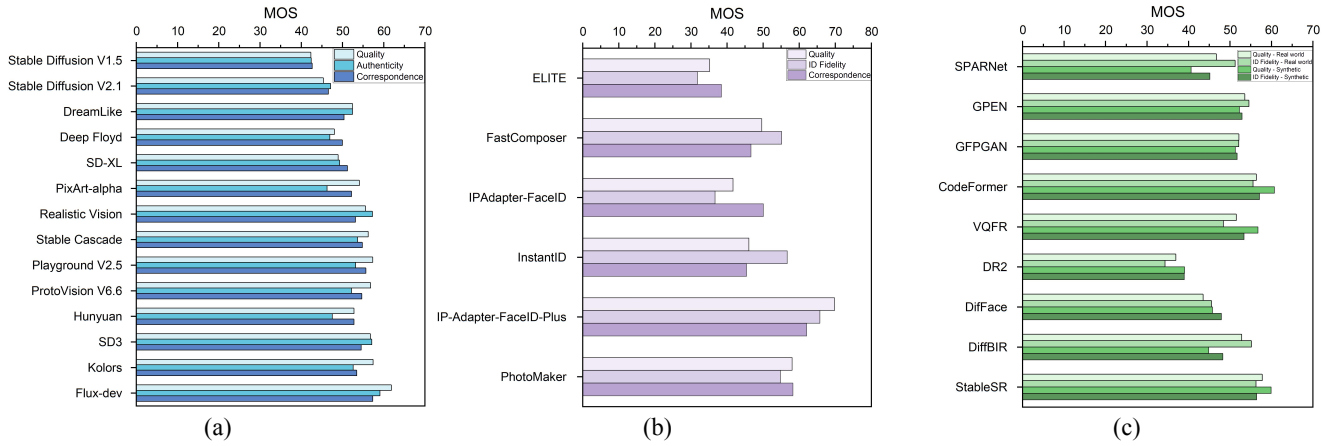

Figure 12. **Comparison of different model rankings based on the averaged MOS** (a) Face generation models. (b) Face customization

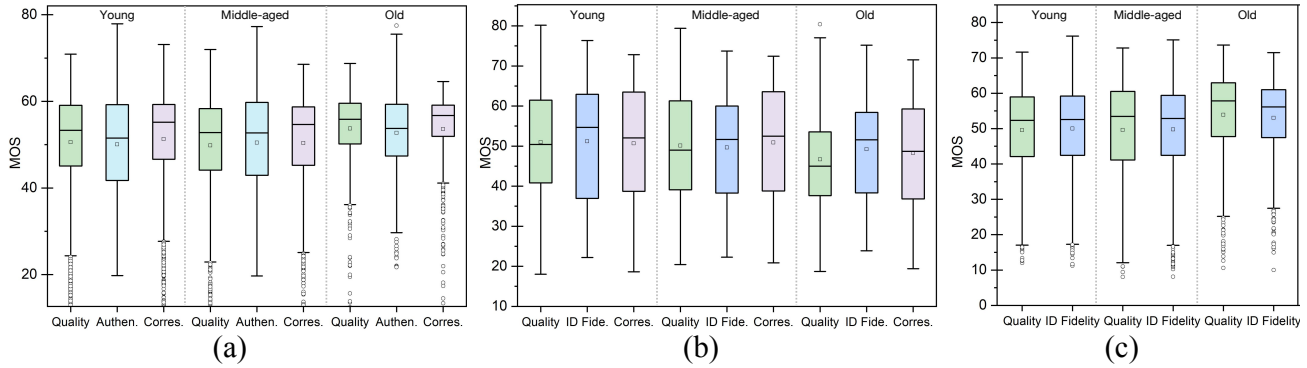

Figure 13. **Comparison of multi-dimensional MOS distributions across age groups.** "Authen.", "Corres."and "ID Fide."denote *Authenticity*, *Correspondence*, and *ID Fidelity* respectively. (a) Face generation models. (b) Face customization models. (c) Face restoration models.

5. More Details of Our F-Eval Model

5.1. Loss Funtion

We use both language loss, L1 loss and cross-entropy loss as the loss functions to optimize the training process. Specifically, the language loss is used to restrict the F-Eval to produce specific quality-related answer patterns. The language loss function can be formulated as:

$$\mathcal{L}_{\text{language}} = -\frac{1}{N}\sum_{i=1}^{N}\log P(y_{\text{label}}|y_{\text{pred}}) \tag{8}$$

where y_{pred} is the predicted token, y_{label} is the ground truth token, $P(y_{\text{label}}|y_{\text{pred}})$ indicates the probability, and N is the number of tokens.

L1 loss is used to regress the quality scores. The L1 loss can be formulated as:

$$\mathcal{L}_1 = \frac{1}{N}\left|q_{\text{pred}} - q_{\text{label}}\right|, \tag{9}$$

where q_{pred} is the predicted quality score, q_{label} is the ground truth quality score, and N is the number of images in a batch.

Cross-entropy loss to predict the dimension ID from the input text tokens. The cross-entropy loss can be formulated as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}), \tag{10}$$

where $y_{i,c}$ is the dimension label for the i-th sample in class c (dimension ID), $p_{i,c}$ is the predicted probability for the i-th sample in class c, and N N is the number of images in a batch. C is 4. The overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{language}} + \mathcal{L}_1 + \mathcal{L}_{\text{CE}}, \tag{11}$$

5.2. Failure Cases

F-Eval's performance may degrade when the face region in the image is either too small or in an extreme side view. In these scenarios, the face features extracted by the face encoder become less accurate, leading to a drop in performance, particularly in the quality and authenticity dimensions. We will address this limitation by integrating a more robust face encoder fine-tuned on side view and small face

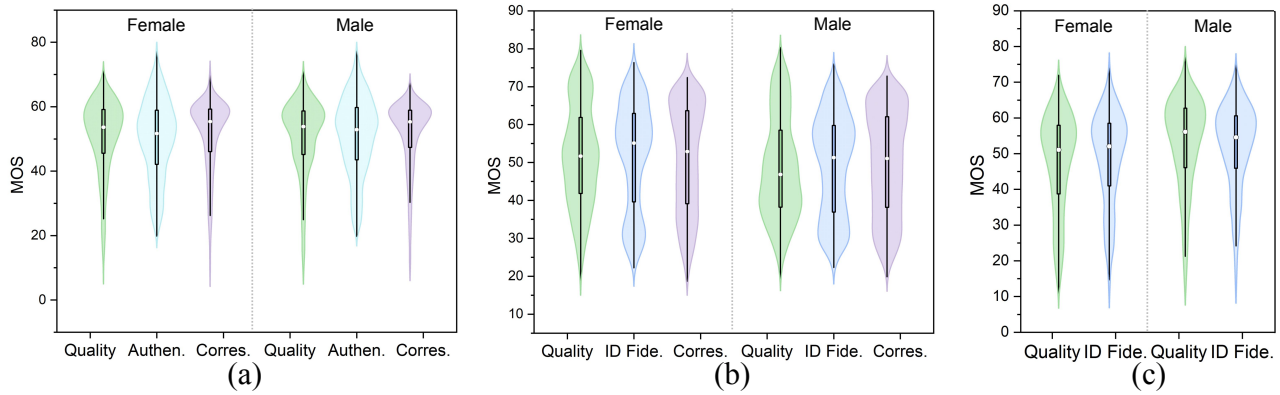

Figure 14. **Comparison of multi-dimensional MOS distributions across genders.** "Authen.", "Corres."and "ID Fide."denote *Authenticity*, *Correspondence*, and *ID Fidelity* respectively. (a) Face generation models. (b) Face customization models. (c) Face restoration models.

Table 3. Complete results in the ablation study.

Task	Face Generation									Face Restoration syn					
Dimension	Quality			Authenticity			Correspondence			Quality			ID Fidelity		
Method	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑
Freeze Projector	0.8080	0.6428	0.8698	0.7518	0.5832	0.8123	0.8475	0.6783	0.9031	0.7769	0.5787	0.8074	0.7757	0.5832	0.8074
Freeze Face encoder	0.8151	0.6450	0.8733	0.7524	0.5834	0.8140	0.8347	0.6654	0.8912	0.8121	0.6120	0.8430	0.7828	0.6021	0.8341
Single LoRA	0.7927	0.6140	0.8659	0.7757	0.5995	0.8285	0.8191	0.6429	0.8760	0.8377	0.6538	0.8874	0.8228	0.6341	0.8556
Task LoRA	0.8231	0.6477	0.8895	0.7803	0.6056	0.8317	0.8377	0.6661	0.8943	0.8244	0.6428	0.8806	0.8257	0.6359	0.8519
F-Eval (Ours)	0.8486	0.6670	0.9085	0.8312	0.6585	0.8578	0.8471	0.6637	0.9106	0.8692	0.6855	0.9009	0.8507	0.6731	0.8726
Task	Face Customization									Face Restoration rw					
Dimension	Quality			Authenticity			Correspondence			Quality			ID Fidelity		
Method	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑	SRCC↑	KRCC↑	PLCC↑
Freeze Projector	0.9242	0.7651	0.9233	0.9269	0.7679	0.9242	0.8429	0.6726	0.8386	0.7067	0.5249	0.7114	0.4778	0.3616	0.5451
Freeze Face encoder	0.9239	0.7662	0.9217	0.9214	0.7612	0.9208	0.8550	0.6836	0.8524	0.7158	0.5421	0.8532	0.6028	0.4386	0.6898
Single LoRA	0.9419	0.7938	0.9421	0.9424	0.7957	0.9430	0.8850	0.7140	0.8974	0.8019	0.6155	0.8455	0.6710	0.5017	0.7768
Task LoRA	0.9421	0.7984	0.9403	0.9422	0.7985	0.9402	0.8727	0.7005	0.8918	0.7540	0.5661	0.7964	0.6376	0.4736	0.7403
F-Eval (Ours)	0.9462	0.7961	0.9461	0.9188	0.7640	0.9322	0.9460	0.7959	0.9457	0.8448	0.6577	0.8705	0.7957	0.6057	0.8366

data in future works. Additionally, F-Eval currently does not support identifying specific distorted regions, which will also be addressed in future works.

5.3. More Experimental Results

The detailed ablation study results are listed in Tab. 3. The results indicate the effectiveness of the key components in our F-Eval.

References

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

[2] Dreamlike V2. https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0, Accessed: 2024-10-03. 1

[3] Flux-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, Accessed: 2024-10-03. 1, 4, 7

[4] IP-Adapter-FaceID-SDXL. https://huggingface.co/h94/IP-Adapter, Accessed: 2024-10-03. 1

[5] IP-Adapter-FaceID-Plus. https://huggingface.co/h94/IP-Adapter-FaceID, Accessed: 2024-10-03. 1, 8

[6] Kolors. https://huggingface.co/Kwai-Kolors/Kolors, Accessed: 2024-10-03. 1, 7

[7] ProtoVision V6.6. https://huggingface.co/stablediffusionapi/protovision-xl-v6.6, Accessed: 2024-10-03. 1

[8] Realistic Vision V5.1. https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE, Accessed: 2024-10-03. 1, 4, 7

[9] SD3. https://huggingface.co/stabilityai/stable-diffusion-3-medium, Accessed: 2024-10-03. 1, 7

[10] 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. 1, 5

[11] 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. 1, 4

[12] 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. 1

[13] 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. 1, 7

[14] 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

[15] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1

[16] 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

[17] 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. 1, 4

[18] 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. 1, 4

[19] 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

[20] Qi Shan, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz. Photo uncrop. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 16–31. Springer, 2014. 8

[21] M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. Responsible research with crowds: pay crowdworkers at least minimum wage. *Communications of the ACM*, 61(3):39–41, 2018. 2

[22] 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. In *Proceedings of the CAAI International Conference on Artificial Intelligence (CICAI)*, pages 46–57. Springer, 2023. 9

[23] 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. 9

[24] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024. 1

[25] 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. 1, 8

[26] 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)*. 2

[27] 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. 1

[28] 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. 1

[29] 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. 1

[30] 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. 1

[31] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision (IJCV)*, 133(3):1175–1194, 2025. 1

[32] 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. 1

[33] Zongsheng Yue and Chen Change Loy. Diffface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022. 1

[34] 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. 1, 8