Contents

Face2Scene: Using Facial Degradation as an Oracle for Diffusion-Based Scene Restoration

Supplementary Material

1. Datasets

In this section, we provide details regarding the datasets considered in our paper: (i) the synthetic dataset generated by InfiniteYou, (ii) real gallery images (synthetically degraded for evaluation), and (iii) the real-world mobile-captured data, with real-world degradations.

1.1. InScene Synthetic Validation Dataset

We first construct a synthetic dataset of 12,595 samples by sampling faces from 1,005 identities in the CelebRef-HQ [13] dataset and using InfiniteYou [10] to generate full-body images in diverse scenes. We use the `Stage 1` weights from InfiniteYou for generation, as they provide stronger identity preservation compared to `Stage 2`. Generation is guided by an anti-blur LoRA[1] to reduce excessive background blur, resulting in varied and visually rich scenes that better highlight the effectiveness of our restoration pipeline. The diversity of these scenes is illustrated via a word-cloud visualization in Figure 1, where semantic tags are extracted from each image using the RAM++ [32] image tagging model.

As illustrated in Figure 2, we also provide example prompts: the *first* and *second* sentences describe the subject, while the *third* and *fourth* sentences describe the background in detail to increase the chance of generating highly detailed scene images.

To ensure identity consistency, we compute the ArcFace [3] embedding similarity between each generated image and its corresponding reference images, and discard samples with identity scores below 0.55. After filtering, we retain 11,266 images from 905 identities for training and 100 identities for the evaluation. For each generated sample, as shown in Figure 2, we store metadata including the *prompt*, *image resolution*, *detected face bounding box*, *facial landmarks*, the *File ID* (identity index in CelebRef-HQ [13]), and the ArcFace *identity similarity score* between the generated face and the reference identity image used during generation.

1.2. InScene Real Validation Dataset

As illustrated in Figure 3, we preprocess the real-world gallery and evaluation images using a three-stage pipeline that standardizes resolution, content, and perceptual quality before training and evaluation.

[1] https://huggingface.co/Shakker-Labs/FLUX.1-dev-LoRA-AntiBlur

Figure 1. **Scene diversity visualization.** Word-cloud representation of the semantic tags extracted from our synthetic dataset using the RAM++ [32] image-tagging model. Larger words indicate tags that occur more frequently across generated images, highlighting the broad diversity of scenes captured in our dataset.

Stage 1: Image size. We first apply a resolution filter and discard any image whose shorter side is smaller than 1000 pixels, ensuring sufficient spatial detail for both face and scene restoration. We then run a YOLOv8-based [17] face detector[2] and reject images in which no face is detected. For the remaining images, we normalize the spatial support by resizing such that $\min(H, W) = 1024$ and cropping to a 1024×1024 square while preserving the full detected face bounding box.

Stage 2: Image content. Next, we refine the face bounding box and measure its area relative to the full image. Images are discarded if the face occupies less than 1% of the image area, which removes cases where the subject is too small or dominated by background. Following prior work [5], we also apply a blur measure based on the variance of the Laplacian, to identify overly blurred samples and reject images that fall below a blur-awareness threshold, keeping only images with sufficiently sharp facial and scene details.

Stage 3: Image quality. In the final stage, we perform no-reference image quality assessment using three metrics: MUSIQ [11], CLIP-IQA [19], and MANIQA [24]. We retain only images that satisfy MUSIQ ≥ 65.0, CLIP-IQA ≥ 0.65, and MANIQA ≥ 0.40, filtering out low-quality or heavily degraded samples. This pipeline yields a curated real dataset with consistent resolution, reliable face content, and high perceptual quality, suitable for robust training and evaluation of our restoration framework.

[2] https://github.com/YapaLab/yolo-face

Prompt: **First**, full-body photography of woman in a bare deciduous forest in winter. **Second**, vertical orientation, direct perspective, wide-angle view 1.2, balanced exposure, distant background fully detailed, small aperture f/16–f/22, ultra high detail, distant canopy detail preserved. **Third**, a stone lanterns and mossy steps in the background, with distant dust plume, distant objects clearly rendered, edge-to-edge clarity. **Fourth**, sea grass blades and barnacle clusters stay legible, consistent detail across near and far surfaces, cinematic realism, ultra high detail.

Image Size: 1024 × 1024
Face BBox: [X_min, Y_min, X_max, Y_max],
Landmark
Identity Number: 00379
File ID: 2.png
Identity Score: 0.735057

Prompt: **First**, full-body photography of woman in a montane cloud forest. **Second**, vertical orientation, direct perspective, wide-angle view 1.2, distant background fully detailed, everything in frame clearly visible, ultra high detail, balanced exposure, textured bark and moss rendered clearly. **Third**, a rusted pipes, grates, and riveted beams in the background, with silver overcast light, distant objects clearly rendered, edge-to-edge clarity. **Fourth**, plaster texture and wood grain remain defined, consistent detail across near and far surfaces, cinematic realism, ultra high detail.

Image Size: 1024 × 1024
Face BBox: [X_min, Y_min, X_max, Y_max],
Landmark
Identity Number: 00945
File ID: 0.png
Identity Score: 0.726001

Prompt: **First**, full-body photography of man in a glass-walled network operations floor. **Second**, vertical orientation, direct perspective, wide-angle view 1.2, ultra high detail, distant background fully detailed, edge-to-edge clarity, everything in frame clearly visible, textured bark and moss rendered clearly. **Third**, a gnarled roots, ferns, and slick stones in the background, with warm window glows at dusk, distant objects clearly rendered, edge-to-edge clarity. **Fourth**, rope fibers and rusted fittings appear clearly resolved, consistent detail across near and far surfaces, cinematic realism, ultra high detail.

Image Size: 1024 × 1024
Face BBox: [X_min, Y_min, X_max, Y_max],
Landmark
Identity Number: 00309
File ID: 14.png
Identity Score: 0.693905

Prompt: **First**, full-body photography of woman in a frozen lake shore. **Second**, vertical orientation, direct perspective, wide-angle view 1.2, ultra high detail, wide-angle lens, everything in frame clearly visible, distant background fully detailed, distant canopy detail preserved. **Third**, an epiphytes clinging to damp trunks in the background, with cloudy soft light, distant objects clearly rendered, edge-to-edge clarity. **Fourth**, glass seams and architectural lines are razor-defined, consistent detail across near and far surfaces, cinematic realism, ultra high detail.

Image Size: 1024 × 1024
Face BBox: [X_min, Y_min, X_max, Y_max],
Landmark
Identity Number: 00063
File ID: 8.png
Identity Score: 0.714440

Figure 2. **Four representative images of our synthetic InScene dataset.** Each sample is shown alongside its structured prompt. We also show the stored metadata, including facial landmarks, bounding boxes, and associated identity labels.

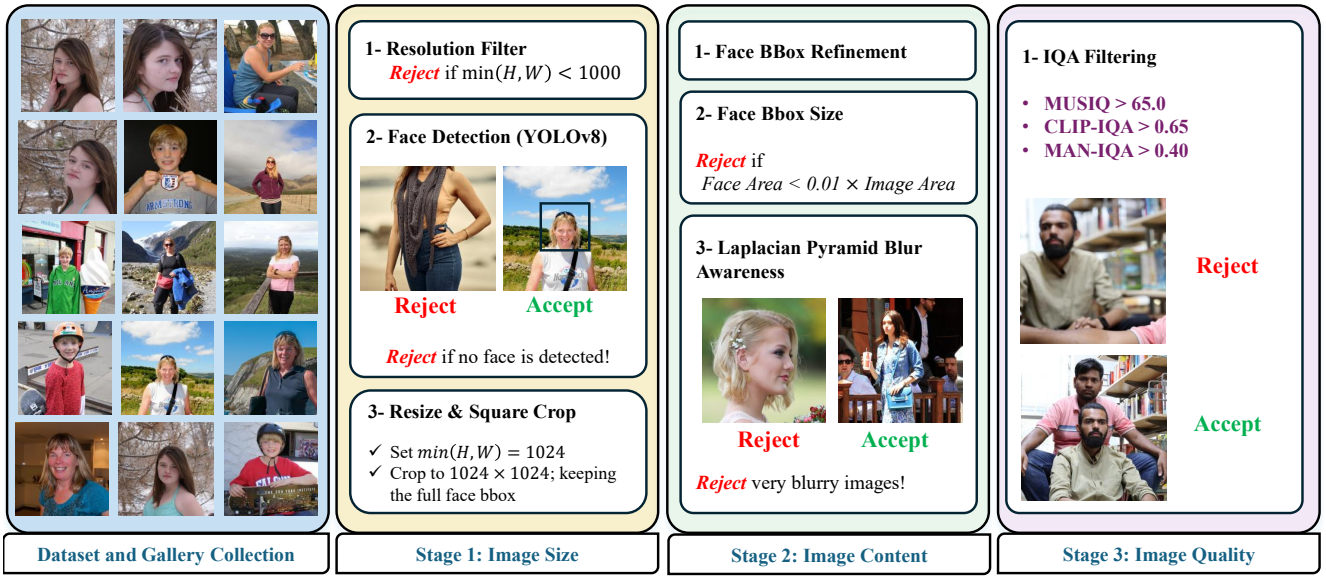

Figure 3. **Real dataset preprocessing pipeline.** Our real-world images are standardized through a three-stage filtering process. **Stage 1** removes images with insufficient resolution and ensures a detectable face is present. **Stage 2** enforces content quality by verifying that the face region is sufficiently large and the image is not overly blurred (e.g., due to defocus blur in the background, as in the example image shown above). **Stage 3** applies no-reference image quality metrics (MUSIQ, CLIP-IQA, and MANIQA) to retain only high-quality samples.

1.3. InScene Real Test Dataset.

We captured 100 real-world test images from 10 different identities using a Samsung S25 Edge smartphone. All images were recorded in RAW format with automatic exposure and ISO values set to 3200, 1600, and 800 to emulate diverse real-world degradations. To further intro-

duce motion-related degradations, we additionally captured a subset of images with a *motion blur*. We then processed the RAW files using the `rawpy` library[3] to obtain RGB images. These processed RGB images serve as the real degraded inputs for evaluating our model under realistic capture conditions.

2. Implementation Details

2.1. FaDeX Details

Figure 4 illustrates the contrastive training setup used for FaDeX, which follows a momentum-encoder design similar to MoCo [6]. Given HQ–LQ face pairs produced by the Ref-FR module, we feed two independent inputs of the concatenated faces, where they share a similar degradation, into a *principal* encoder branch and a *momentum* encoder branch. Both branches share the same architecture: a stack of 3×3 convolution, BatchNorm [9], and LeakyReLU blocks followed by a projection head (Linear + LeakyReLU + Linear). The principal branch is updated by standard backpropagation, while the momentum branch is an exponential moving average of the principal encoder parameters and is never directly optimized. The output of each branch is passed through the projector to obtain d-dimensional face-degradation embeddings. Embeddings from the momentum branch are enqueued into a first-in–first-out *negative queue* that maintains a large, constantly refreshed set of past samples. At each iteration, for a given sample i in the current mini-batch, the embedding from the principal branch acts as the query, q_i, embeddings from other samples with the same degradation operator, $\mathcal{G}$, form the positive set, and embeddings drawn from the queue (and non-matching samples in the batch) serve as negatives. The contrastive degradation loss, $\mathcal{L}_{\text{Deg}}$, is then computed over this large set of negatives. After the gradient update on the principal branch, we update the momentum branch via EMA and refresh the negative queue by enqueuing the current momentum embeddings and dequeuing the oldest ones. This implementation stabilizes training, enlarges the pool of negative degradations without increasing batch size, and yields robust, degradation-discriminative FaDeX features that we subsequently freeze and use as conditioning for the one-step restorer.

2.2. Additional One-Step Diffusion Model Implementation Details

During training, we adopt an online negative prompt learning scheme inspired by previous work [30]. For each mini-batch, with probability, p_n, we replace the ground-truth HR target with its synthesized LR counterpart to form a negative target, while the remaining samples keep their original HR targets as positives. We pair negative targets with a fixed negative prompt (e.g., "oil painting, cartoon, blur, dirty, messy, low quality, deformation, low resolution, oversmooth") and positive targets with a generic high-quality prompt ("a high-resolution, 8K, ultra-realistic sharp image, vibrant colors, and natural lighting"). During inference, we apply classifier-free guidance [7] with both prompts: the VAE encoder, $\mathbf{E}_\theta$, and UNet denoiser, ϵ_θ, produce features for the positive and negative prompts, $z_{\text{pos}} = \epsilon_\theta(\mathbf{E}_\theta(I_{\text{LR}}), t_{\text{pos}})$ and $z_{\text{neg}} = \epsilon_\theta(\mathbf{E}_\theta(I_{\text{LR}}), t_{\text{neg}})$, and the final representation is obtained as

$$z_{\text{out}} = z_{\text{neg}} + \lambda_{\text{cfg}}(z_{\text{pos}} - z_{\text{neg}}),$$

where λ_{cfg} is the guidance scale. This teaches the model to associate negative prompts with low-quality artifacts and positive prompts with desired restoration quality, while reusing synthesized LR images and thus adding no extra overhead to the training pipeline. We use $\lambda_{\text{cfg}} = 1.10$ in all our experiments.

Degradation Attention Implementation We implement the degradation attention module, $\mathrm{DegAttn}$ with eight heads, 512 input channels, and 1024-dimensional output tokens that match the cross-attention embedding size of the diffusion model. Starting from the FaDeX feature map $Z_{\text{face}} \in \mathbb{R}^{B\times256\times H\times W}$, we first apply an overlapping patch embedding, which halves the spatial resolution and expands the channels to 512. We then split this tensor along the channel dimension into two branches, F_1 and F_2, and flatten them into token sequences that serve as the two inputs to $\mathrm{DegAttn}$. Inside $\mathrm{DegAttn}$, we build separate queries and keys from F_1 and F_2, each using half of the eight heads (four heads per branch), yielding two self-attention maps A_1 and A_2. Values are computed from the concatenation $[F_1; F_2]$, so that each head operates on a richer $2d_h$ value vector that jointly encodes both streams. A learnable scalar, λ, controls how much the second branch's attention should be subtracted from the first. The result Y_{ch} is normalized with RMSNorm and projected back to 256 channels. From Y_{ch}, we derive multi-scale degradation tokens using three pooling stages described in the method section.

2.3. Reference-Based Face Restoration Model Selection

Among the public and open-source reference-based restoration models that outperform simple restoration models, we found InstantRestore [31], FaceMe [16], Ref-LDM [8], and RestorerID [26]. Unfortunately, most methods are not publicly available. Among these, FaceMe gave the best results while being efficient at inference. Therefore, we adopt FaceMe as our Ref-FR model to generate a high-quality face given reference images.

[3] https://github.com/letmaik/rawpy

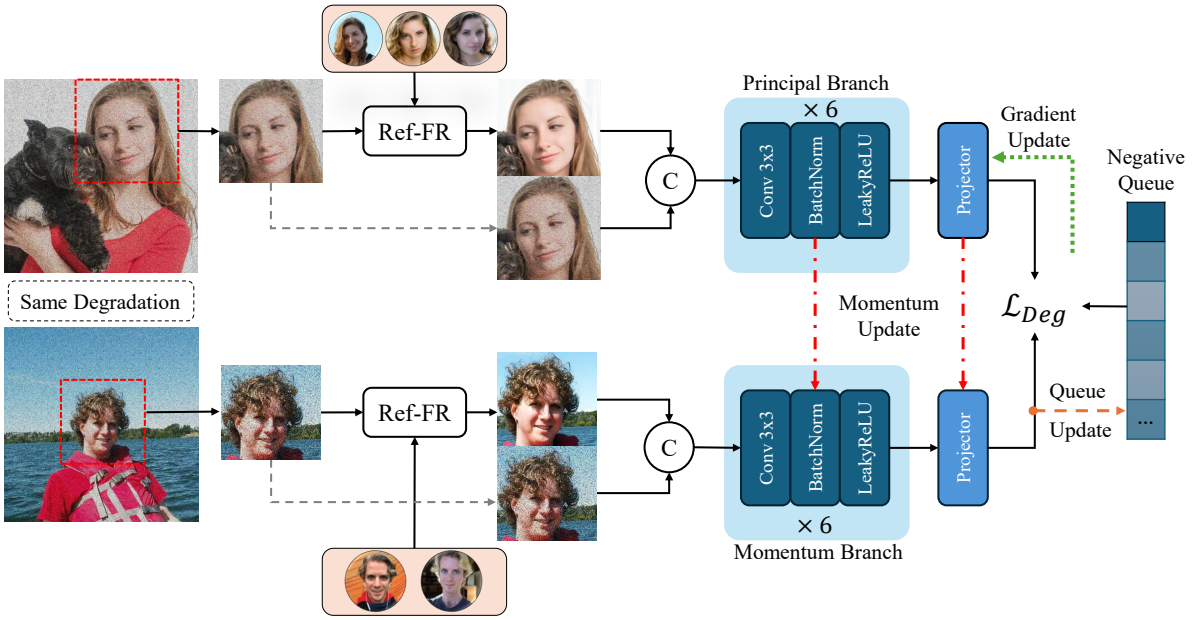

Figure 4. **Overview of FaDeX contrastive learning.** Momentum-encoder contrastive training setup with dual encoders and a queue of negatives to learn degradation-discriminative face embeddings.

2.4. FaDeX Cosine-Similarity Experiment

In the paper, we evaluate whether FaDeX encodes degradations while being invariant to image content. To this end, we apply four fixed RealESRGAN-style degradation presets to 10 randomly selected images and compute cosine similarities in the FaDeX embedding space. We denote these degradation settings by $d_1, \ldots, d_4$ and keep all hyperparameters deterministic, so each d_i corresponds to a reproducible degradation operator. Each preset uses the standard two-stage RealESRGAN pipeline.

- d_1**: strong downscale + strong blur + high JPEG.** Stage one applies a strong downscale, isotropic Gaussian blur, gray Gaussian noise, Poisson noise, and high-quality JPEG compression. Stage two keeps the resolution, applies a milder isotropic blur and very high JPEG quality, followed by a final sinc filter. This yields heavily smoothed, downsampled images with strong blur and mild JPEG artifacts.
- d_2**: strong upscale + low JPEG, then downscale + anisotropic blur.** Stage one strongly upscales the image, adds colored Poisson noise, anisotropic Gaussian blur, and low JPEG quality. Stage two downsamples the image, adds additional Poisson noise, applies anisotropic blur, and uses low–mid JPEG quality, without a final sinc filter. Overall, d_2 produces moderately blurred images with strong Poisson noise and noticeable JPEG artifacts.
- d_3**: almost-clean identity with sinc filtering.** Both stages keep the original resolution and use maximum JPEG quality. Stage one has no Gaussian noise and minimal Poisson noise with a sinc filter; Stage two adds only a very small Gaussian noise and another sinc filter. This preset produces nearly artifact-free images with very mild sharpening/sinc effects and negligible noise.
- d_4**: moderate downscale + light noise + very low JPEG, then upscale + plateau blur.** Stage one moderately downsamples the image, adds light colored Gaussian noise, mild Poisson noise, plateau-anisotropic blur, and very low JPEG quality. Stage two then upscales the image, applies plateau-anisotropic blur again, very low JPEG quality, and a final sinc filter. This yields images with structured blur, strong JPEG artifacts, and a slight scale inconsistency.

3. Experimental Results

3.1. Quantitative Results: InScene Test Dataset

In Table 1, we compare models on the InScene real test dataset captured by the Samsung S25 Edge phone. As there is no ground-truth image in this case, we report on commonly used non-reference image quality assessment metrics (MUSIQ, Clip-IQA, MANIQA, LIQE, and TOPIQ). The results show that our approach compares favorably to other baselines, being either best or second-best according to four of the five models. In particular, compared to the second best model in the table, DiffBIR, our model outputs sharper images with more details, as observed in Figures 15, 16, and 17.

Table 1. **Quantitative comparison on the InScene Test Dataset.** Since real-world photographs do not have a ground-truth image against which to compare, we quantitatively evaluate on our InScene Test data via no-reference IQA models. Arrows indicate whether lower (↓) or higher (↑) values are better. C-IQA and M-IQA denote CLIP-IQA and MANIQA, respectively. Each cell is color-coded to represent the best and second-best performance.

InScene Test Dataset					
Methods	**MUSIQ↑**	**C-IQA↑**	**M-IQA↑**	**LIQE↑**	**TOPIQ↑**
SUPIR [27]	60.4119	0.3506	0.3108	2.8347	0.4210
DiffBIR [15]	71.4031	0.6701	0.5238	4.3273	0.6751
ResShift [28]	65.8328	0.5414	0.3827	3.7667	0.4916
PASD [25]	71.6666	0.5920	0.4849	4.3644	0.6483
OSEDiff [23]	72.6963	0.6327	0.4819	4.4864	0.6583
SinSR [21]	71.6175	0.5965	0.4947	4.3157	0.6746
InvSR [29]	72.7069	0.4559	0.4145	4.3888	0.6665
S3Diff [30]	69.4941	0.5423	0.4014	4.1888	0.5805
Face2Scene	73.3047	0.6407	0.4859	4.6973	0.6846

3.2. Qualitative Results

3.2.1. InScene Synthetic and Real Validation Datasets

As shown in the qualitative comparisons on the InScene synthetic (Figures 7, 8, 9, 10, 11, and 12), and real validation datasets (Figures 13, and 14) conditioning only on the face-derived degradation prior enables our method to restore substantially more coherent backgrounds with richer textures, directly aligning with our core motivation. In the facial regions, our approach also produces more human-like appearances with fewer distortions than existing restoration or diffusion-based generative models. These qualitative observations hold consistently for both the synthetic and real validation sets and align with the quantitative trends, where our method achieves higher perceptual scores and stronger no-reference image quality metrics. Together, these results confirm that estimating degradation solely from faces provides an effective prior that improves both global scene quality and facial fidelity across synthetic and real data.

3.2.2. InScene Test Dataset

Figures 15, 16, and 17 show real test examples captured by the Samsung S25 Edge mobile device and restored by different methods. As can be observed in the examples, our method reconstructs a higher quality face with less artifacts. Further, on the background regions, our model yields sharper outputs with more details and less noise. In particular, the reconstructed face generated by our method tends to maintain better fidelity to the LQ input, meaning it can recover from blurriness and noise, yet also avoid hallucinated artifacts. For example, in Figure 16 (rightmost column), our model is able to deblur the face with plausible details, whereas either output a blurry face (most other methods) or an artifacted one (OSEDiff and S3Diff). Aside from the face, we also see high frequency detail recovery is improved; for instance, in Figure 15 (left columns), the horizontal stripes are preserved sharply by our method, while others blur them, restore them in a slightly cartoonish manner, or remove them entirely.

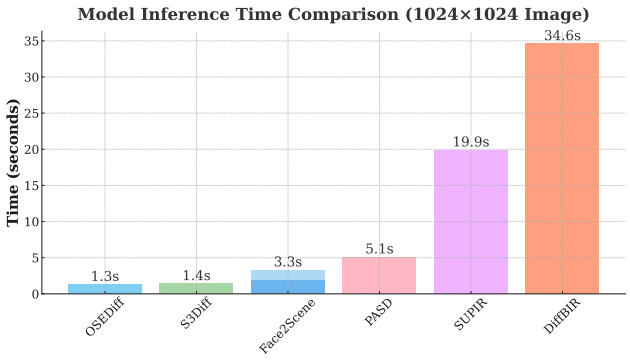

Figure 5. **Model complexity comparison.** We compare inference time across different diffusion-based restoration models. For Face2Scene, we show the separate time costs of the reference-based preprocessing and our final restoration as two different shades. While our method is not as efficient as standard one-step models, which do not require an initial face restoration step, it is still considerably faster than current multistep models, despite performing on-par or better than them, in terms of image quality and restoration fidelity.

3.3. Model Complexity

To contextualize the efficiency of our approach, we compare the inference time of Face2Scene against several state-of-the-art restoration baselines (OSEDiff [23], S3Diff [30], PASD [25], SUPIR [27], and DiffBIR [15]), as shown in Figure 5. All timings are measured on an NVIDIA A6000 48GB GPU using 1024×1024 input images, averaged over 100 samples. For each model, we follow the official inference pipeline. Notably, in our method, the degradation extractor operates on the cropped face region, whose size varies across images. Since the module is fully resolution-agnostic, its runtime naturally scales with the face-crop resolution: smaller faces lead to faster processing than the reported average, whereas larger faces incur a slightly higher cost.

4. Ablation Studies

4.1. Impact of CFG scaling factor

We study the effect of the classifier-free guidance (CFG) scaling factor on our restoration pipeline. To isolate this behavior, we sweep the scaling value from 1.00 (using only the positive prompt) up to 1.16 and evaluate performance on the InScene Synthetic Validation set. Figure 6 presents the results, where the first row shows no-reference metrics (CLIP-IQA, MUSIQ, MANIQA, TOPIQ, LIQE) and the second row shows reference-based metrics (PSNR, LPIPS, DISTS,

SSIM, FID).

Overall, we observe a consistent trend: increasing the CFG scaling factor improves no-reference metrics, indicating that the generated images become visually sharper and more perceptually appealing. However, this comes at the cost of reduced reference-based performance, where fidelity to the ground-truth image gradually decreases as the scaling factor increases. This trade-off reflects the classical balance between perceptual sharpness and reconstruction accuracy [2]; higher guidance pushes the model toward more confident, high-frequency outputs, while lower guidance yields more faithful but slightly softer reconstructions. Based on this analysis, we set the default value to $\lambda_{\text{cfg}} = 1.10$ for all experiments, which provides a strong perceptual gain while retaining competitive fidelity.

4.2. Impact of LoRA Rank

Table 2 shows how the LoRA rank for the VAE encoder and UNet affects restoration quality. Very low ranks (4/4, 8/8) substantially degrade both fidelity (PSNR/SSIM) and perceptual/no-reference metrics. Increasing the rank to 16/16 already yields strong gains, and the asymmetric setting (VAE: 16, UNet: 32) further improves DISTS, LPIPS, FID, and all no-reference scores, while keeping PSNR and SSIM competitive. We therefore adopt the 16/32 configuration as our default trade-off between performance and parameter efficiency.

4.3. Impact of Multi-scale Degradation Tokens

To understand how the granularity of degradation encoding influences the restoration process, we ablate the number of degradation tokens used for conditioning the diffusion model. As shown in Table 3, we compare three configurations: using only a single global token, adding a set of intermediate-scale tokens, and finally incorporating the full set of global, intermediate, and local tokens.

We observe a clear performance trend: increasing the number of degradation tokens consistently improves most perceptual and no-reference metrics. While global-only conditioning captures coarse degradation properties, adding intermediate tokens leads to modest gains and more stable fidelity metrics (e.g., PSNR, SSIM). The full multi-scale configuration, which uses 21 tokens, yields the strongest performance across nearly all perceptual measures, indicating that richer degradation representations provide the diffusion model with more precise conditioning.

In practice, the number of tokens is constrained by the spatial embedding size of the degradation features. Our design uses a 4×4 grid at the latest scale, which implies that the smallest usable face crop is 16×16 pixels; already a very small face. Taking this constraint into account, we adopt the 21-token configuration as a good balance between expressive multi-scale conditioning and robustness to small face crops.

4.4. Identity Preservation

In Table 4, we show the face identity preservation accuracy of the different methods. In particular, for each model output, we cropped the face region of the super-resolved image and computed the ArcFace [3] cosine similarity between the GT and restored face. The results show that our approach most reliably preserves identity information.

4.5. Performance Comparison: Our Model vs. S3Diff on Our Dataset

To provide a fairer comparison to our approach, in Table 5, we compare against the closest model to ours (S3Diff [30]) trained on the same InScene dataset and using the exact same experimental setup. We call this variant S3Diff**. While the results of the two models are close on the synthetic validation dataset, there is a noticeable gap in the real validation dataset between our approach and S3Diff variants, especially for reference-based metrics. This is likely due to having more diverse test examples in the real validation set, compared to the synthetic validation set. Indeed, compared to the synthetic dataset (derived from celebrity photos), the real data not only has a more complex distribution of people, but the content of the images themselves (e.g., the pose of the inserted people, or the level of detail in the background) is more complicated. Overall, the results suggest that the improvement gained by our model is more generalizable to out-of-domain data. We hypothesize this is due to more accurate degradation estimation and injection, which avoids overfitting to the training data, in part by explicitly disentangling content from degradation.

4.6. Robustness to Reference Quality

We analyze the sensitivity of Face2Scene to the quality of the reference-based face restoration module (Ref-FR) on InScene Synthetic Validation Dataset. We generate Ref-FR outputs with three different quality levels: *good*, *medium* (default), and *bad*, using a controlled synthetic perturbation, and compare the resulting scene restoration in Table 6. The corresponding restored-face quality is shown in the last row of the table, where the effect of the synthetic perturbation is reflected in the quantitative results. As shown, Face2Scene degrades *gracefully* as Ref-FR quality worsens and remains competitive with S3Diff, indicating that Stage II does not require near-perfect Ref-FR to be effective.

4.7. Comparison with Face Restoration Models

Face-only restoration methods are not directly comparable to our primary setting, as they assume cropped and aligned face inputs and do not reconstruct the full scene. Nevertheless, we evaluate Face2Scene on the dedicated face restoration benchmark CelebA-Ref, which contains 100 identi-

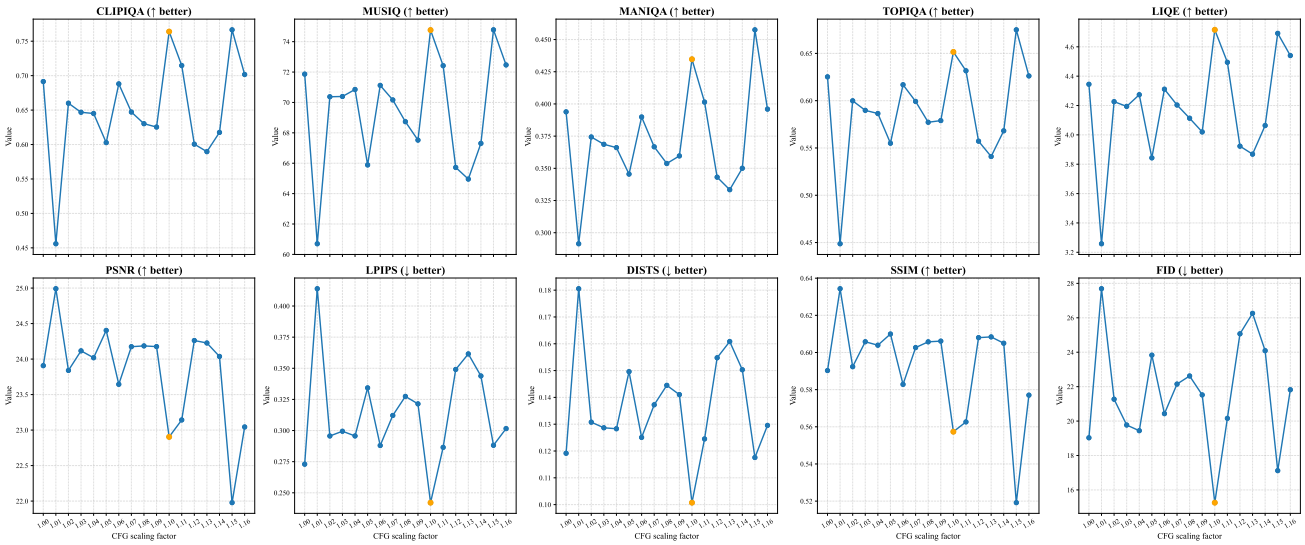

Figure 6. **Effect of CFG scaling.** We vary the CFG scaling factor from 1.00 to 1.16 on the InScene synthetic validation set. The top row shows no-reference metrics and the bottom row shows reference-based metrics, illustrating the trade-off between perceptual quality and fidelity. In particular, although there is some noise in the trends, one can see that the no-reference quality metrics tend to improve as the CFG weight increases, while the low-level fidelity (PSNR and SSIM) tend to decline. The orange dot signifies the CFG weight that we chose, which obtains high perceptual quality while maintaining fidelity (particularly according to the full-reference perceptual metrics).

Table 2. **Impact of LoRA rank on the VAE encoder and UNet.** We evaluate different LoRA rank configurations for the VAE and UNet to analyze their effect on restoration quality across perceptual, fidelity, and no-reference metrics on the InScene synthetic validation set. We observe that low ranks heavily damage performance, on both perceptual distances (LPIPS and DISTS) and no-reference quality. Our 16/32 configuration performs quite well on both types, though it is at the expense of pixel-level metrics (PSNR and SSIM). Increasing the rank beyond this causes a slight decline according to most metrics.

InScene Synthetic Validation Dataset											
VAE LoRA Rank	**UNet LoRA Rank**	**DISTS↓**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**FID↓**	**MUSIQ↑**	**C-IQA↑**	**M-IQA↑**	**LIQE↑**	**TOPIQ↑**
4	4	0.1771	0.4034	24.1619	0.6153	29.15	62.0885	0.5289	0.3093	3.6797	0.4913
8	8	0.1523	0.3505	23.9881	0.6035	23.28	65.8073	0.6009	0.3433	3.9864	0.5469
16	16	0.1244	0.2881	23.4638	0.5834	19.58	71.1710	0.6928	0.3906	4.3963	0.5984
16	32	0.1007	0.2421	22.9040	0.5574	15.26	74.7630	0.7640	0.4347	4.7157	0.6515
32	32	0.1076	0.2536	23.3672	0.5736	16.61	73.7835	0.7545	0.4434	4.6329	0.6607

ties (1064 images) that we used for the InScene synthetic validation split and are unseen during training. Despite not being designed as a specialized face restoration model, Face2Scene achieves the best overall performance on most metrics compared with strong non-reference face restoration baselines (Table 7). This improvement is expected, as our method leverages identity reference images that provide consistent appearance and identity cues, whereas competing methods must hallucinate missing facial details from a single degraded face crop.

4.8. Spatially Varying Degradation

To evaluate robustness to spatially non-uniform degradations, we modify the *first blur stage* of the Real-ESRGAN degradation pipeline to apply a spatially varying blur kernel to generate new LQ samples for InScene Synthetic Validation Dataset. Following Lin *et al.* [14], we model blur as a position-dependent Point Spread Function (PSF) field, yielding milder blur near the image center and progressively stronger blur toward the image corners. Quantitative results are reported in Table 8. Although Face2Scene is trained under the assumption of globally consistent degradations, it remains robust in this setting and outperforms S3Diff on most perceptual and no-reference quality metrics. Representative qualitative examples are also shown in Figure 15. We find that Face2Scene restores facial regions reliably even under this more challenging degradation, whereas most competing methods struggle to recover clear facial details. Some background regions still exhibit mild residual blur, highlighting a limitation of the current formulation when the degradation is not spatially uniform.

Table 3. **Ablation on hierarchical components.** We incrementally enable global, intermediate, and local components to assess their impact on restoration quality across perceptual, fidelity, and no-reference metrics. Our complete configuration, which combines tokens across scales, performs the best across all perceptual evaluation measures (i.e., excluding the low-level distortion measures, PSNR and SSIM).

InScene Synthetic Validation Dataset										
Configuration	**DISTS↓**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**FID↓**	**MUSIQ↑**	**C-IQA↑**	**M-IQA↑**	**LIQE↑**	**TOPIQ↑**
Global only (1 Token)	0.1180	0.2671	23.6223	0.5862	19.50	72.0117	0.6497	0.3769	4.2772	0.5942
Global + Intermediate (5 Tokens)	0.1294	0.2997	23.7422	0.5921	21.92	69.9288	0.6582	0.3665	4.2190	0.5820
Global + Intermediate + Local (21 Tokens)	0.1007	0.2421	22.9040	0.5574	15.26	74.7630	0.7640	0.4347	4.7157	0.6515

Table 4. **Identity similarity on the InScene validation datasets.** Our ID score measures similarity in terms of identity, meaning higher values indicate better identity preservation. Since our reference-aware degradation estimator explicitly discards non-degradation information, including identity, we investigated identity preservation as a metric specifically. We find that our method still performs best, in terms of retaining or inferring identity. This suggests our degradation estimation is accurate enough to help the model recover the true underlying face details, thereby preserving the correct identity as well.

Dataset	**Metric ↑**	SUPIR [27]	DiffBIR [15]	ResShift [28]	PASD [25]	OSEDiff [23]	SinSR [21]	InvSR [29]	S3Diff [30]	**Face2Scene**
Synthetic	ID Score	0.4434	0.4433	0.4477	0.4030	0.4477	0.4239	0.4055	0.4684	**0.4867**
Real		0.4333	0.4286	0.4692	0.4333	0.3766	0.4334	0.3892	0.4598	**0.4881**

Table 5. **Quantitative comparison on the InScene synthetic and real validation sets.** To investigate the importance of training with our dataset and provide a fairer comparison, we retrain the S3Diff model on our InScene dataset under the same experimental settings, resulting in the model denoted S3Diff**. Although doing so does result in improved performance over the standard S3Diff, our approach still outperforms S3Diff** by a significant margin, particularly according to the full-reference metrics on the Real set. Since our method is built upon S3Diff, this suggests our novel reference-derived degradation estimation module is necessary to obtain strong performance. Arrows indicate whether lower (↓) or higher (↑) values are better. C-IQA and M-IQA denote CLIP-IQA and MANIQA, respectively. Each cell is color-coded to represent the best and second-best performance.

InScene Synthetic Validation Dataset										
Methods	**DISTS↓**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**FID↓**	**MUSIQ↑**	**C-IQA↑**	**M-IQA↑**	**LIQE↑**	**TOPIQ↑**
S3Diff [30]	0.1131	0.2557	23.5955	0.5916	18.06	72.1764	0.6980	0.3858	4.4248	0.6233
S3Diff**	0.1039	0.2486	22.4990	0.5382	15.81	74.5945	0.7605	0.4384	4.7127	0.6450
Face2Scene (ours)	0.1007	0.2421	22.9040	0.5574	15.26	74.7630	0.7640	0.4347	4.7157	0.6515
InScene Real Validation Dataset										
Methods	**DISTS↓**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**FID↓**	**MUSIQ↑**	**C-IQA↑**	**M-IQA↑**	**LIQE↑**	**TOPIQ↑**
S3Diff [30]	0.2231	0.5149	17.1439	0.4894	38.64	73.8209	0.6734	0.4480	4.7060	0.6627
S3Diff**	0.2176	0.5069	16.9445	0.4702	37.63	74.6644	0.6872	0.4745	4.7249	0.6613
Face2Scene (ours)	0.1178	0.2502	22.8975	0.6197	42.21	75.3739	0.7015	0.4714	4.8044	0.6777

Method	DISTS↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	MUSIQ↑	C-IQA↑	M-IQA↑	LIQE↑	TOPIQ↑
Face2Scene	0.099 / 0.101 / 0.106	0.239 / 0.242 / 0.251	23.01 / 22.90 / 22.86	0.562 / 0.557 / 0.562	15.03 / 15.26 / 17.83	74.40 / 74.76 / 74.81	0.762 / 0.764 / 0.770	0.431 / 0.435 / 0.442	4.70 / 4.72 / 4.76	0.643 / 0.652 / 0.652
S3Diff	0.113	0.256	23.60	0.592	18.06	72.18	0.698	0.386	4.42	0.623
FaceMe	0.219 / 0.227 / 0.238	0.375 / 0.403 / 0.452	25.65 / 24.73 / 22.98	0.729 / 0.711 / 0.515	58.74 / 62.73 / 68.37	74.21 / 74.57 / 74.29	0.676 / 0.684 / 0.688	0.558 / 0.564 / 0.562	4.90 / 4.93 / 4.89	0.679 / 0.681 / 0.680

Table 6. **Robustness to reference restoration quality on the InScene Synthetic validation set.** For Face2Scene, the Stage I Ref-FR model produces reference faces at three quality levels (*good / medium / bad*). The table reports the resulting scene restoration performance when these references are used in Stage II. As the reference quality decreases, Face2Scene shows only moderate degradation across perceptual and no-reference quality metrics, indicating that the scene restoration stage is robust to imperfect face restoration. In particular, even with degraded references, Face2Scene remains competitive with S3Diff, demonstrating that Stage II does not require near-perfect reference faces to operate effectively.

5. Failure Modes

We show some failure cases of our model in Figure 18. In particular, we observe that our model has difficulty restoring the text regions and small faces. Though such content is inherently difficult, these issues are also due in part to using stable diffusion (SD) 2.1 [18] as our diffusion backbone, which we observe has limitations in handling text and reconstructions with small contexts. Using FLUX [1] or

Method	DISTS↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	MUSIQ↑	LIQE↑
CodeFormer [33]	0.1665	0.2504	26.0058	0.7089	26.27	75.3111	4.8531
OSDFace [20]	0.1606	0.2485	24.7144	0.7009	22.31	75.4782	4.8672
RestoreFormer++ [22]	0.2012	0.3925	20.4579	0.5954	26.67	74.5379	4.7613
Face2Scene	0.1432	0.2180	26.4253	0.7367	19.21	74.6088	4.8789

Table 7. **Face restoration comparison on CelebA-Ref.** Face2Scene achieves the best performance on most metrics against dedicated face restoration baselines. This is consistent with the use of reference images, which provide additional identity and appearance cues.

Method	DISTS↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	MUSIQ↑	C-IQA↑	M-IQA↑	LIQE↑	TOPIQ↑
Face2Scene	**0.1137**	**0.2599**	22.9552	0.5571	**17.88**	**73.9026**	**0.7586**	**0.4269**	**4.6433**	**0.6405**
S3Diff	0.1275	0.2794	**23.8162**	**0.5923**	19.55	70.9485	0.6870	0.3800	4.2837	0.6143

Table 8. **Comparison under spatially varying blur.** We replace the first blur stage of the Real-ESRGAN degradation pipeline with a position-dependent blur field, such that blur is weaker near the center and stronger toward the corners. Although trained assuming globally consistent degradations, Face2Scene remains robust and outperforms S3Diff on most perceptual and no-reference metrics, indicating better perceptual restoration under spatially varying blur.

Method	DISTS↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	MUSIQ↑	C-IQA↑	M-IQA↑	LIQE↑	TOPIQ↑
Face2Scene	**0.1007**	**0.2421**	22.9040	0.5574	**15.26**	**74.7630**	**0.7640**	**0.4347**	**4.7157**	**0.6515**
RedegNet [12]	0.1675	0.3868	**24.7300**	**0.6290**	36.06	66.9653	0.4893	0.3056	3.1354	0.5181

Table 9. **Comparison between Face2Scene and RedegNet [12] on InScene Synthetic Validation Dataset.** Face2Scene achieves better perceptual quality and no-reference image quality metrics, while RedegNet attains higher PSNR and SSIM. ReDegNet learns realistic degradation representations from face images and transfers them to natural images to synthesize degraded training data for blind image super-resolution.

SD3.5 [4], which generate bigger images with more details and can also better handle text, would help overcome these limitations. Separately, another limitation is restoring different degrees of blur due to depth of field. As our method's degradations are based on the face region, it does not account for different levels of degradations (i.e., spatial variation). This problem can be alleviated, for example, by adding a depth-dependent blur estimation module to compensate for such cases.

References

[1] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 9

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 7

[3] 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*, pages 4690–4699, 2019. 2, 7

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 10

[5] Jue Gong, Jingkai Wang, Zheng Chen, Xing Liu, Hong Gu, Yulun Zhang, and Xiaokang Yang. Human body restoration with one-step diffusion model and a new benchmark. In *International Conference on Machine Learning*, 2025. 2

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on computer vision and pattern recognition*, 2020. 4

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[8] Chi-Wei Hsiao, Yu-Lun Liu, Cheng-Kun Yang, Sheng-Po Kuo, Kevin Jou, and Chia-Ping Chen. Ref-ldm: A latent diffusion model for reference-based face image restoration. *Advances in Neural Information Processing Systems*, 37: 74840–74867, 2024. 4

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 2015. 4

[10] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. InfiniteYou: Flexible photo recrafting while preserving your identity. *International conference on computer vision (ICCV)*, 2025. 2

[11] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 2

[12] Xiaoming Li, Chaofeng Chen, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. From face to natural image: Learning real degradation for blind image super-resolution. In *European Conference on Computer Vision*, pages 376–392. Springer, 2022. 10

[13] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5904–5917, 2022. 2

[14] Esther YH Lin, Zhecheng Wang, Rebecca Lin, Daniel Miau, Florian Kainz, Jiawen Chen, Xuaner Zhang, David B Lindell, and Kiriakos N Kutulakos. Learning lens blur fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 8

[15] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pages 430–448. Springer, 2024. 6, 9

[16] Siyu Liu, Zheng-Peng Duan, Jia OuYang, Jiayi Fu, Hyunhee Park, Zikun Liu, Chun-Le Guo, and Chongyi Li. Faceme:

Robust blind face restoration with personal identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5567–5575, 2025. 4

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[18] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 9

[19] 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*, pages 2555–2563, 2023. 2

[20] Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. Osdface: One-step diffusion model for face restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12626–12636, 2025. 10

[21] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25796–25805, 2024. 6, 9

[22] Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. Restoreformer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15462–15476, 2023. 10

[23] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37:92529–92553, 2024. 6, 9

[24] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022. 2

[25] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European conference on computer vision*, pages 74–91. Springer, 2024. 6, 9

[26] Jiacheng Ying, Mushui Liu, Zhe Wu, Runming Zhang, Zhu Yu, Siming Fu, Si-Yuan Cao, Chao Wu, Yunlong Yu, and Hui-Liang Shen. Restorerid: Towards tuning-free face restoration with id preservation. *arXiv preprint arXiv:2411.14125*, 2024. 4

[27] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25669–25680, 2024. 6, 9

[28] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6, 9

[29] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23153–23163, 2025. 6, 9

[30] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024. 4, 6, 7, 9

[31] Howard Zhang, Yuval Alaluf, Sizhuo Ma, Achuta Kadambi, Jian Wang, and Kfir Aberman. Instantrestore: Single-step personalized face restoration with shared-image attention. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 4

[32] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 2

[33] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 10

Figure 7. **Visual comparison on InScene Synthetic dataset across seven methods.**

Figure 8. **Visual comparison on InScene Synthetic dataset across seven methods.**

Figure 9. **Visual comparison on InScene Synthetic dataset across seven methods.**

Figure 10. **Visual comparison on InScene Synthetic dataset across seven methods.**

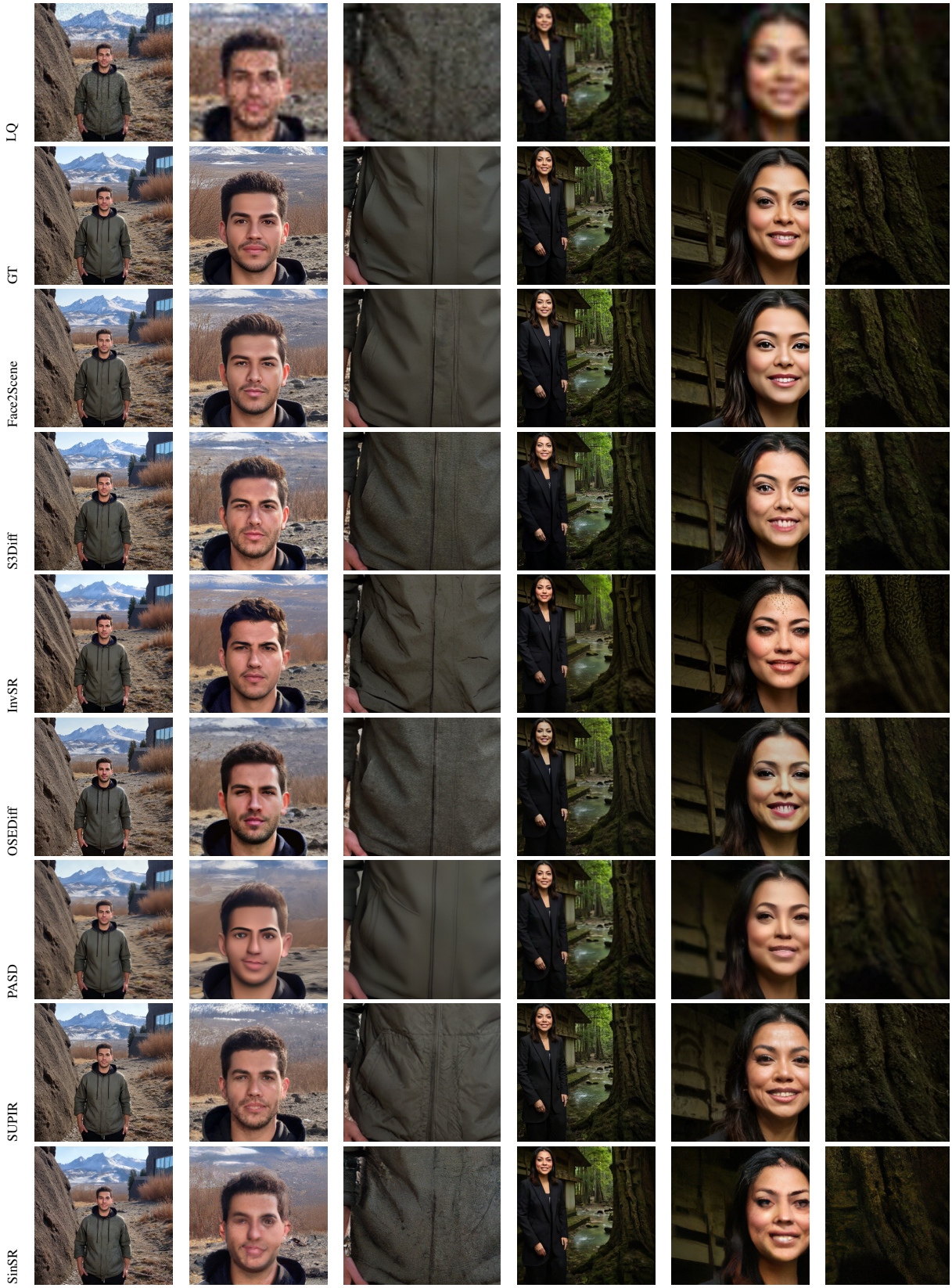

Figure 11. **Visual comparison on InScene Synthetic dataset across seven methods.**

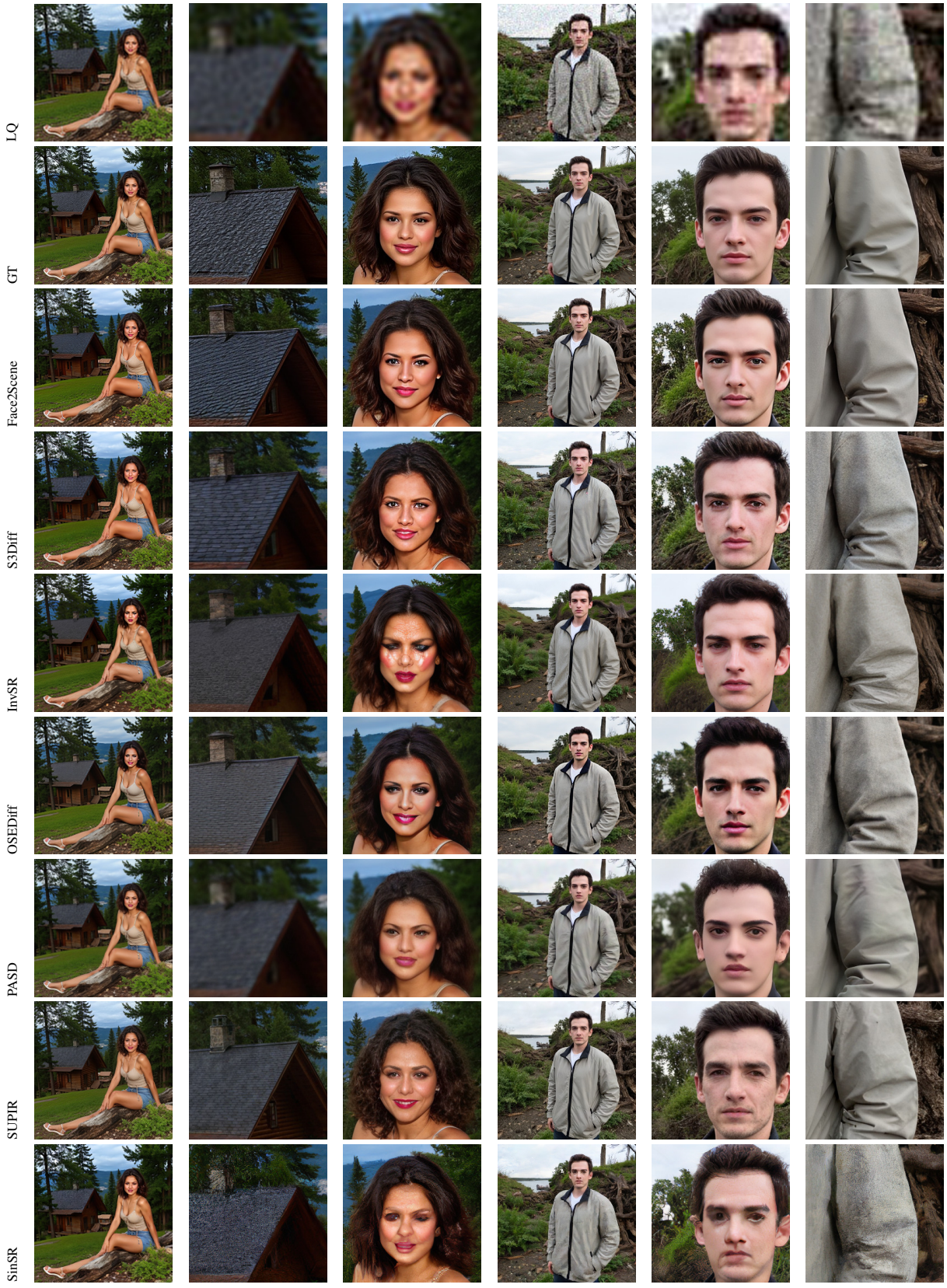

Figure 12. **Visual comparison on InScene Synthetic dataset across seven methods.**

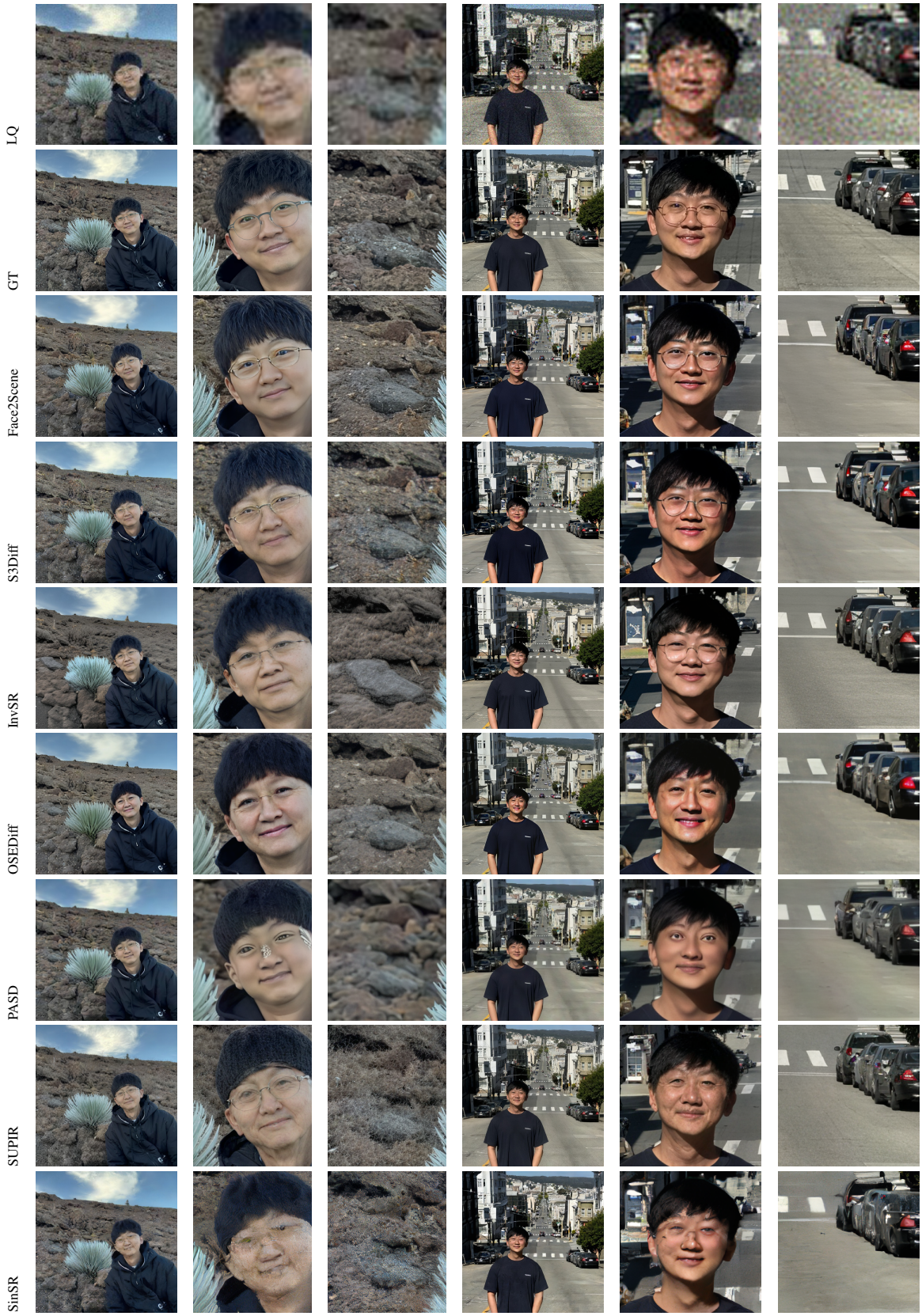

Figure 13. **Visual comparison on real validation across seven methods.**

Figure 14. **Visual comparison on real validation across seven methods.**

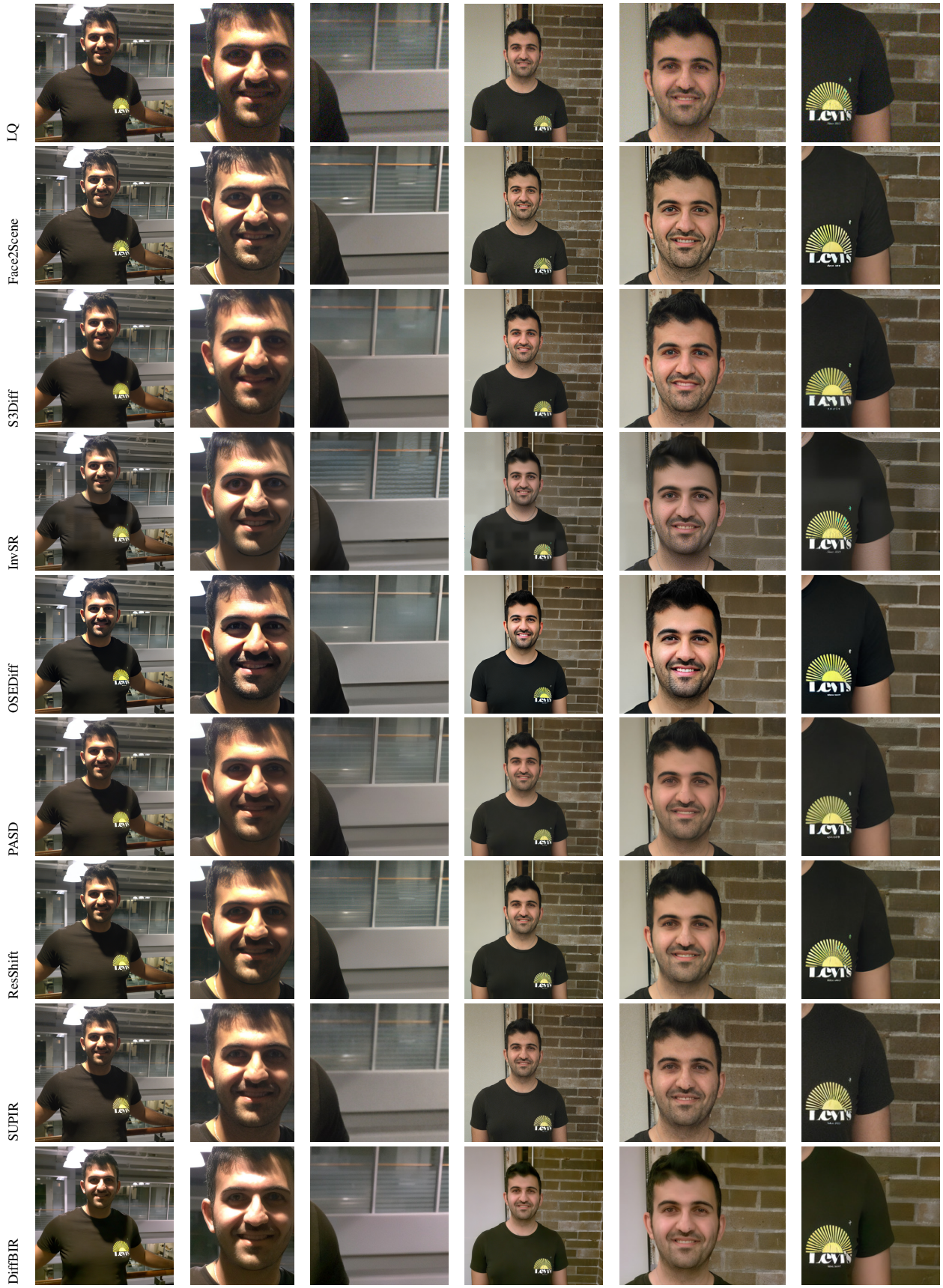

Figure 15. **Visual comparison across restoration methods on real test samples.**

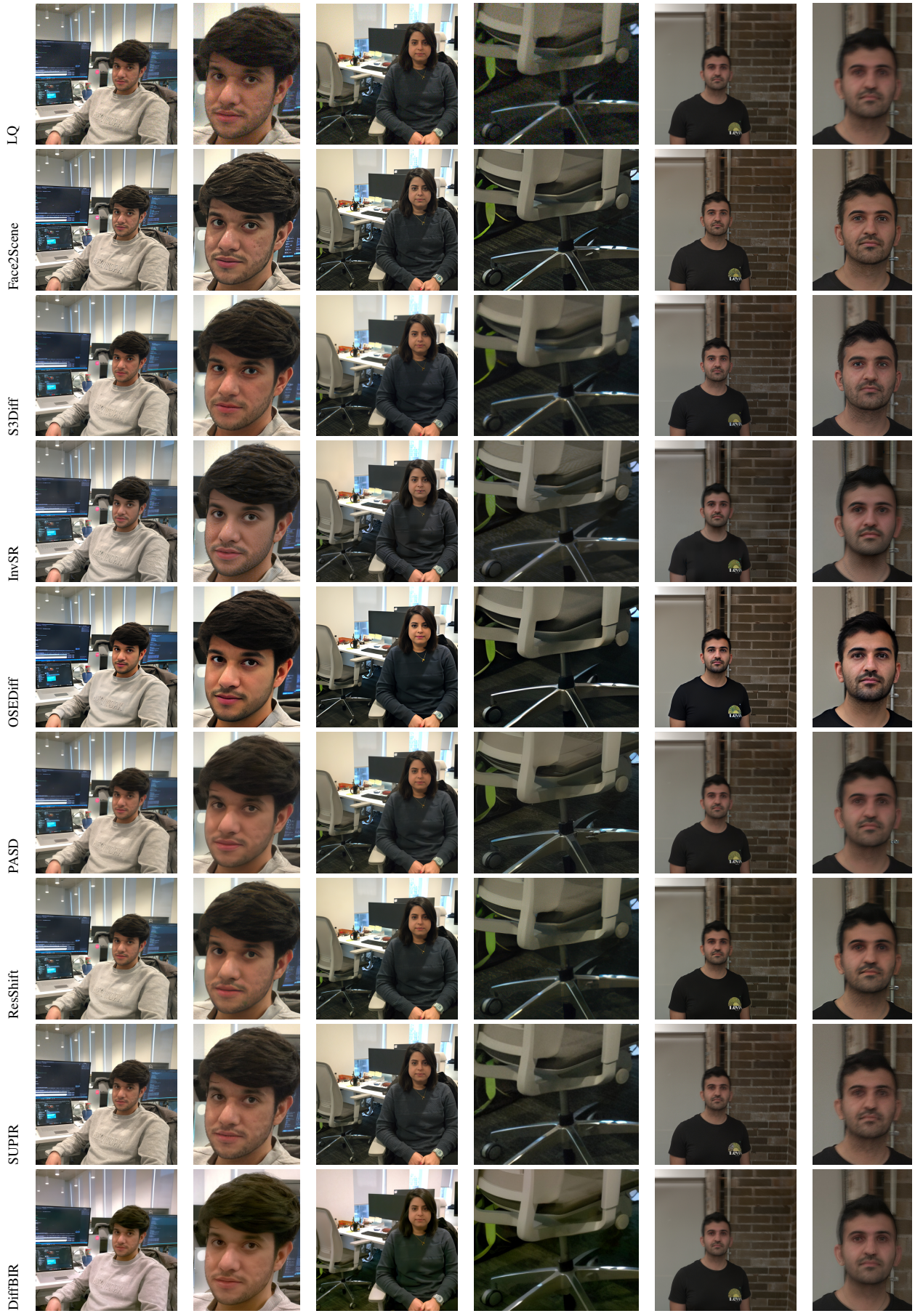

Figure 16. **Visual comparison across restoration methods on real test samples.**

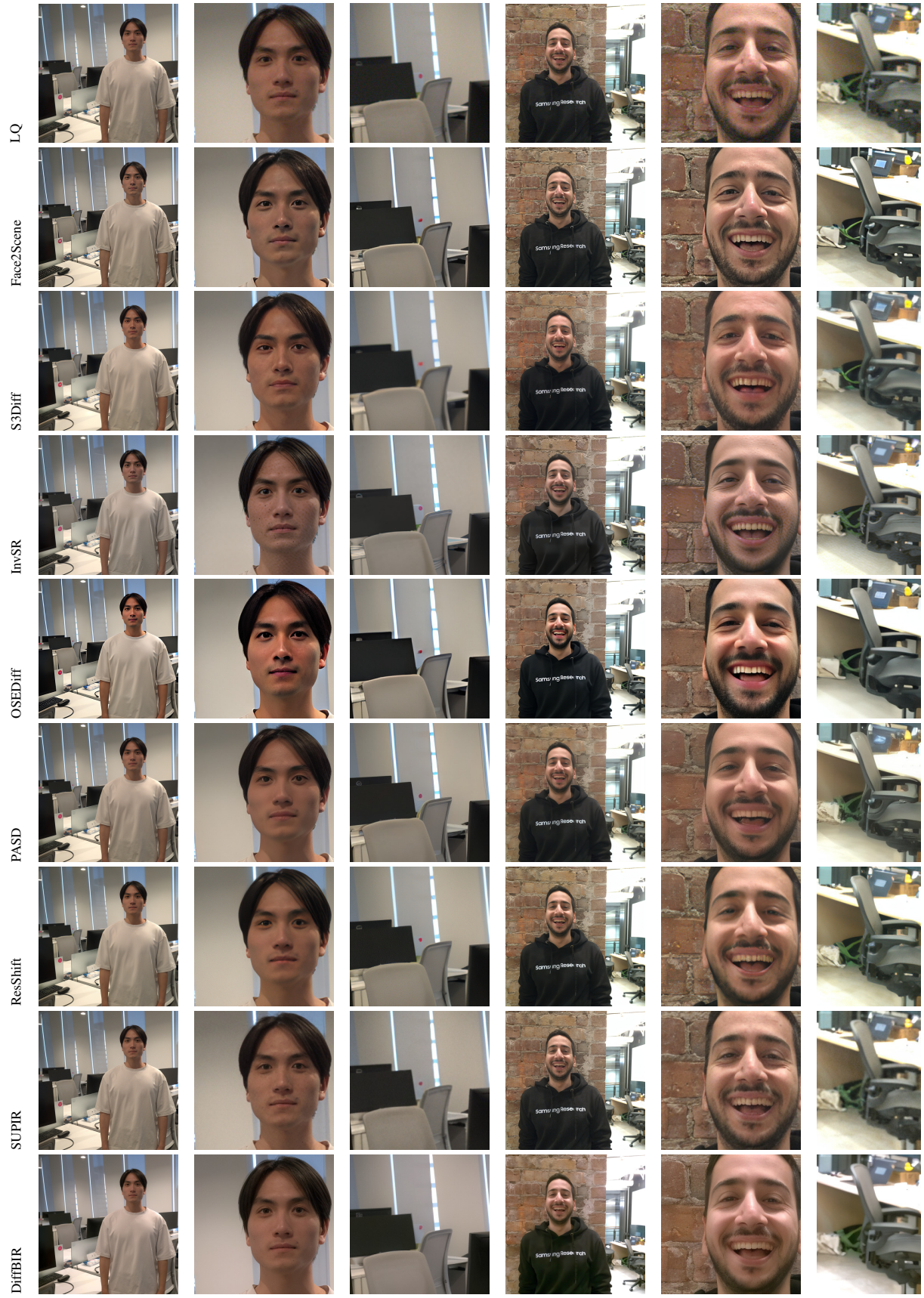

Figure 17. **Visual comparison across restoration methods on real test samples.**

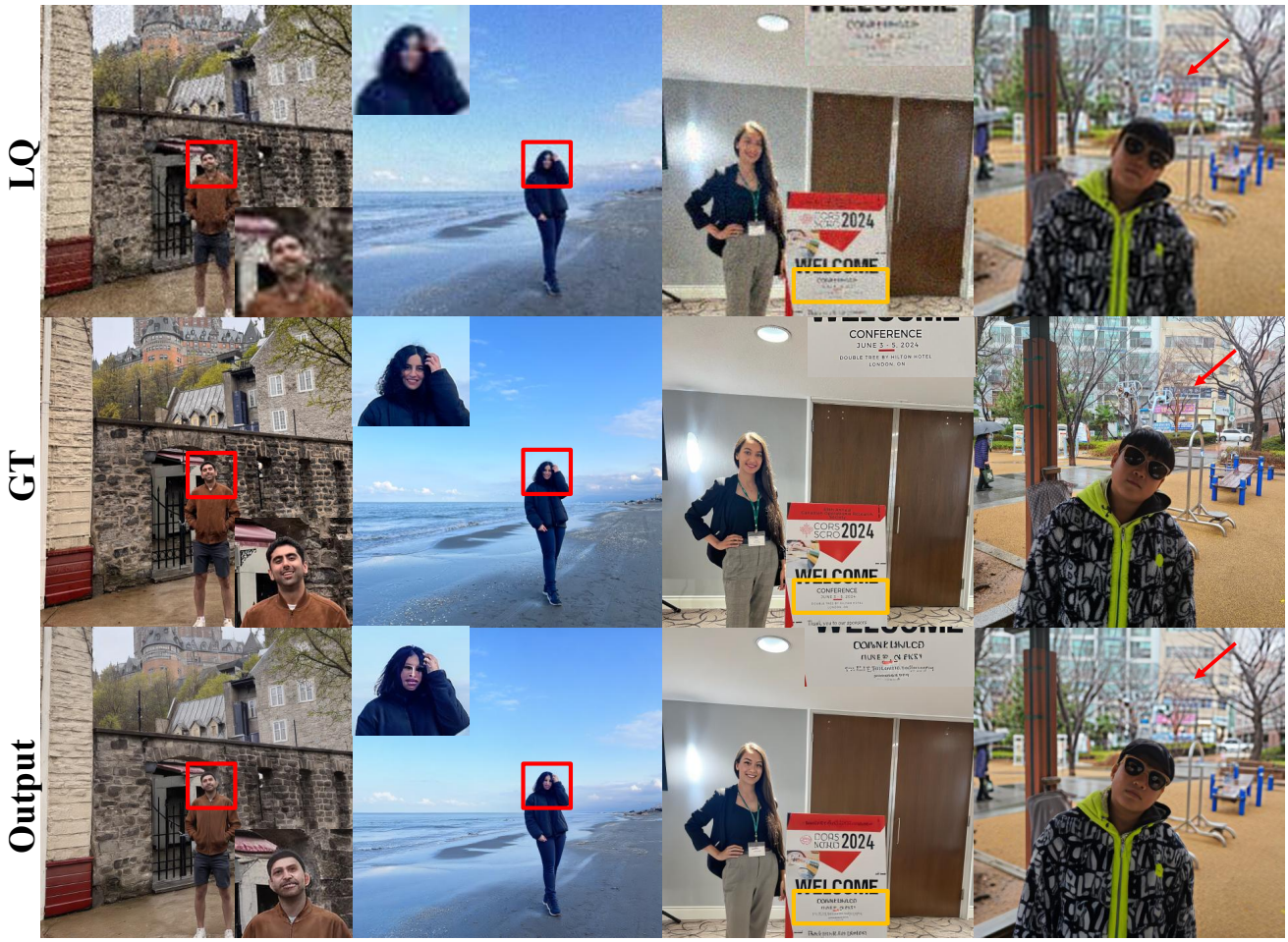

Figure 18. **Failure Cases.** Our method, like other SD 2.1-based restorers, struggles to recover fine text and very small faces, reflecting the backbone's limited text rendering ability and difficulty preserving details in tiny spatial regions. More advanced backbones such as FLUX or SD3.5, which provide higher native resolution and stronger text/face modeling, could alleviate these issues.