WHAT MAKES A FACE EDIT RELIABLE? BENCHMARKING RELIABILITY IN TEXT-GUIDED FACE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-guided face editing requires fulfilling editing instructions while preserving the subject's identity. However, existing evaluation practices primarily emphasize edit effectiveness, leaving identity preservation insufficiently examined and potentially favoring models that follow instructions at the expense of identity. To address this gap, we introduce **FaceEdit-Bench**, a benchmark that evaluates identity preservation and edit effectiveness as complementary dimensions of editing reliability. FaceEdit-Bench comprises approximately 110K annotated face edits generated by 15 editing models across 20 instruction types. We assess identity preservation by aggregating scores from multiple face recognition systems. To evaluate edit effectiveness, we develop an instruction-conditioned agent that decomposes each instruction into verifiable requirements and assesses their fulfillment using tool-derived visual evidence. Our analysis reveals divergent model rankings across the two dimensions: *stronger instruction fulfillment does not necessarily imply better identity preservation*, highlighting the limitations of instruction fulfillment alone as a measure of editing reliability. We further introduce **FaceEditEval**, a vision-language model that consolidates supervision from the recognition systems and evaluation agent into a single evaluator, enabling efficient prediction of identity preservation and edit effectiveness as separate scores. FaceEditEval achieves higher correlations with the automated reference scores than existing quality assessment methods. By accounting for both the requested changes and the preservation of identity, FaceEdit-Bench and FaceEditEval enable more comprehensive comparisons of editing models and support research on reliable text-guided face editing.

1 INTRODUCTION

Text-guided image editing (Brooks et al., 2023; Zhang et al., 2023a; Zhao et al., 2024) enables users to modify visual content through natural-language instructions. When editing faces, however, making the requested change is only part of the task: the edited image should also retain the identity of the person in the source image. These objectives are related but not interchangeable. For example, adding a beard may satisfy an instruction while also altering identity-defining facial features; conversely, leaving a face largely unchanged may preserve identity while failing to perform the requested edit. As illustrated in Figure 1, successful instruction fulfillment alone is therefore insufficient to establish reliable face editing. This motivates our central question: *what makes a face edit reliable, and how can this reliability be systematically evaluated?*

Existing evaluations of text-guided face editing (Liu et al., 2025a) provide only a partial answer: they often emphasize instruction fulfillment, leaving identity preservation insufficiently examined. This emphasis can favor editors that achieve the requested changes at the expense of the subject's identity. To address this gap, we assess reliability along two complementary dimensions, **identity preservation** and **edit effectiveness**. Identity preservation assesses whether the edited image still depicts the same person as the source image, whereas edit effectiveness measures the extent to which the requested modification is achieved. Evaluating these dimensions separately distinguishes reliable edits from those that satisfy one requirement while failing the other.

Towards this goal, we introduce **FaceEdit-Bench**, comprising approximately 110K face edits generated by 15 editing models across 20 instruction types. For identity preservation, face recognition embeddings provide identity-oriented representations for comparing source and edited images. We

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

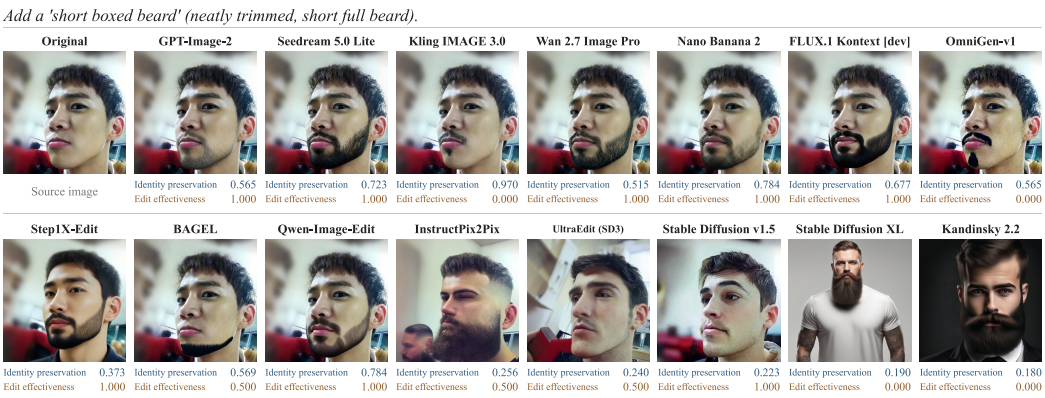


Figure 1: Visualization of face editing results across models under the same source image and instruction. Reference scores indicate identity preservation and edit effectiveness.

aggregate similarity scores from nine recognizers to reduce dependence on any single model. For edit effectiveness, assessing the requested change requires instruction-specific visual evidence. We therefore develop an instruction-conditioned agent that decomposes each instruction into atomic visual requirements and assesses their fulfillment through source–edit comparisons supported by visual analysis tools. These complementary annotation procedures provide reference scores grounded in identity similarity and instruction-specific visual evidence for the two reliability dimensions.

Our benchmark analysis reveals divergent model rankings across the two dimensions: *stronger instruction fulfillment does not necessarily imply better identity preservation*. This relationship also varies across editing categories. Specifically, style editing achieves the highest average edit effectiveness but the lowest identity preservation, while facial attribute editing exhibits the opposite pattern. These findings highlight failure modes that evaluation along a single dimension can obscure.

Building on these annotations, we introduce FaceEditEval, which learns dimension-specific score predictors by fine-tuning Qwen3-VL-8B-Instruct (Bai et al., 2025) with LoRA (Hu et al., 2021). For each dimension, we train separate evaluators under source-available (SA) and source-unavailable (SU) settings, depending on whether the source image is provided. In the SU setting, identity preservation is estimated rather than directly verified. Experiments show that FaceEditEval achieves stronger correlations with the automated reference scores than the evaluated image quality methods, face quality methods, and vision-language models, including on outputs from unseen closed-source editors.

Our main contributions are:

- **Reliability formulation and benchmark for face editing.** We formulate face editing reliability through identity preservation and edit effectiveness, and introduce FaceEdit-Bench with separate automated annotations for approximately 110K edits across 15 models and 20 instruction types.
- **Systematic analysis of face editing reliability.** We reveal divergent model rankings and category-dependent performance across the two dimensions, demonstrating that instruction fulfillment alone is insufficient to assess face editing reliability.
- **Learned reliability evaluators of face editing.** We introduce FaceEditEval, which achieves stronger agreement with reference scores than the evaluated baselines under source-available and source-unavailable settings, including on unseen editors.

2 Related Work

Benchmarks for face generation and editing. Existing benchmarks primarily emphasize perceptual quality and edit correctness, while identity preservation is either treated as part of a broader fidelity score or examined only within a restricted task setting. F-Bench (Liu et al., 2025a) evaluates

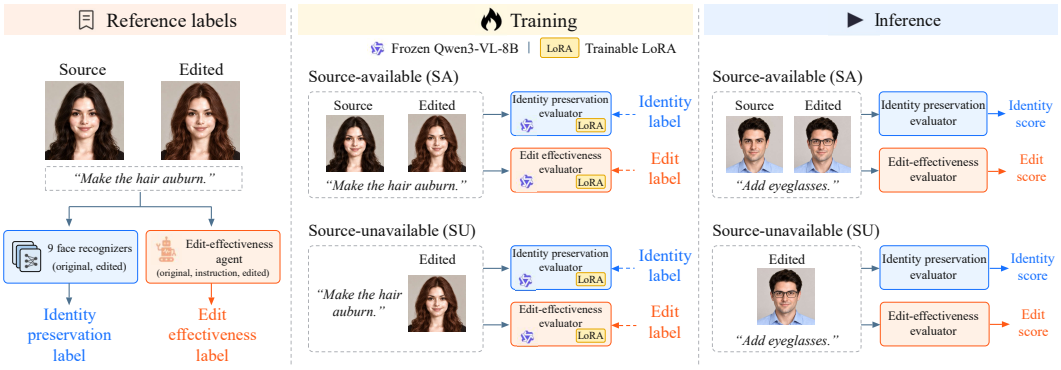


Figure 2: Overview of the annotation pipeline and FaceEditEval for assessing identity preservation and edit effectiveness.

face generation, customization, and restoration on the FaceQ dataset using human judgments of visual quality, authenticity, identity fidelity, and text–image correspondence, but does not specifically address paired instruction-guided face editing. FED-Bench (Xue et al., 2026) evaluates fidelity, instruction alignment, and expression-change magnitude using curated source–instruction–reference triplets, but is limited to facial expression editing and does not isolate identity preservation as an independent evaluation dimension. Our FaceEdit-Bench instead covers 20 instruction types across six categories and separately evaluates identity preservation and edit effectiveness, enabling their trade-off to be analyzed explicitly.

Face image quality assessment. Face image quality assessment (FIQA) estimates the utility of a face image for downstream recognition and is therefore related to, but distinct from, identity preservation. Representative methods derive quality from relative classifiability (Boutros et al., 2023), recognition scores against high-quality references (Hernandez-Ortega et al., 2019), embedding magnitude (Meng et al., 2021), or intra- and inter-class similarity distributions (Ou et al., 2021). A complementary line of work focuses on facial degradations and visual quality (Jo et al., 2023; Sun et al., 2025). Although higher recognition utility may yield more reliable identity features, these methods score each image independently rather than comparing an edited face with its source. Consequently, a high-quality edited face may receive a high FIQA score even when its identity has changed, motivating an explicit source–edit evaluation of identity preservation.

3 FaceEdit-Bench

FaceEdit-Bench establishes a unified framework for evaluating reliability in text-guided face editing through two complementary dimensions: identity preservation and edit effectiveness. By considering both dimensions, it enables more comprehensive comparisons of editing models and reveals limitations that either dimension alone may overlook. The following subsections present the data collection process, editing taxonomy, annotation methodology, and benchmark analysis.

3.1 Dataset Collection

FaceEdit-Bench comprises approximately 110K annotated edits generated from 10,000 source images: 2,000 from CelebAMask-HQ (Lee et al., 2020) and 8,000 from LAION-Face (Zheng et al., 2022). These two sources supply both high-quality studio portraits and in-the-wild face photographs, ensuring diversity in pose, lighting, and resolution.

The benchmark incorporates outputs from 15 editing models, including 10 open-source and five closed-source systems. The open-source models comprise Stable Diffusion v1.5 (Rombach et al., 2022), InstructPix2Pix (Brooks et al., 2023), Kandinsky 2.2 (Shakhmatov et al., 2023), SDXL (Podell et al., 2024), UltraEdit (SD3) (Zhao et al., 2024), BAGEL (Deng et al., 2025), FLUX.1 Kontext (Labs et al., 2025), OmniGen-v1 (Xiao et al., 2025), Qwen-Image-Edit (Wu et al., 2025), and Step1X-

Edit (Liu et al., 2025b). Each model is applied to all 10,000 source images, yielding approximately 100K edits in total.

The closed-source models comprise GPT-Image-2 (OpenAI, 2026), Kling IMAGE 3.0 (Kling AI, 2026), Nano Banana 2 (Fortin & Alessio, 2026), Seedream 5.0 Lite (ByteDance Seed, 2026), and Wan 2.7 Image Pro (Alibaba Cloud, 2026). Each is applied to a subset of 2,000 source images, contributing approximately 10K additional edits as a held-out test set.

Every output receives separate identity-preservation and edit-effectiveness annotations following the procedures described in Section 3.3. Detailed model descriptions are provided in Appendix A.

3.2 Editing Taxonomy

Text-guided face editing can modify facial features, the subject's appearance, the surrounding scene, or the image's visual style. To reflect these different scopes, we organize the 20 instruction types into six categories according to the main target and nature of the requested change: facial attributes, appearance, expression, accessories, background, and style. As illustrated in Figure 4(a), facial attributes concern local facial features, while appearance covers changes such as hair and facial hair. Expression concerns facial configuration and emotional state, and accessories involve objects worn by the subject. Background and style cover changes to the surrounding scene and the image's overall rendering, respectively. Together, these categories span edits to individual facial features, the subject's overall appearance, and the broader visual context, covering both localized changes and transformations of the whole image.

3.3 Reliability Annotation

We annotate identity preservation and edit effectiveness separately. For identity preservation, scores from multiple face recognizers are aggregated to limit the influence of model-specific errors. For edit effectiveness, an annotation agent breaks the instruction into individual requirements and uses visual analysis tools to check whether each has been met. The two procedures are described below.

Identity preservation. Given an original face x^o and an edited face x^e, we measure identity similarity using nine pretrained face recognition models implemented with InsightFace and DeepFace (Deng et al., 2019; Serengil & Ozpinar, 2020). Their identity-oriented representations provide a basis for comparing the subject before and after editing. Since models can respond differently to changes in facial appearance, we aggregate their scores to reduce dependence on any single model. Let $\tilde{s}_j(x^o, x^e) \in [0, 1]$ denote the normalized similarity from the j-th model. The identity-preservation target is

$$s^*_{\mathrm{id}} = \frac{1}{M} \sum_{j=1}^{M} \tilde{s}_j(x^o, x^e), \qquad M = 9. \tag{1}$$

Each model contributes equally, and the resulting score lies in $[0, 1]$, with higher values indicating stronger identity similarity. Appendix C details the recognition models and scores.

Edit effectiveness. Instruction fulfillment depends on the specific change requested, including its location, extent, and appearance. The annotation agent therefore decomposes instruction p into atomic visual requirements $\mathcal{A}(p) = \{a_1, \ldots, a_K\}$ that can be checked individually, as shown in Figure 3. For each requirement, it identifies the attribute type and target region, then selects relevant visual analysis tools. These provide region-level color statistics, facial landmark changes, region–text similarity, and focused comparisons of source and edited crops. This procedure directs assessment toward the requested modification rather than the presence of a target attribute alone.

Qwen3-VL-32B-Instruct (Bai et al., 2025) then reviews both images, the instruction, and the tool observations to assign a binary satisfaction decision z_k to each requirement. Tool observations inform the judgment but are not treated as ground truth. For outputs containing a visible human face and $K > 0$, the edit-effectiveness target is

$$s^*_{\mathrm{edit}} = \frac{1}{K} \sum_{k=1}^{K} z_k, \qquad z_k \in \{0, 1\}. \tag{2}$$

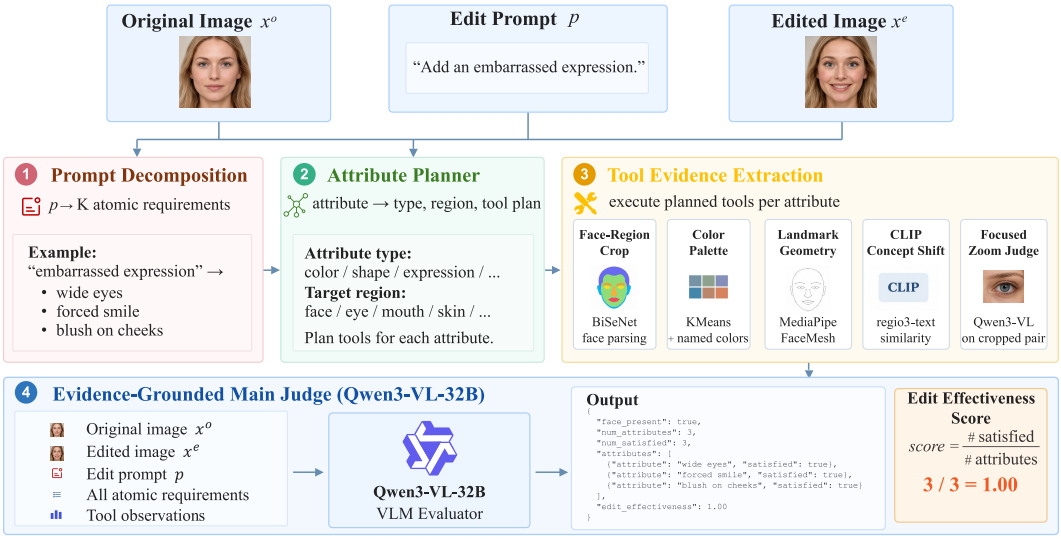


Figure 3: Agent-based annotation of edit effectiveness. The agent decomposes the editing instruction, collects tool evidence, and evaluates the fulfillment of each requirement.

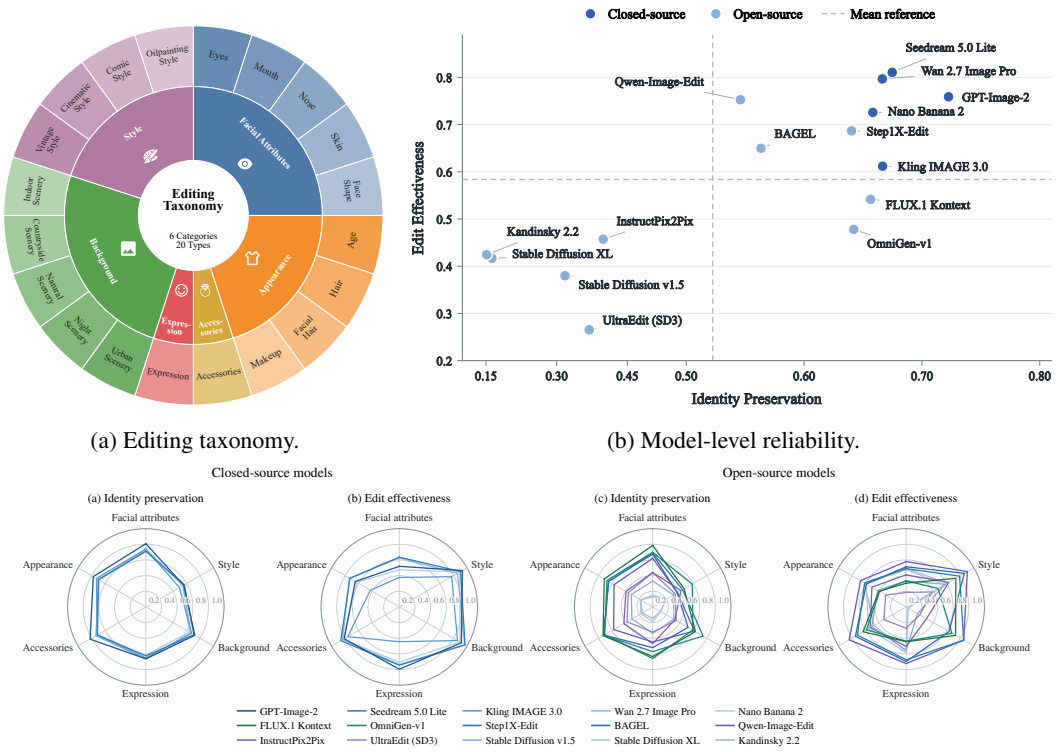
The final score ranges from 0 to 1 and reflects the proportion of instruction requirements satisfied, with equal weight assigned to each requirement. Appendix D provides the tool details and annotation prompts.

3.4 Benchmark Analysis

We compare all 15 editing models on edits generated from the shared set of 2,000 source images under the same editing instructions. We examine how model rankings differ between the two reliability dimensions and how performance varies across editing categories.

Model-level reliability. Figure 4(b) shows that proprietary models generally occupy the high-performing region for both identity preservation and edit effectiveness, whereas open-source models exhibit a wider performance spread. Nevertheless, the best-performing model depends on the evaluation dimension: GPT-Image-2 achieves the strongest identity preservation, while Seedream 5.0 Lite and Wan 2.7 Image Pro show stronger edit effectiveness. A similar ranking inconsistency appears among open-source models. Qwen-Image-Edit follows editing instructions effectively but preserves identity less reliably, while FLUX.1 Kontext exhibits the opposite tendency. Step1X-Edit provides a more balanced result across the two dimensions and lies closer to the proprietary models. The earlier diffusion-based baselines generally perform less favorably, suggesting that recent instruction-aware models improve not only edit execution but also the balance between editing and content preservation. Overall, these results demonstrate that no single dimension adequately characterizes model reliability: evaluating only edit effectiveness may favor aggressive edits, whereas identity preservation alone may reward insufficient modifications.

Category-wise reliability. Figure 4(c) further demonstrates that the relationship between identity preservation and edit effectiveness varies across editing categories. Style editing achieves the strongest edit effectiveness but the weakest identity preservation, while facial attribute editing shows the reverse tendency. This contrast is consistent with the scope of the requested modification. Style transformations typically introduce global changes to texture, color, and appearance, which make the edit visually salient but can also alter identity-relevant facial cues. Facial attribute edits are more localized and therefore preserve the overall identity more effectively, although precisely executing such local changes remains challenging. The category profiles also indicate that proprietary models are generally more consistent across editing types, whereas open-source models show larger category-dependent variation. These results suggest that aggregate scores can obscure important failure modes and motivate reporting both reliability dimensions at the category level.



(a) Editing taxonomy.

(b) Model-level reliability.

(c) Category-wise reliability.

Figure 4: Editing taxonomy and analysis on FaceEdit-Bench. (a) Six editing categories covering 20 editing types. (b) Overall identity preservation and edit effectiveness across 15 editing models. (c) Category-wise scores for both dimensions, grouped by closed-source and open-source models.

4 FACEEDITEVAL

The annotation procedures in FaceEdit-Bench provide complementary but computationally intensive supervision: identity preservation is derived from an ensemble of face recognizers, whereas edit effectiveness is assessed by an instruction-conditioned agent with tool-derived evidence. We introduce **FaceEditEval**, which transfers these heterogeneous supervision signals into a scalable learned evaluator by adapting a shared pretrained vision-language backbone.

Dimension-specific supervision. For each triplet consisting of a source image x^o, an editing instruction p, and an edited image x^e, FaceEdit-Bench provides two reference scores, s^*_{id} and s^*_{edit}. The former consolidates identity evidence from multiple face recognition systems, while the latter captures the fulfillment of instruction-specific visual requirements. We train a separate evaluator for each dimension so that the model learns the visual evidence relevant to its assigned criterion, instead of optimizing an ambiguous aggregate notion of editing quality.

Source-aware evaluation. The source image is not always accessible in practical evaluation pipelines. We therefore consider two complementary settings: **source-available (SA)**, which directly compares the source and edited images, and **source-unavailable (SU)**, which operates only on the instruction and edited result. For each dimension $d \in \{\text{id}, \text{edit}\}$,

$$\begin{aligned} \hat{s}^{\text{SA}}_d &= f^{\text{SA}}_d(x^o, p, x^e), \\ \hat{s}^{\text{SU}}_d &= f^{\text{SU}}_d(p, x^e). \end{aligned} \tag{3}$$

The SA setting supports explicit source-edit comparison, whereas the SU setting learns to estimate reliability from the edited output and instruction alone. In particular, SU identity preservation is an estimate rather than a direct verification of whether the original subject is retained.

Parameter-efficient adaptation. We formulate score prediction as conditional response generation. Each reference score is serialized as a textual target and used to fine-tune the corresponding evaluator with LoRA (Hu et al., 2021). Let $u_{d,m}$ denote the complete input for dimension d under setting $m \in \{\mathrm{SA}, \mathrm{SU}\}$, and let $y_d = (y_{d,1}, \ldots, y_{d,T})$ denote the tokenized target response. The training objective is

$$\mathcal{L}_{d,m} = -\mathbb{E}_{(u_{d,m}, y_d) \sim \mathcal{D}_{d,m}} \left[\sum_{t=1}^{T} \log P_{\theta_0, \phi_{d,m}} \left(y_{d,t} \mid u_{d,m}, y_{d,<t} \right) \right], \tag{4}$$

where θ_0 denotes the frozen backbone parameters and $\phi_{d,m}$ denotes the trainable LoRA parameters for the dimension-setting pair. The loss is applied only to the target-response tokens. This lightweight specialization retains a common visual-language representation while allowing each evaluator to focus on its corresponding reliability signal. Training configurations are provided in Appendix E.

5 Experiments

5.1 Experimental Setup

Baselines. We compare FaceEditEval with three families of baselines. General-purpose IQA methods include HyperIQA (Su et al., 2020), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), CLIP-IQA (Wang et al., 2023), LIQE (Zhang et al., 2023b), and TOPIQ-NR (Chen et al., 2024). FIQA methods include FaceQnet (Hernandez-Ortega et al., 2019; 2020), MagFace (Meng et al., 2021), SDD-FIQA (Ou et al., 2021), CR-FIQA (Boutros et al., 2023), IFQA (Jo et al., 2023), and Efficient-FIQA (Sun et al., 2025). These image-quality baselines operate on the edited image alone and are not conditioned on the editing instruction or source image.

We additionally evaluate three pretrained VLMs: InternVL3-9B-Instruct (Zhu et al., 2025), Qwen3-VL-8B-Instruct, and Qwen3-VL-32B-Instruct (Bai et al., 2025). Each VLM predicts both reliability dimensions without task-specific fine-tuning and is evaluated under both source-available (SA) and source-unavailable (SU) settings. Implementation details and prompts are provided in Appendix B.

Training and evaluation protocol. FaceEdit-Bench contains approximately 110K edited images generated from 10,000 source images. The training set comprises approximately 80K edits of 8,000 source images produced by the 10 open-source editors. The remaining 2,000 source images are shared by all 15 editors and yield approximately 30K test edits. We divide them into an open-source subset of approximately 20K edits and a proprietary subset of approximately 10K edits. Because outputs from proprietary editors are excluded from training, the proprietary subset evaluates out-of-distribution generalization to unseen editors. All four FaceEditEval variants are initialized from Qwen3-VL-8B-Instruct and adapted with LoRA. We report SRCC and PLCC separately for identity preservation and edit effectiveness on both the open-source and proprietary test subsets.

5.2 Main Results

Can image-quality metrics measure editing reliability? Table 1 shows that general-purpose IQA and FIQA methods have limited agreement with the reference scores. MagFace provides the strongest identity-preservation signal among these image-only baselines, but correlates poorly with edit effectiveness. More broadly, none of these methods performs consistently across identity preservation, edit effectiveness. This confirms that image or face quality alone is insufficient for determining whether an edit fulfills the instruction while retaining the original identity.

How well do VLMs assess editing reliability? Under the SU setting, VLMs predict reliability from the editing instruction and edited image alone. Their correlations are generally higher for edit effectiveness than for identity preservation, suggesting that instruction fulfillment is easier to assess without the source image. The SA setting additionally provides the original image. For both Qwen3-VL models, source access consistently improves edit-effectiveness prediction on the open-source and proprietary test subsets, but its effect on identity preservation is mixed. InternVL3-9B likewise shows no consistent improvement from SU to SA. These results indicate that source access alone does not enable reliable source-edit comparison without task-specific adaptation. Increasing the model size from Qwen3-VL-8B to Qwen3-VL-32B also provides no consistent benefit.

Table 1: Correlation with FaceEdit-Bench reference scores, reported as SRCC / PLCC. Open-source and Proprietary denote the two test subsets. Overall Avg. is computed across the two evaluation dimensions and the two test subsets. **Bold** and underlined indicate the best and second-best results, respectively.

Category	Method	Identity preservation		Edit effectiveness		Overall Avg.
		Open-source	Proprietary	Open-source	Proprietary	
IQA	HyperIQA	0.146 / 0.121	0.083 / 0.076	0.062 / 0.066	0.107 / 0.104	0.100 / 0.092
	MUSIQ	0.067 / 0.010	0.065 / 0.051	0.095 / 0.126	0.032 / 0.019	0.065 / 0.052
	MANIQA	0.178 / 0.158	0.095 / 0.134	0.050 / 0.063	0.070 / 0.069	0.098 / 0.106
	CLIP-IQA	0.081 / 0.063	0.138 / 0.127	0.064 / 0.102	0.143 / 0.139	0.107 / 0.108
	LIQE	0.101 / 0.082	0.013 / 0.021	0.023 / 0.032	0.064 / 0.053	0.050 / 0.047
	TOPIQ-NR	0.204 / 0.179	0.084 / 0.098	0.047 / 0.066	0.113 / 0.108	0.112 / 0.113
FIQA	FaceQnet	0.035 / 0.026	0.112 / 0.095	0.036 / 0.027	0.038 / 0.034	0.055 / 0.046
	MagFace	0.644 / 0.616	0.379 / 0.320	0.054 / 0.066	0.191 / 0.168	0.317 / 0.293
	SDD-FIQA	0.116 / 0.182	0.189 / 0.222	0.008 / 0.006	0.068 / 0.064	0.095 / 0.119
	CR-FIQA	0.034 / 0.107	0.112 / 0.184	0.022 / 0.043	0.006 / 0.028	0.044 / 0.091
	IFQA	0.073 / 0.075	0.032 / 0.030	0.051 / 0.055	0.037 / 0.037	0.048 / 0.049
	Efficient-FIQA	0.112 / 0.116	0.186 / 0.187	0.101 / 0.116	0.070 / 0.065	0.117 / 0.121
VLM	InternVL3-9B (SU)	0.188 / 0.116	0.136 / 0.093	0.336 / 0.359	0.492 / 0.523	0.288 / 0.273
	InternVL3-9B (SA)	0.054 / 0.012	0.141 / 0.119	0.406 / 0.419	0.456 / 0.502	0.264 / 0.263
	Qwen3-VL-8B (SU)	0.164 / 0.137	0.163 / 0.153	0.358 / 0.370	0.485 / 0.509	0.292 / 0.292
	Qwen3-VL-8B (SA)	0.278 / 0.294	0.137 / 0.118	0.573 / 0.582	0.574 / 0.654	0.391 / 0.412
	Qwen3-VL-32B (SU)	0.204 / 0.182	0.173 / 0.192	0.362 / 0.415	0.510 / 0.556	0.312 / 0.336
	Qwen3-VL-32B (SA)	0.135 / 0.129	0.208 / 0.183	0.549 / 0.593	0.614 / 0.726	0.376 / 0.408
Ours	FaceEditEval (SU)	0.797 / 0.785	0.474 / 0.478	0.813 / 0.818	0.705 / 0.736	0.697 / 0.704
	FaceEditEval (SA)	**0.861 / 0.854**	**0.611 / 0.615**	**0.849 / 0.854**	**0.729 / 0.763**	**0.763 / 0.772**

How effective is FaceEditEval? FaceEditEval outperforms all IQA, FIQA, and VLM baselines for both identity preservation and edit effectiveness on the open-source and proprietary test subsets. FaceEditEval (SA) achieves the best overall performance, with average SRCC and PLCC values of 0.763 and 0.772, compared with 0.391 and 0.412 for the strongest VLM baseline. FaceEditEval (SU) remains the second-best method despite operating without the source image. The larger improvement from SU to SA for identity preservation highlights the importance of directly comparing the source and edited faces.

Does FaceEditEval generalize to unseen proprietary editors? FaceEditEval is trained only on edits from open-source models, making the proprietary test subset an out-of-distribution evaluation. Although correlations decrease on this subset, FaceEditEval remains stronger than all evaluated baselines. The decrease is more pronounced for source-unavailable identity preservation than for edit effectiveness, suggesting that instruction-related supervision transfers more reliably across editor families than source-free identity estimation. Providing the source image reduces this generalization gap, demonstrating the benefit of explicit source-edit comparison for unseen editors.

5.3 Human Study

To complement correlation-based evaluation, we conduct a human study following the group maximum differentiation (gMAD) principle (Ma et al., 2016). Using the predictions of FaceEditEval and the evaluated baselines, we select 200 edited-output pairs that exhibit the largest ranking disagreements. Each pair contains two outputs generated from the same source image and editing instruction. This gMAD-based protocol focuses the evaluation on challenging cases that are most informative for distinguishing competing methods.

Human annotators independently select the output that better preserves the source identity and the output that more faithfully fulfills the editing instruction. We measure each method's accuracy by comparing its pairwise predictions with the human preferences. As reported in Table 2, FaceEditEval (SA) achieves the highest agreement with human judgments, reaching 84.0% accuracy for identity preservation and 76.0% for edit effectiveness. These results exceed the strongest corresponding

Table 2: Human preference accuracy (%) on the gMAD-selected pairs. SU/SA denote source-unavailable/source-available evaluation. "–" indicates not applicable.

Category	Method	Identity preservation		Edit effectiveness	
		SU	SA	SU	SA
FIQA	CR-FIQA	56.0	–	49.5	–
	FaceQnet	54.0	–	41.5	–
	MagFace	65.0	–	45.5	–
	SDD-FIQA	57.0	–	51.5	–
VLM	InternVL3-9B	32.0	36.0	34.0	54.0
	Qwen3-VL-8B	34.0	42.0	34.0	61.0
	Qwen3-VL-32B	38.0	44.0	40.0	62.0
Ours	FaceEditEval	**74.0**	**84.0**	**59.0**	**76.0**

baselines by 19.0 and 14.0 percentage points, respectively. The results confirm that FaceEditEval reliably captures both dimensions of editing reliability on challenging cases where the evaluated methods disagree.

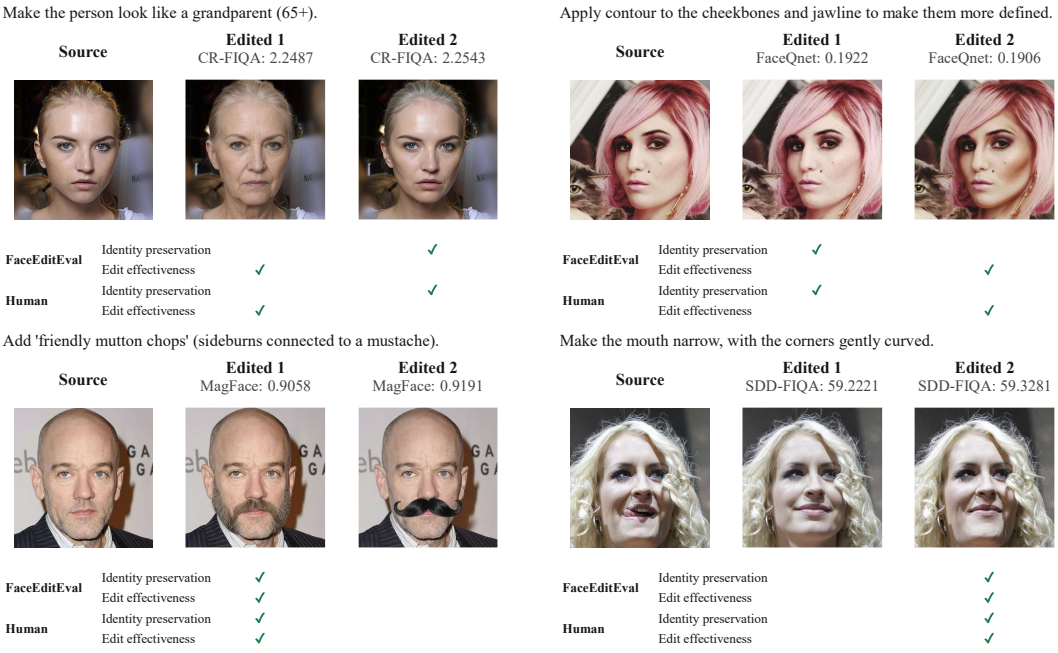


Figure 5: Human study examples where each edited pair receives similar scores from the indicated FIQA method. Checkmarks show agreement between FaceEditEval and human preferences for identity preservation and edit effectiveness.

6 Conclusion

We introduced FaceEdit-Bench, a benchmark of approximately 110K annotated face edits for assessing identity preservation and edit effectiveness. Our analysis of 15 editing models shows that stronger instruction fulfillment does not necessarily imply better identity preservation. We also developed FaceEditEval, which achieves stronger correlations with automated reference scores than the evaluated IQA, FIQA, and pretrained VLM baselines under both SA and SU settings. A pairwise human study provides further evidence of agreement with human preferences. Together, FaceEdit-Bench and FaceEditEval support systematic evaluation of face editing reliability by separately assessing instruction fulfillment and identity preservation.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and language polishing, and to explore figure design ideas. AI-based image generation and annotation used in this study are described in Section 3. The authors take responsibility for the final manuscript, figures, and reported results.

ETHICS STATEMENT

This work uses existing face datasets, with data use and release subject to their original usage terms and privacy constraints.

REPRODUCIBILITY STATEMENT

The main text and appendix describe the dataset composition, annotation procedures, data splits, and training and testing settings. Together, these details document how the benchmark is constructed and how the evaluators are trained and assessed to support reproducibility.

REFERENCES

Alibaba Cloud. Wan-Image Generation and Editing 2.7 API Reference. Alibaba Cloud Model Studio documentation, 2026. URL https://www.alibabacloud.com/help/en/model-studio/wan-image-generation-and-editing-api-reference. Accessed September 20, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

ByteDance Seed. Seedream 5.0 Lite. Official model page, 2026. URL https://seed.bytedance.com/en/seedream5_0_lite. Accessed September 20, 2026.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

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*, pp. 4690–4699, 2019.

Alisa Fortin and Bea Alessio. Build with Nano Banana 2, our best image generation and editing model. Google Blog, 2026. URL https://blog.google/innovation-and-ai/technology/developers-tools/build-with-nano-banana-2/. Accessed September 20, 2026.

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)*, pp. 1–8. IEEE, 2019.

Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *arXiv preprint arXiv:2006.03298*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Byungho Jo, Donghyeon Cho, In Kyu Park, and Sungeun Hong. Ifqa: Interpretable face quality assessment. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3433–3442. IEEE, 2023.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Kling AI. Kling IMAGE 3.0 Model User Guide. Official model documentation, 2026. URL `https://kling.ai/quickstart/klingai-image-3-model-user-guide`. Accessed September 20, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5548–5557. IEEE, 2020.

Lu Liu, Huiyu Duan, Qiang Hu, Liu Yang, Chunlei Cai, Tianxiao Ye, Huayu Liu, Xiaoyun Zhang, and Guangtao Zhai. F-bench: Rethinking human preference evaluation metrics for benchmarking face generation, customization, and restoration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10982–10994. IEEE, 2025a.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group mad competition-a new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1664–1673, 2016.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14220–14229. IEEE, 2021.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012b.

OpenAI. GPT-Image-2. OpenAI API documentation, 2026. URL `https://developers.openai.com/api/docs/models/gpt-image-2`. Accessed September 20, 2026.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7666–7675. IEEE, 2021.

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. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (ASYU)*, pp. 1–5. IEEE, 2020.

Arseniy Shakhmatov, Anton Razzhigaev, Aleksandr Nikolich, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 2.2, 2023.

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)*, pp. 3664–3673. IEEE, 2020.

Wei Sun, Weixia Zhang, Linhan Cao, Jun Jia, Xiangyang Zhu, Dandan Zhu, Xiongkuo Min, and Guangtao Zhai. Efficient face image quality assessment via self-training and knowledge distillation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3394–3402. IEEE, 2025.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304. IEEE, 2025.

Fengjian Xue, Xuecheng Wu, Heli Sun, Yunyun Shi, Shi Chen, Liangyu Fu, Jinheng Xie, Dingkang Yang, Hao Wang, Junxiao Xue, et al. Fed-bench: A cross-granular benchmark for disentangled evaluation of facial expression editing. *arXiv preprint arXiv:2603.29697*, 2026.

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. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199. IEEE, 2022.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European conference on computer vision*, pp. 334–349. Springer, 2018.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in neural information processing systems*, 36:31428–31449, 2023a.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081. IEEE, 2023b.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18676–18688. IEEE, 2022.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

A FaceEdit-Bench Details

The benchmark includes edited images from ten open-source and five closed-source editing models. We summarize each model and dataset details below.

A.1 Open-Source Editing Models

FLUX.1 Kontext. FLUX.1 Kontext (Labs et al., 2025) uses flow matching to integrate text and image context within a unified generation and editing model. It encodes reference images into latent tokens and concatenates them with the target image tokens, allowing the transformer to condition generation on the supplied visual context. This sequence-based formulation accommodates differences in resolution and aspect ratio between the reference and output images. The model supports local modifications, broader contextual transformations, and successive edits that reuse previous outputs as visual references. We evaluate the publicly released [dev] variant.

OmniGen-v1. OmniGen (Xiao et al., 2025) is a unified diffusion model that accepts interleaved text and image inputs. Its architecture combines a variational autoencoder for image representation with a transformer initialized from Phi-3. Text tokens and image latents are processed within the same network, allowing interactions between the editing instruction and visual content without separate task-specific conditioning modules. Multitask training enables the model to handle image editing, subject-driven generation, and other visual conditioning tasks through a shared input format. Users specify the desired operation through the multimodal instruction rather than selecting a dedicated model for each task.

Step1X-Edit. Step1X-Edit (Liu et al., 2025b) combines a multimodal language model with a diffusion image decoder. The language model jointly interprets the reference image and editing instruction to extract semantic information relevant to the requested modification. A lightweight connector transforms these representations into conditioning features for the diffusion transformer. In parallel, a variational autoencoder encodes the reference image into latent tokens, which are concatenated with the noisy target image tokens. These complementary inputs provide semantic guidance for the edit while retaining access to the source image's visual details.

BAGEL. BAGEL (Deng et al., 2025) is a unified multimodal model supporting both understanding and generation. It adopts a mixture-of-transformer-experts architecture with separate parameters for understanding and image generation. The understanding expert processes text and visual semantic tokens, whereas the generation expert processes image latents. Pretraining on interleaved text, image, video, and web data enables the model to learn relationships across modalities and contextual sequences. This shared framework supports instruction-driven image manipulation alongside visual reasoning, subject-conditioned generation, and other multimodal tasks.

Qwen-Image-Edit. Qwen-Image-Edit (Wu et al., 2025) extends Qwen-Image to instruction-based editing. It processes the input image through both Qwen2.5-VL for semantic conditioning and a VAE encoder for appearance conditioning. The semantic pathway supports changes such as object rotation and style transformation, while the appearance pathway supplies visual details for localized modifications. The model also supports adding, removing, and replacing Chinese and English text within images, with attention to the existing typography and layout. Its editing capabilities cover changes to image content, clothing, backgrounds, visual style, and rendered text.

InstructPix2Pix. InstructPix2Pix (Brooks et al., 2023) learns image editing from synthetic original-image, instruction, and edited-image triplets. Its data construction pipeline combines a language model that produces editing instructions with a text-to-image model that generates corresponding image pairs. The editor is initialized from Stable Diffusion and conditioned jointly on the source image and the editing instruction. Separate image and text guidance scales control the balance between retaining the input appearance and following the requested change. At inference time, the model directly generates the edited result without per-image fine-tuning or inversion.

UltraEdit. UltraEdit (Zhao et al., 2024) develops instruction-based image editors using a large automatically constructed dataset of approximately four million editing examples. The data construction process starts from real photographs and artworks, providing diverse visual content for learning edits. Editing instructions are generated using large language models guided by human-provided examples, broadening the range of requested operations. Automatically generated region annotations additionally support localized editing with explicit spatial constraints. Diffusion-based editors trained on these data support both general image modifications and region-based changes.

Stable Diffusion v1.5. Stable Diffusion v1.5 belongs to the latent diffusion family (Rombach et al., 2022), which performs text-conditioned denoising in a compressed image representation. An autoencoder maps images into a latent space, where a U-Net iteratively predicts and removes noise. Text representations guide this process through cross-attention, and the resulting latents are decoded into an output image. Operating in latent space reduces the computational cost compared with diffusion directly over high-resolution pixels. We include it as a general image-generation baseline alongside models trained specifically for instruction-based editing.

Stable Diffusion XL. SDXL (Podell et al., 2024) extends latent diffusion with a larger denoising backbone and richer text conditioning for high-resolution synthesis. Its U-Net contains more attention blocks, and two text encoders provide complementary representations of the input prompt. Additional conditioning on image size and crop coordinates, together with training across multiple aspect ratios, gives the model greater control over output composition. The framework also includes an optional refinement model that further improves the visual detail of generated images. SDXL provides a higher-capacity general diffusion baseline for comparison with specialized editing models.

Kandinsky 2.2. Kandinsky 2.2 (Shakhmatov et al., 2023) combines a text-conditioned image embedding prior with a latent diffusion decoder. The prior maps text representations into the image embedding space, and the decoder synthesizes an image conditioned on the resulting semantic representation. Version 2.2 adopts a CLIP ViT-G image encoder and uses a MoVQ autoencoder to represent images in the decoder's latent space. Its released components support text-to-image generation, image-to-image generation, inpainting, and depth-conditioned generation through ControlNet. This architecture provides an image-embedding conditioning mechanism distinct from the text conditioning used by the Stable Diffusion baselines.

A.2 Closed-Source Editing Models

GPT-Image-2. GPT-Image-2 (OpenAI, 2026) is OpenAI's image generation and editing model, accepting text and image inputs. For editing, a natural-language instruction specifies the desired modification while the input image provides visual context. The model automatically processes image inputs at high fidelity, supporting edits that require attention to source-image details. It also supports flexible output dimensions and aspect ratios, along with configurable generation quality. These capabilities allow the same model to serve both prompt-based image creation and reference-conditioned editing.

Seedream 5.0 Lite. Seedream 5.0 Lite (ByteDance Seed, 2026) is ByteDance's unified multimodal image generation and editing model. It combines visual understanding with reasoning to interpret editing requests in the context of the supplied images. Its editing capabilities include handling underspecified instructions and using reference-image features to guide the requested transformation. The model also supports multi-step visual reasoning and applies world knowledge to the content and relationships depicted in generated images. An additional online retrieval capability provides access to current information for generation tasks that require it.

Kling IMAGE 3.0. Kling IMAGE 3.0 (Kling AI, 2026) supports image generation and reference-based editing through natural-language instructions. Its editing interface covers object addition, removal, attribute modification, and restyling. Multiple reference images can supply complementary subject, appearance, and style information for the output. For portraits, supported operations include changes to hairstyles, expressions, accessories, lighting, and backgrounds. The interface also accepts visual guidance such as selections, arrows, and sketches to specify the location or form of an edit.

Wan 2.7 Image Pro. Wan 2.7 Image Pro (Alibaba Cloud, 2026) is Alibaba's commercial image generation and editing model. Its multimodal interface supports text-to-image generation, editing of input images, and generation conditioned on multiple image references. Text and reference images can be supplied together to specify the desired content and visual characteristics of the output. The interface additionally supports interactive editing and the generation of image sets. The Pro variant supports up to 4K output for text-to-image generation, while image editing and image-set generation support up to 2K resolution.

Nano Banana 2. Nano Banana 2 (Fortin & Alessio, 2026), also named Gemini 3.1 Flash Image, is Google's image generation and editing model. It combines text and image conditioning for prompt-driven visual creation and modification. The model supports maintaining character resemblance and object appearance across related outputs, enabling repeated edits within a shared visual context. It also provides text rendering and translation capabilities for modifying written content within images. Its Flash-based design emphasizes rapid generation and iteration, with support for multiple output resolutions and aspect ratios.

A.3 Dataset Details

Public datasets. CelebAMask-HQ (Lee et al., 2020) contains 30,000 high-resolution face images derived from CelebA, with manually annotated masks covering 19 semantic classes, including facial components, hair, and accessories. LAION-Face (Zheng et al., 2022) is a face-containing subset of LAION-400M, comprising approximately 50 million image–text pairs collected from the web for large-scale facial representation learning. Together, these datasets provide high-resolution portraits and web-sourced face imagery for constructing our source pool.

Source images. The source pool contains 10,000 face images: 2,000 from CelebAMask-HQ (Lee et al., 2020) and 8,000 from LAION-Face (Zheng et al., 2022). The training set uses 8,000 source images, comprising 1,600 from CelebAMask-HQ and 6,400 from LAION-Face. The test set uses 2,000 source images, comprising 400 from CelebAMask-HQ and 1,600 from LAION-Face.

Gender and skin-tone balance. The source images are balanced by gender and skin tone, with equal numbers of male and female subjects and equal numbers of subjects with darker and lighter skin tones. This sampling strategy prevents either gender or skin-tone group from dominating the source pool.

B Evaluation Baselines

We compare FaceEditEval with eight general IQA methods, four recognition-oriented FIQA methods, two perceptual FIQA methods, and three pretrained VLMs. The IQA and FIQA baselines assess only the edited image, without access to the source image or editing instruction. The pretrained VLMs are evaluated under both SA and SU settings. We examine the association of their predictions with the two FaceEdit-Bench reference targets.

B.1 General Image Quality Assessment

BRISQUE. BRISQUE (Mittal et al., 2012a) extracts natural-scene statistics from locally normalized luminance and neighboring pixel relationships. It fits generalized Gaussian distributions to the normalized coefficients and their pairwise products, obtaining features that describe changes in local image structure. A support vector regression model trained on subjective quality scores maps these features to an overall quality estimate.

CLIP-IQA. CLIP-IQA (Wang et al., 2023) uses pretrained vision-language representations to assess image quality through comparisons with paired textual descriptions. The image embedding is compared with text embeddings representing opposing attributes, such as good versus poor quality. The corresponding cosine similarities are normalized into a score that expresses the image's relative alignment with the positive description. This prompt-pairing strategy enables zero-shot quality assessment without training a separate quality regressor.

HyperIQA. HyperIQA (Su et al., 2020) uses a hypernetwork to generate parameters for an image-quality prediction network according to image content. A content-understanding branch extracts semantic information, which guides the hypernetwork in adapting the quality predictor. The resulting target network processes distortion-related features to estimate the perceptual quality of the input image. This content-dependent adaptation helps account for diverse scenes and authentic image distortions.

LIQE. LIQE (Zhang et al., 2023b) formulates blind quality assessment through vision-language correspondence and multitask learning. It constructs textual descriptions that combine scene categories, distortion types, and quality levels, and compares their embeddings with visual features using CLIP. A numerical quality estimate is obtained by aggregating the probabilities assigned to different quality levels. Learning quality assessment alongside scene classification and distortion recognition allows semantic content and distortion information to inform quality prediction.

MANIQA. MANIQA (Yang et al., 2022) combines vision-transformer features with attention across channel and spatial dimensions. Its transposed attention blocks model dependencies across feature channels, while scale Swin transformer blocks capture spatial interactions between image regions. A dual-branch prediction head estimates a quality score and an importance weight for each patch. The final score is their weighted aggregation, accounting for the unequal contribution of image regions to perceptual quality.

MUSIQ. MUSIQ (Ke et al., 2021) is a multiscale image-quality transformer designed to process images with varying resolutions and aspect ratios. It extracts patches from the native-resolution image and resized versions, preserving fine details while providing broader context. A hash-based spatial embedding encodes patch locations, and a separate scale embedding identifies the resolution of each patch. The transformer integrates these tokens to capture quality cues at different spatial granularities.

NIQE. NIQE (Mittal et al., 2012b) measures the departure of an image's natural-scene statistics from a statistical model fitted to pristine images. Its quality score is a covariance-normalized distance between multivariate Gaussian models of the test-image features and pristine-image features, with larger distances indicating greater deviations from natural image statistics. It requires neither human quality ratings nor distorted-image examples to construct the pristine-image model.

TOPIQ-NR. TOPIQ (Chen et al., 2024) uses high-level semantic features to guide attention toward relevant low-level distortions. Its coarse-to-fine network extracts hierarchical features with a convolutional backbone and propagates information from higher to lower levels. Cross-scale attention uses semantic representations to emphasize local regions whose distortions matter for perceptual quality. We evaluate the no-reference variant, which estimates quality from the input image without a pristine counterpart.

B.2 Recognition-Oriented Face Image Quality Assessment

CR-FIQA. CR-FIQA (Boutros et al., 2023) predicts a face sample's relative classifiability from its position with respect to its own class center and the nearest competing class center in angular embedding space. During training, this relationship provides supervision for a quality prediction branch alongside the recognition objective. At inference time, the learned predictor estimates quality from a single face image without requiring its identity label or access to class centers.

FaceQnet. FaceQnet (Hernandez-Ortega et al., 2019; 2020) learns a single-image quality predictor using supervision derived from face recognition comparisons. Training targets are constructed by comparing each face image with a selected high-quality reference of the same identity. A ResNet-50-based regression network learns to predict these recognition-derived quality targets directly from image content. Once trained, it produces a scalar quality score without requiring a reference image at inference time.

MagFace. MagFace (Meng et al., 2021) learns face representations whose feature magnitudes encode sample quality through a magnitude-dependent recognition objective. Its adaptive angular margin and magnitude regularization encourage a structured embedding space in which more

recognizable samples have larger feature norms. At inference time, the feature magnitude before normalization provides a quality indicator, while the normalized embedding is used for identity comparison. Quality estimation therefore requires no separate regression network or human quality annotations.

SDD-FIQA. SDD-FIQA (Ou et al., 2021) constructs quality pseudo-labels from the Wasserstein distance between intra-class and inter-class similarity distributions. For each training image, these distributions summarize its similarities to images of the same identity and to images of other identities. A regression network learns the resulting pseudo-labels to estimate face-image quality without human quality ratings. Although label construction involves comparisons across images, inference requires only a single input face.

B.3 Perceptual Face Image Quality Assessment

IFQA. IFQA (Jo et al., 2023) is a face-oriented perceptual quality assessment method developed for evaluating restored facial images. It adopts an adversarial framework in which a generator simulates face restoration, while a per-pixel discriminator produces a spatial quality map for the input image. The method places greater emphasis on perceptually important facial regions, such as the eyes, nose, and mouth, and aggregates the pixel-level predictions into an overall quality score. This design enables IFQA to capture localized facial degradations while providing an interpretable visualization of their spatial distribution.

Efficient-FIQA. Efficient-FIQA (Sun et al., 2025) is a lightweight perceptual face-quality predictor based on self-training and knowledge distillation. It first trains a teacher model on labeled face images and then uses the teacher to assign pseudo-quality labels to additional unlabeled images. The pseudo-labeled samples are used both to refine the teacher and to transfer its quality-prediction capability to a compact student model. The resulting student model predicts a perceptual quality score from a single face image while substantially reducing the computational cost of inference.

In our evaluation, both methods receive only the edited image and produce a single perceptual face-quality score, without access to the source image or editing instructions.

B.4 Vision-Language Models

Qwen3-VL-8B-Instruct. Qwen3-VL-8B-Instruct (Bai et al., 2025) serves as a pretrained baseline for assessing identity preservation and edit effectiveness under both source-available (SA) and source-unavailable (SU) settings. It predicts scores from the supplied images and instructions without fine-tuning on the FaceEditEval training annotations. As the backbone of FaceEditEval, it also provides the unadapted baseline for evaluating the effect of fine-tuning.

Qwen3-VL-32B-Instruct. Qwen3-VL-32B-Instruct (Bai et al., 2025) provides a larger pretrained baseline within the same model family. It assesses both reliability dimensions under the SA and SU settings without task-specific fine-tuning. Comparison with the 8B variant allows us to examine whether increased model capacity improves agreement with the reference annotations.

InternVL3-9B-Instruct. InternVL3-9B-Instruct (Zhu et al., 2025) provides a pretrained baseline from a different vision-language model family. It evaluates identity preservation and edit effectiveness under both SA and SU settings, using the supplied visual inputs and scoring instructions. The model is evaluated without fine-tuning on the FaceEditEval training annotations.

C Identity-Preservation Annotation

Recognition models. Table 3 lists the nine recognition models used to construct the identity-preservation annotations: three from InsightFace and six from DeepFace (Deng et al., 2019; Serengil & Ozpinar, 2020). The ensemble includes multiple models from the same recognition family; for example, `Facenet` and `Facenet512` are two FaceNet models with different embedding dimensions.

Table 3: Recognition models used for identity annotation.

Library	Model	Architecture	Training data
InsightFace	`antelopev2`	ResNet100	Glint360K
	`buffalo_l`	ResNet50	WebFace600K
	`buffalo_s`	MobileFaceNet	WebFace600K
DeepFace	VGG-Face	VGG16	VGGFace
	Facenet	Inception-ResNet-v1	100-200M Face images
	Facenet512	Inception-ResNet-v1(512 dimensions)	100-200M Face images
	OpenFace	Inception-based CNN	CASIA-WebFace + FaceScrub
	DeepID	Four-layer CNN	YouTube Faces
	ArcFace	ResNet34	CASIA-WebFace

Score presentation. The identity-preservation annotations are stored on a $[0, 5]$ scale and used directly as training targets. For benchmark visualizations that display both dimensions on a common $[0, 1]$ scale, the identity scores are divided by five.

D EDIT-EFFECTIVENESS ANNOTATION

Tool observations. The annotation agent uses BiSeNet (Yu et al., 2018) for face parsing, MediaPipe FaceMesh (Kartynnik et al., 2019) for facial landmarks, and CLIP (Radford et al., 2021) for region–text similarity, together with color palette extraction and focused crop comparison. Table 4 summarizes their outputs.

Table 4: Tools and their outputs for edit-effectiveness annotation.

Tool	Output
BiSeNet	Facial-region masks
K-means color palettes	Dominant colors and their proportions
MediaPipe FaceMesh	Relative changes in facial landmark ratios
CLIP	Changes in region–text similarity
Focused crop comparison	Satisfaction decision and brief explanation

Face-presence check. In addition to requirement-level decisions, the main judge returns a `face_present` flag indicating whether the edited image contains at least one clearly visible human face. If the flag is false, edit effectiveness is set to zero regardless of the individual requirement decisions.

Prompt templates. Figures 6 and 7 present the instruction decomposition, attribute routing, main judgment, and focused crop judgment prompts. Braced placeholders are populated with the editing instruction, individual requirements, target regions, and collected observations. The main judge is instructed to treat tool observations as supporting evidence rather than ground truth. The crop judge assesses only the visible region and compares it with the source crop when the requirement is comparative. Both judges return structured JSON responses containing satisfaction decisions and brief explanations.

01 Instruction Decomposition Prompt

SYSTEM

You are an image-editing instruction analyzer. Given an instruction in natural language, decompose it into a list of ATOMIC, independently-verifiable visual requirements that the edited image is expected to satisfy. Output JSON only, no markdown, no extra text.

USER

Edit instruction: """{prompt}"""

Decompose this instruction into atomic visual requirements. Rules:

- Each item is ONE concrete, independently checkable visual change.
- Do not merge several changes into one item.
- Do not invent requirements that the instruction does not ask for.
- Keep each item short, plain, and visual (a sentence or noun phrase).
- If a requirement is comparative (e.g. "fuller lips", "narrower nose"), KEEP it comparative.
- Typical count is 1-5 items.

Return ONLY this JSON object:

```
{
  "attributes": ["requirement 1", "requirement 2", "..."]
}
```

02 Attribute Router Prompt

SYSTEM

You are an attribute-routing classifier for an image-editing evaluator. Given ONE visual requirement (a single short phrase), you classify its type and decide which face region it concerns. Output JSON only.

USER

Visual requirement to classify: """{attribute}"""

Classify the requirement into ONE type and ONE region.

TYPES: color, shape, object, region_style, expression, texture, global_style, unknown

REGIONS: face, skin, hair, eye, l_eye, r_eye, brow, nose, mouth, lip, ear, earring, glasses, neck, cloth, hat, background, full, beard, chin, cheek, head, hand

For region_style/object/color/texture, you must ALSO provide a SHORT natural-language concept (3-8 words) that an image-text similarity model could probe.

Output ONLY this JSON:

```
{
  "type": "<one of TYPES>",
  "region": "<one of REGIONS>",
  "concept": "<3-8 word probe phrase, or empty string>"
}
```

Figure 6: Prompt templates for instruction decomposition and attribute routing.

03 Main Judgment Prompt

You are a meticulous image-editing evaluator. You are shown two images: first the ORIGINAL, then the EDITED. You are given a list of atomic visual requirements and, for each, AUTOMATED TOOL OBSERVATIONS (color palette of the relevant region, geometric landmark deltas, CLIP-similarity shift, and/or a focused zoom judge from a second VLM pass on the cropped region). Use these observations as EVIDENCE, not as ground truth -- they can be wrong. Judge each requirement, then also report face_present. Answer with a single JSON object only.

Original edit instruction:
"""{prompt}"""

Pre-decomposed atomic requirements (copy each verbatim in your output):
{attrs_block}

Per-requirement automated tool observations (treat as advisory evidence):
{obs_block}

Tasks:
1. "face_present": true iff EDITED contains at least one clearly visible human face.
2. For each requirement, decide "satisfied" true/false using BOTH the visible image content AND the tool observations. The observations are summaries: colors[name(pct)] = top dominant colors in the region; geom[metric +/-pct%] = signed % change of a face landmark ratio; clip[d=+/-0.xx] = CLIP image-text cosine shift toward the concept; zoom=yes/no = focused VLM judgment on the region crop.

Output ONLY this JSON (copy each "attribute" verbatim):

```
{"face_present": true_or_false, "attributes": [
{"attribute": "<copy item 1>", "satisfied": true_or_false,
"reason": "<one sentence>"},
...
]}
```

04 Focused Crop Judgment Prompt

You are a meticulous image-editing evaluator focused on ONE attribute. You are shown two cropped images (ORIGINAL region first, EDITED region second) and a single visual requirement. Judge strictly and visually whether the requirement is satisfied in the EDITED crop, relative to the ORIGINAL crop when the requirement is comparative. Answer with a single JSON object and nothing else.

The two images above are CROPS of the {region} region of the same scene.
The first crop is from the ORIGINAL image; the second is from the EDITED image.

Single visual requirement to verify:
"""{attribute}"""

Judging rules:
- Mark "satisfied" true ONLY if the requested change is clearly visible in the EDITED crop.
- If the requirement is unchanged or wrong in the EDITED crop, mark false.
- Confine your judgment to what is visible in the crop; do not assume context outside the crop.

Output ONLY this JSON:

```
{"satisfied": true_or_false, "reason": "<one short sentence>"}
```

Figure 7: Prompt templates for the main judge and focused crop judgment.

E Evaluator Configuration and Evaluation Details

Training configuration. FaceEditEval is adapted with LoRA through `swift sft`. The vision encoder and multimodal aligner are frozen, and the adapters are applied to the language component's attention and feed-forward projections. Training uses bfloat16 computation, FlashAttention, and gradient checkpointing. Table 5 groups the settings by adapter, input, and training configuration.

Table 5: Configuration of FaceEditEval.

Parameter	Description	Value
LoRA configuration		
LoRA rank	Adapter rank	32
LoRA alpha	Adapter scaling parameter	64
LoRA dropout	Adapter dropout probability	0.05
Input configuration		
Maximum sequence length	Token limit per sequence	4,096
Minimum image tokens	Lower image token limit	128
Maximum image tokens	Upper image token limit	1,024
Training configuration		
Training epochs	Passes over training data	4
Learning rate	Parameter update step size	2×10^{-5}
Warmup ratio	Fraction of warmup steps	0.03
Per-device batch size	Samples per device per step	1
Gradient accumulation steps	Steps before each update	16
Effective batch size	Samples per optimizer update	64
Precision	Computation data type	bfloat16

Batch configuration. With four training processes, a per-device batch size of one, and 16 gradient accumulation steps, each optimizer update uses an effective batch of 64 examples.

Adapter placement. The adapted projections are `q_proj`, `k_proj`, `v_proj`, and `o_proj` in attention, and `gate_proj`, `up_proj`, and `down_proj` in the feed-forward layers. These module names specify the locations of the trainable LoRA adapters.

Input limits. The sequence-length limit applies to the complete tokenized input and target response. The image-token limits control the visual input budget for each image. In the SA setting, the source and edited images are both supplied; in the SU setting, only the edited image is supplied alongside the instruction.

Training targets. The identity-preservation and edit-effectiveness evaluators use separate numerical targets on $[0, 5]$ and $[0, 1]$, respectively. Each target is serialized as a text response. Supervised fine-tuning uses token-level cross-entropy over that response, teaching the model to produce the score in the requested format. The scoring instruction, editing prompt, and available images form the conditioning input; the reference score supplies the response that the model learns to generate.

Separate evaluator weights. The two dimensions and two reference settings are trained independently from the same pretrained backbone, giving four sets of adapted parameters. Each evaluator has its own scoring task and input format. SA and SU use the same target for a given edit and dimension, while differing in whether the source image is included in the input. At inference time, the corresponding evaluator generates a response from which the numerical score is extracted.

F SCOPE AND LIMITATIONS OF THE REFERENCE TARGETS

Identity-preservation labels depend on the selected recognition models and face preprocessing; edit-effectiveness labels depend on instruction decomposition, tool observations, and the annotation judge. Correlation with these labels measures agreement with automated criteria. The human study provides a separate comparison with human preferences on the annotated triplets, but broader perceptual validation and thresholds for practical use require further evaluation.