How far have we gone in Generative Image Restoration? A study on its capability, limitations and evaluation practices

Supplementary Material

Appendix

This Appendix provides additional details, quantitative and qualitative results that complement the findings presented in the main paper. It is organized as follows: Appendix A provides comprehensive details of our dataset construction and implementation details of GIR models. Appendix B shows additional score distributions across semantic scenes and degradation types. The section also reports extended quantitative and qualitative results for parameter sensitivity analysis. Appendix C presents qualitative demonstrations of the limitations of existing IQA methods. Appendix D provides additional failure cases of current GIR models. Appendix E reports detailed per-model performance across all evaluation dimensions. Appendix F procides training details and more qualitative results of our IQA model.

A. Details of Evaluation Design

A.1. Details of the dataset

The images in dataset are collected from both open-access online repositories [1–4], our own captured photographs and existing degradation datasets [5–18]. Together, these sources provide rich semantic diversity and a broad range of real-world degradations, providing a solid foundation for evaluating the behavior of GIR models under different conditions.

All images are processed to a unified resolution of 1024×1024. Each restoration model receives a 1024×1024 degraded image as input and produces a restored output of the same resolution, ensuring consistent comparison across all GIR models.

For the synthetic subset of 147 high-quality images across 21 semantic categories, we apply degradations using the official RealESRGAN pipeline [19]. The two-stage degradation framework of Real-ESRGAN can span an extremely severe degradation space, so severe that the entire image becomes unrecognizable to humans. This goes beyond realistic image degradation and loses practical meaning for restoration. To ensure that input images remain meaningful for evaluation, we have made several modifications to the default degradation settings. We fix the blur kernel size to 17×17 for both stages and constrain the blur standard deviations to $[0.2, 1.5]$ and $[0.2, 0.8]$ respectively. The probability of applying the final sinc filter is reduced from 0.8 to 0.4. In both degradation stages, only Gaussian noise is added, with noise levels set to $[1, 10]$ in stage 1 and $[1, 5]$ in stage 2. The resize operation is applied within scale ranges of $[0.5, 1.5]$ in stage 1 and $[0.8, 1.2]$ in stage 2. JPEG compression levels are restricted to $[50, 95]$ and $[60, 95]$, producing moderate compression artifacts compared with the original broader ranges. Overall, these adjustments intentionally lighten the default RealESRGAN degradation strength, allowing us to introduce diverse but not overly destructive degradations. This ensures that the synthetic inputs remain realistic and challenging while still preserving enough visual information for meaningful generative restoration.

To better demonstrate the semantic and degradation diversity covered in our benchmark, we include extended visual examples in Figs. 2 to 5.

A.2. Implementation Details of Restoration Models

Most models were evaluated using their official default configurations, including pretrained weights and recommended inference hyperparameters, to ensure fair comparison and reproducibility. For SUPIR, we adopt the configuration `s_stage2 = 0.85`, `s_cfg = 4.0`, `spt_linear_CFG = 1.0`, and `s_noise = 1.007`. For HYPIR, we use the Flux-based variant and set `sharpness = 800` and `noise = 200`, without any prompt provided. For PASD, we evaluate the model using its official SD-XL version.

Generation models require prompts for image-to-image restoration. To ensure consistency, we use the same prompt across both generation models:

```
Perform high-quality image
restoration and super-resolution
in this image: enhance clar-
ity, remove noise and scratches,
reduce blur, and sharpen fine
details, while preserving a nat-
ural, realistic appearance with-
out over-processing or artifi-
cial artifacts.
```

To verify that our choice of prompt does not bias the evaluation, we further tested each generation model using several alternative prompts. Across all tested prompts, we observed only minor variations in overall perceptual quality and no systematic improvement over the main prompt we used. Representative comparisons of different prompts are shown in the Figs. 6 and 7, demonstrating that our chosen prompt provides stable and representative restoration behavior.

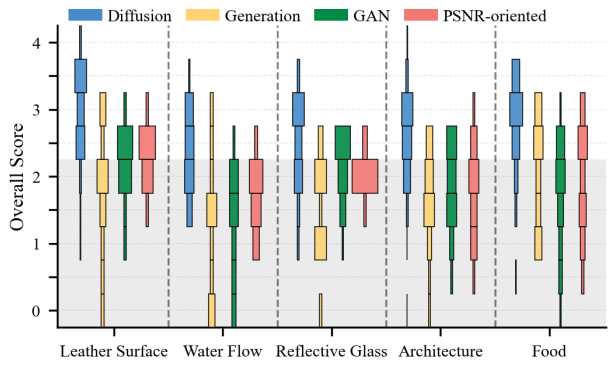


Figure 1. Distribution of annotated overall scores across semantic scene groups. The horizontal width of each box indicates the percentage of samples within each score interval. The light gray region indicates low overall scores, representing generally unacceptable results.

A.3. Detailed Criteria of Human Evaluation

Here, we provide the full detailed criteria used by annotators during manual evaluation, expanding upon the definitions of **Detail**, **Sharpness**, and **Semantics**.

Detail. This dimension evaluates the amount of fine-grained textures. -3: The restored image exhibits almost no improvement in local or global detail. Fine patterns, textures, or surface structures are largely absent, and most objects appear flat and featureless. -2: The restoration introduces limited additional detail compared to the degraded input. Fine textures and local variations remain largely missing or underdeveloped. -1: The image shows moderate enhancement in detail, but fine textures are still partially missing or over-smoothed in certain regions. 0: The level of detail is appropriate and consistent with the content of the input image. Textures are well-balanced, yielding a natural and realistic appearance. 1: Additional details beyond the original content are generated. These details slightly deviate from the ground truth and reduce overall realism. 2: A considerable amount of excessive or inconsistent detail is introduced, producing cluttered textures or unnatural structures that degrade visual authenticity. 3: The restoration introduces an overwhelming number of artificial and incoherent details, leading to an overly complex, chaotic, and unrealistic appearance.

Sharpness. -3: The restored image remains as blurry as the low-quality input, showing no perceptual improvement in clarity. -2: The image shows a slight overall improvement in clarity but remains noticeably blurry across most regions. -1: The image appears generally clear, yet minor blurriness persists in specific local regions 0: Sharpness is appropriate; object boundaries appear clear, continuous, and natural. 1: Object boundaries are slightly sharper than those in natural scenes. The excessive edge enhancement slightly degrades perceptual quality. 2: Object boundaries appear overly sharp and unnatural, sometimes producing artifacts such as white halos or excessive contrast. 3: The entire image appears excessively sharp and highly unnatural, with very strong white edges and exaggerated contrast.

Semantics. 0: Total absence of the required category's semantics. Object completely missing, unrecognizable, or the content is pure hallucination relative to the input 1: Semantic generation was attempted but resulted in a highly chaotic structure, making the intended object unidentifiable or leading to critical structural failure. 2: Semantic structure is recognizable, but significant texture or geometric distortions are present, severely impacting the perception of the semantic meaning. 3: Semantic structure is largely correct and fully recognizable. Minor texture or structural distortions exist, but they do not lead to misinterpretation, and there are no major structural errors. 4: Generated texture and structure are natural, reasonable, and fully consistent with the input image.

Model	Config ID	Configurations
HYPIR	1	`sharpness:800, noise:600`
	2	`sharpness:4000, noise:200`
	3	`sharpness:4000, noise:600`
PiSA-SR	1	`lambda_pix:1.0, lambda_sem:0.3`
	2	`lambda_pix:2.0, lambda_sem:1.0`
	3	`lambda_pix:0.3, lambda_sem:1.0`
SeeSR	1	`conditioning_scale:1.0`
	2	`conditioning_scale:1.5`
	3	`conditioning_scale:2.0`
DiffBIR	1	`cfg_scale:1, noise_aug:0`
	2	`cfg_scale:10, noise_aug:0`
	3	`cfg_scale:1, noise_aug:5`
	4	`cfg_scale:10, noise_aug:5`
SUPIR	1	`s_stage2:1.0, s_noise:1.030 s_cfg:6.5, spt_linear_CFG:3.5`
	2	`s_stage2:0.7, s_noise:1.030 s_cfg:6.5, spt_linear_CFG:3.5`
	3	`s_stage2:0.58, s_noise:1.007 s_cfg:4.0, spt_linear_CFG:1.0`
	4	`s_stage2:1.0, s_noise:1.007 s_cfg:4.0, spt_linear_CFG:1.0`

Table 1. Configurations used in our parameter sensitivity analysis. All other parameters remain at their official defaults.

A.4. Consistency of Human Evaluation

To ensure annotation consistency, we recruited professional annotators from the digital visual art industry, standardized the annotation workflow via a custom Gradio system with unified training, and implemented expert inspection to filter anomalies. To analyze inter-annotator agreement, a random 10% subset of

Model	Config ID	SmallFace Face	Crowd	Vehicles	Street View	Aerial View	Hands/ Feet	Complex Texture	Text	Digital Zoom	OP (Color)
HYPIR	1	2.83	2.00	3.00	2.71	**2.90**	2.68	2.58	2.50	2.17	2.23
	2	**2.94**	**2.50**	3.10	**2.86**	**2.90**	**3.05**	**2.75**	**2.61**	**2.61**	**2.43**
	3	2.61	2.10	**3.20**	2.71	2.70	2.73	2.58	2.50	2.26	2.14
PiSA-SR	1	**1.78**	**1.60**	2.30	2.00	**1.80**	2.64	**2.17**	**2.22**	**2.15**	1.67
	2	1.67	1.30	**2.40**	**2.21**	1.70	**2.68**	2.00	2.11	2.04	**1.94**
	3	1.28	1.10	2.00	1.79	1.40	2.41	2.08	2.06	1.96	1.44
SeeSR	1	**2.28**	**1.70**	**2.20**	**2.21**	**1.70**	2.30	**2.17**	1.78	**2.20**	**2.30**
	2	1.78	1.30	2.10	1.79	**1.70**	**2.59**	2.08	**1.94**	1.65	1.66
	3	1.72	1.40	2.00	1.71	**1.70**	2.41	1.92	**1.94**	1.57	1.54
DiffBIR	1	**2.11**	1.30	2.50	2.36	1.80	**2.82**	**2.58**	**2.39**	1.70	2.11
	2	2.06	**1.40**	2.70	2.29	**2.40**	2.59	2.17	2.17	1.98	**2.26**
	3	**2.11**	1.30	2.50	2.21	1.90	2.77	**2.58**	**2.39**	1.65	2.01
	4	2.06	**1.40**	**2.80**	**2.50**	2.30	2.50	2.17	2.22	**2.07**	**2.26**
SUPIR	1	2.44	1.40	**2.80**	**2.71**	2.50	2.68	2.08	2.17	1.85	1.77
	2	**2.50**	1.30	**2.80**	2.36	2.10	2.32	1.75	1.78	1.50	**1.87**
	3	**2.50**	**1.70**	2.70	2.64	**2.70**	**2.86**	**2.33**	**2.33**	**1.87**	1.44
	4	2.33	1.10	2.30	2.14	1.60	2.32	1.83	1.89	1.33	1.20

Table 2. Average overall scores of different parameter configurations across semantic categories. Bold numbers indicate the best-performing configuration within each model for a given category

the dataset was re-annotated. The Krippendorff's Alpha coefficients [20] were 0.804/0.729/0.780/0.807 for Detail/Sharpness/Semantics/Overall dimensions. Statistically, these scores are well above the standard acceptance threshold of 0.667, demonstrating that our data reflects rigorous expert consensus rather than subjective noise.

B. Additional Results

B.1. More Score Distributions

We further provide score distributions for all remaining semantic categories that were not visualized in the main paper. Results for Overall Detail, Sharpness, and Semantics are shown in Figs. 1 and 8. We also report complete score distributions across all degradation types in Fig. 9.

B.2. More Results of Parameter Configurations

To support the parameter sensitivity study in the main paper, we evaluate multiple inference configurations for several diffusion-based GIR models. These configurations, as shown in Tab. 1 are chosen to intentionally vary the level of fidelity and generation, allowing us to obtain a richer and more diverse set of restoration outcomes. We further report the performance of each configuration across different semantic scenes and degradation types, as summarized in Tab. 2. Qualitative comparisons illustrating how different parameters affect restoration behavior can be found in Figs. 10 and 11. All configuration identifiers exactly match those reported in the main paper.

C. Qualitative Analysis of IQA

To complement the quantitative results in the main paper, we provide qualitative examples illustrating the limitations of existing IQA methods when evaluating generative image restoration. Although recent learning-based IQA models achieve strong performance on traditional distortions, they remain unreliable for GIR scenarios, particularly when semantic inconsistencies or generative artifacts are present.

In the Figs. 12 to 14, each case includes two restored outputs produced by different GIR models. For each restored image, we report the human-annotated overall scores, along with the predictions from several representative IQA models and our IQA model. These examples provide side-by-side comparisons of how IQA models and human annotators evaluate the same pair of restored images, allowing to directly observe the differences in their judgments.

D. More Failure Cases

To further support the observations discussed in the main paper, we present additional failure cases of current GIR models across a wide range of semantic scenes and degradation types. These examples cover various representative failure modes, including semantic errors in Figs. 15 and 16, over-generation of details in Fig. 17, and insufficient degradation removal in Fig. 18, providing a more complete visual reference of the behaviors discussed in the main paper.

E. Detail Results of each model

In the main paper, we primarily analyzed restoration behaviors at the model-family level. Here, we provide model-

Model	Avgerage Overall Score↑	Avgerage Rank↓
HYPIR	2.775	3.127
PiSA-SR	2.508	5.547
TSD-SR	2.467	7.487
DiffBIR	2.428	5.181
PASD	2.38	6.382
OSEDiff	2.377	6.173
Invsr	2.279	7.635
SUPIR	2.235	9.062
SeeSR	2.201	8.870
CCSR	2.195	10.365
S3Diff	2.181	8.484
FLUX	1.935	12.099
StableSR	1.925	11.133
SwinIR	1.625	15.569
RealESRGAN	1.548	15.734
ResShift	1.466	13.858
BSRGAN	1.435	15.037
HAT	1.426	16.159
Nano Banana	1.401	15.552
CAL-GAN	1.246	16.544

Table 3. Overall performance ranking of all evaluated restoration models. We report each model's mean overall quality score across the full dataset and its average per-image ranking position.

specific results, reporting the detailed performance of all 20 restoration models across all evaluation dimensions.

We first present the overall performance ranking of each model in Tab. 3, computed in two complementary ways: (1) the average overall quality score across the entire dataset, and (2) the average per-image ranking position, which reflects how frequently a model outperforms others on individual samples.

We then report per-degradation and per-semantic results for every evaluation dimension in Tabs. 5 to 11. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores. And since 0 is the point of balance, a lower score indicates a better performance for these two dimensions. This includes the mean score of each model under every degradation type and semantic category, allowing comparison between different models.

F. More Details of our IQA Model

F.1. Training Details

To evaluate the restored images across four distinct dimensions, we train a dedicated model for each specific direction. Notably, the original **Detail** and **Sharpness** assessments are bipolar metrics, with subjective scores ranging from -3 to 3, where 0 represents the optimal perceptual quality. To facilitate effective learning, we introduce a data preprocessing step that decouples these bipolar metrics into unipolar ones. For the **Detail** dimension, we utilize the data subset with **Detail** scores ≥ 0 to train a model for predicting **Content Conciseness**, while the subset with scores ≤ 0 is used to train the model for assessing **Detail Completeness**. For the **Sharpness** dimension, we utilize the data subset with **Sharpness** scores ≥ 0 to train a model for predicting **Visual Clarity**, while the subset with scores ≤ 0 is used to train the model for assessing **Visual Clarity**. The overall training protocol and hyperparameter configurations for our IQA models strictly follow those established in DeQA[21].

Test set	NIQE	MANIQA	MUSIQ	CLIP-IQA	DeQA-Score	Ours
LIVEW	0.81/0.77	0.85/0.83	0.79/0.83	0.83/0.81	**0.89/0.88**	0.83/0.81
AGIQA	0.72/0.65	0.72/0.64	0.72/0.63	0.74/0.69	**0.81**/0.73	**0.81/0.76**
CSIQ	0.70/0.65	0.62/0.63	0.77/0.71	0.77/0.72	0.79/0.74	**0.86/0.83**

Table 4. Performance comparison of our proposed method against popular IQA models on three external datasets. Results are reported in terms of SRCC/PLCC. The best and second-best results are highlighted in **bold** and underlined respectively.

F.2. Generalization

We evaluated our model on three distinct external datasets; the results (SRCC/PLCC) are as shown in Tab. 4. While other models predict only overall quality, we provide a multi-dimensional IQA model that independently assesses Detail, Sharpness, and Semantics, enabling fine-grained evaluation.

References

[1] Wikimedia Commons. Wikimedia commons: a collection of 129,990,166 freely usable media files to which anyone can contribute. 1

[2] Unsplash. Unsplash: Beautiful free images & pictures.

[3] Pixabay. Pixabay: Beautiful free images & royalty free stock.

[4] International Photography Magazine & Grant. International photography magazine. 1

[5] Hajime Nada, Vishwanath Sindagi, He Zhang, and Vishal M Patel. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. *arXiv preprint arXiv:1804.10275*, 2018. 1

[6] Josue Anaya and Adrian Barbu. Renoir–a dataset for real low-light image noise reduction. *Journal of Visual Communication and Image Representation*, 2018.

[7] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*, 2021.

[8] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *CVPR*, 2014.

[9] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *ECCV*, 2020.

[10] Jie Liang, Radu Timofte, Qiaosi Yi, Zhengqiang Zhang, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Lei Zhang, et al. Ntire 2025 the 2nd restore any image model (raim) in the wild challenge. In *CVPR Workshops*, 2025.

[11] Aashish Sharma and Robby T Tan. Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In *CVPR*, 2021.

[12] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017.

[13] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015.

[14] Aoxiang Ning, Minglong Xue, Yiting Wei, Mingliang Zhou, and Senming Zhong. Artistic-style text detector and a new movie-poster dataset. *Expert Systems with Applications*, 2025.

[15] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *ECCV*, 2022.

[16] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

[17] Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10k: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*, 2020.

[18] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 1

[19] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV Workshops*, 2021. 1

[20] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007. 3

[21] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *CVPR*, 2025. 4

Model	Large Face				Medium Face				Small Face			
	Overall↑	Sharpness ↓	Detail ↓	Semantics↑	Overall↑	Sharpness ↓	Detail ↓	Semantics↑	Overall↑	Sharpness ↓	Detail ↓	Semantics↑
HYPIR	3.07	0.57	0.57	3.14	3.00	0.94	0.78	3.17	2.78	1.17	1.00	2.89
SUPIR	2.57	1.14	0.93	2.79	2.67	1.39	1.06	2.72	2.33	1.56	1.06	2.56
PiSA-SR	2.64	1.07	1.36	2.86	2.78	1.22	1.06	2.94	2.50	1.44	1.44	2.67
SeeSR	2.57	1.43	1.36	2.71	2.39	1.44	1.44	2.72	2.28	1.39	1.39	2.22
OSEDiff	2.64	1.36	1.43	3.21	2.44	1.28	1.44	2.50	1.94	1.61	1.67	2.28
CCSR	2.57	0.29	1.00	2.93	2.56	0.67	0.94	2.72	2.06	1.28	1.50	2.17
DiffBIR	3.00	0.79	0.93	3.14	2.78	0.78	1.17	2.94	2.22	1.22	1.50	2.44
StableSR	2.21	1.64	1.50	2.79	2.06	1.83	1.61	2.44	1.94	2.06	1.83	2.22
PASD	2.71	0.79	0.86	2.93	2.22	1.11	1.11	2.44	2.06	0.72	0.67	2.11
Invsr	2.43	1.07	1.07	2.79	2.44	1.17	1.17	2.61	2.22	1.28	1.50	2.39
S3Diff	3.00	0.93	0.79	3.07	2.78	0.72	0.72	2.89	2.39	1.06	1.22	2.44
TSD-SR	2.50	1.07	1.07	2.71	2.56	1.11	1.06	2.72	2.39	1.33	1.28	2.50
ResShift	1.86	2.07	1.79	2.14	1.61	2.06	1.89	2.22	1.33	1.83	1.83	1.72
FLUX	2.14	1.21	1.14	2.71	2.44	0.89	1.00	2.78	1.78	1.44	1.44	2.39
Nano Banana	1.71	1.71	1.57	2.79	1.89	2.00	1.39	2.50	0.78	2.44	2.28	2.11
BSRGAN	2.00	2.00	1.71	2.71	1.83	2.06	2.11	2.39	1.22	2.17	2.11	1.39
CAL-GAN	1.79	2.00	1.71	2.50	1.50	2.22	1.94	2.06	0.83	2.17	2.28	1.67
RealESRGAN	1.71	2.14	2.00	2.43	1.83	1.89	1.83	2.39	1.39	1.94	2.17	1.56
HAT	1.71	2.21	1.93	2.36	1.56	2.11	2.06	2.06	0.78	2.39	2.39	1.44
SwinIR	2.07	1.71	1.64	2.57	1.83	1.94	1.83	2.28	1.50	1.78	2.17	1.67

Table 5. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Large Face, Large Face and Small Face. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

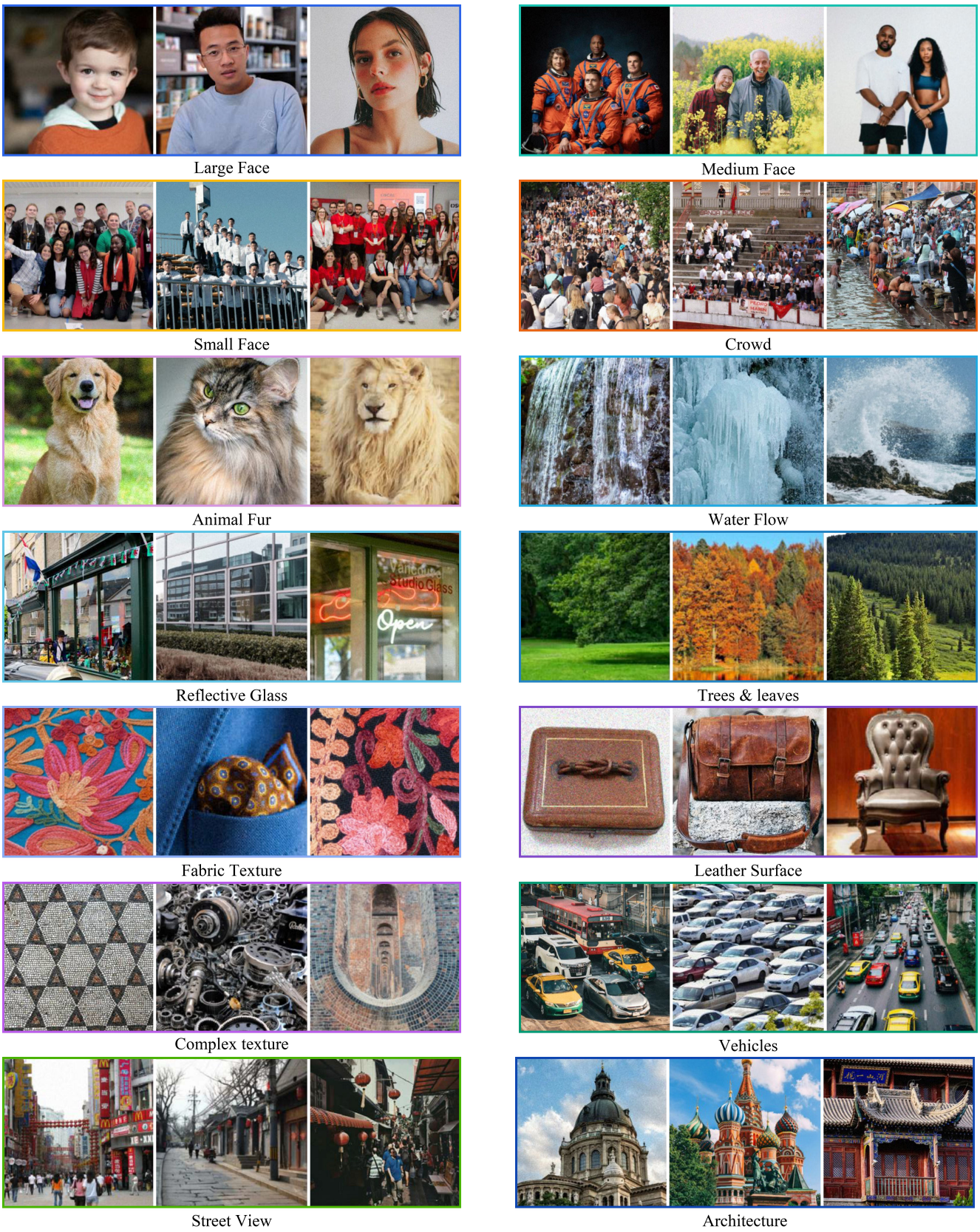


Figure 2. More low-quality examples of different semantic categories. Zoom in for a better view.

Figure 3. More low-quality examples of different semantic categories. Zoom in for a better view.

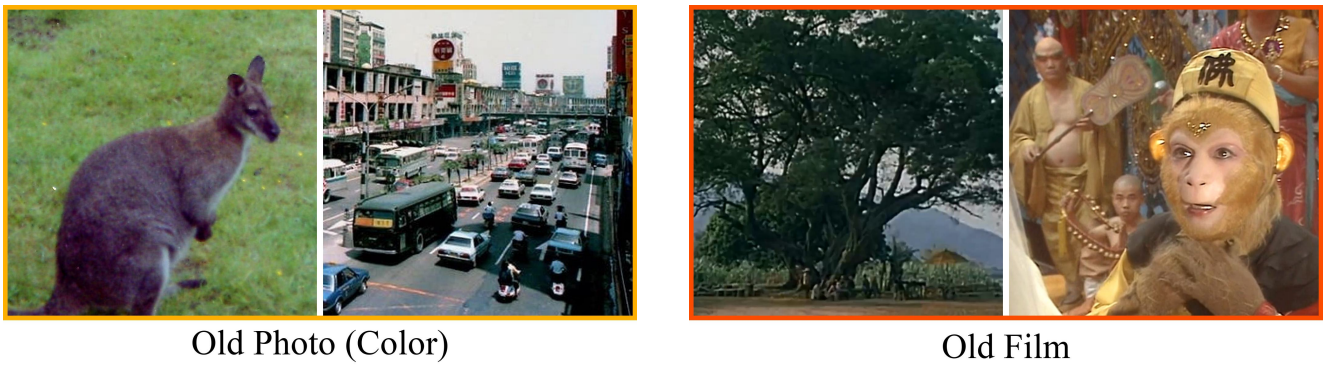


Figure 4. More visual examples of different degradation categories. Zoom in for a better view.

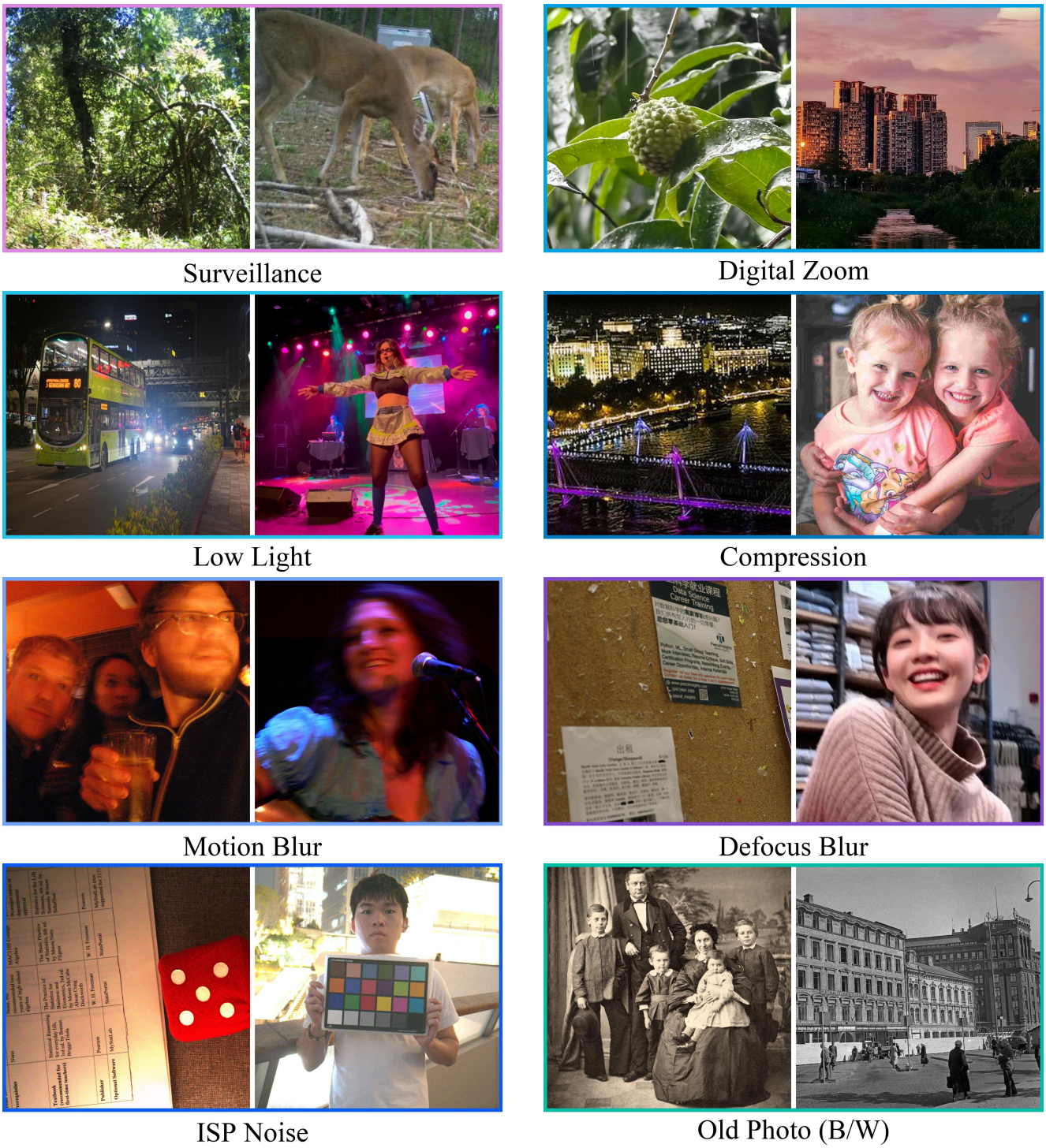


Figure 5. More visual examples of different degradation categories. Zoom in for a better view.

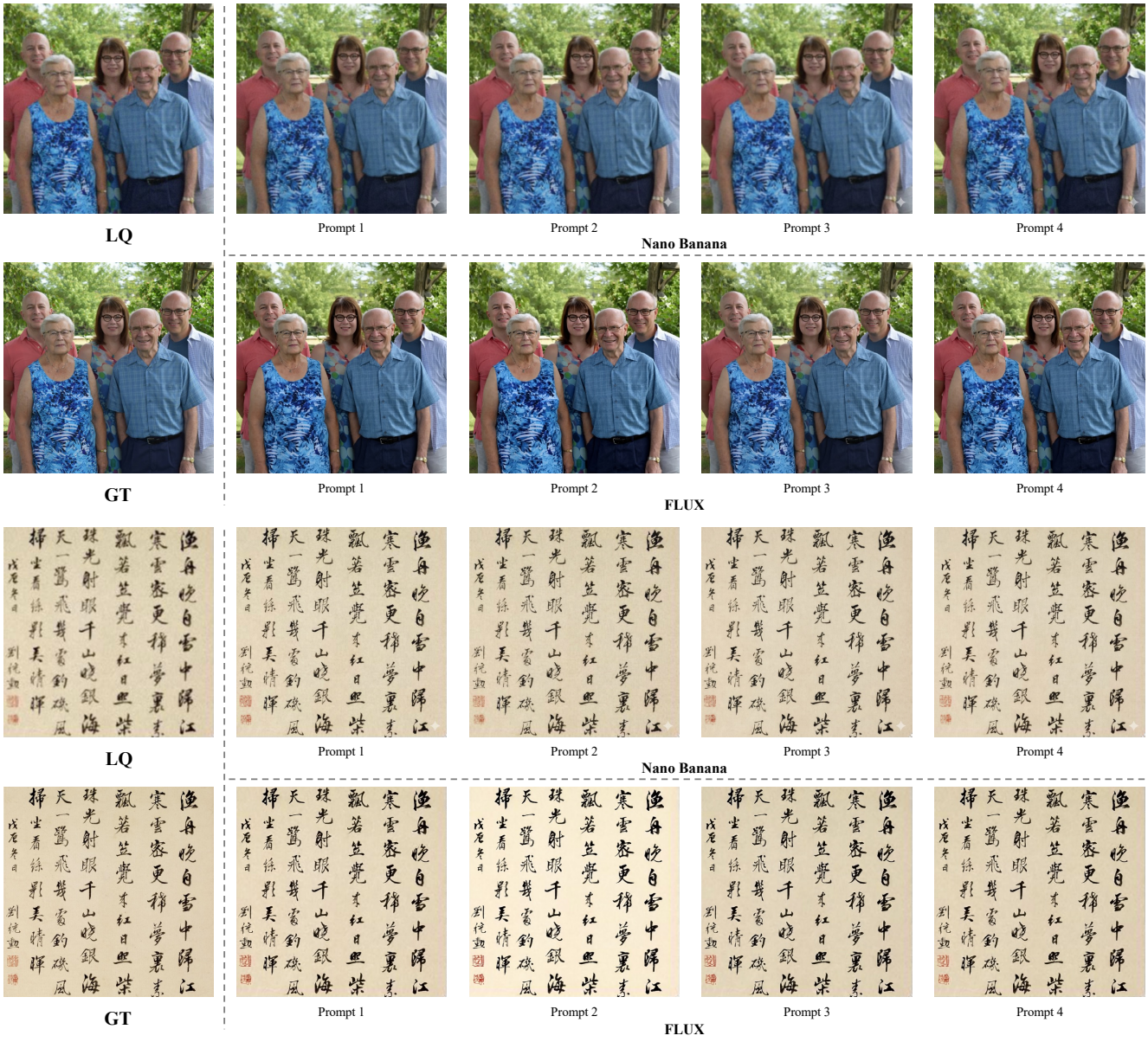


Figure 6. Results of different prompts used in generation models. Zoom in for a better view.

Prompt 1: `Perform high-quality image restoration and super-resolution in this image: enhance clarity, remove noise and scratches, reduce blur, and sharpen fine details|while preserving a natural, realistic appearance without over-processing or artificial artifacts.`

Prompt 2: `Restore this image with enhanced clarity and sharpness. Reduce noise, fix compression artifacts, and improve detail while strictly preserving the original content, style, and semantics.`

Prompt 3: `Clean and restore the image by removing artifacts, noise, and blur. Maintain full fidelity to the original composition and avoid creating new visual elements.`

Prompt 4: `Improve this image's perceptual quality: refine edges, correct degradation, and enhance fine details. Do not alter identity, expression, or any semantic attributes.`

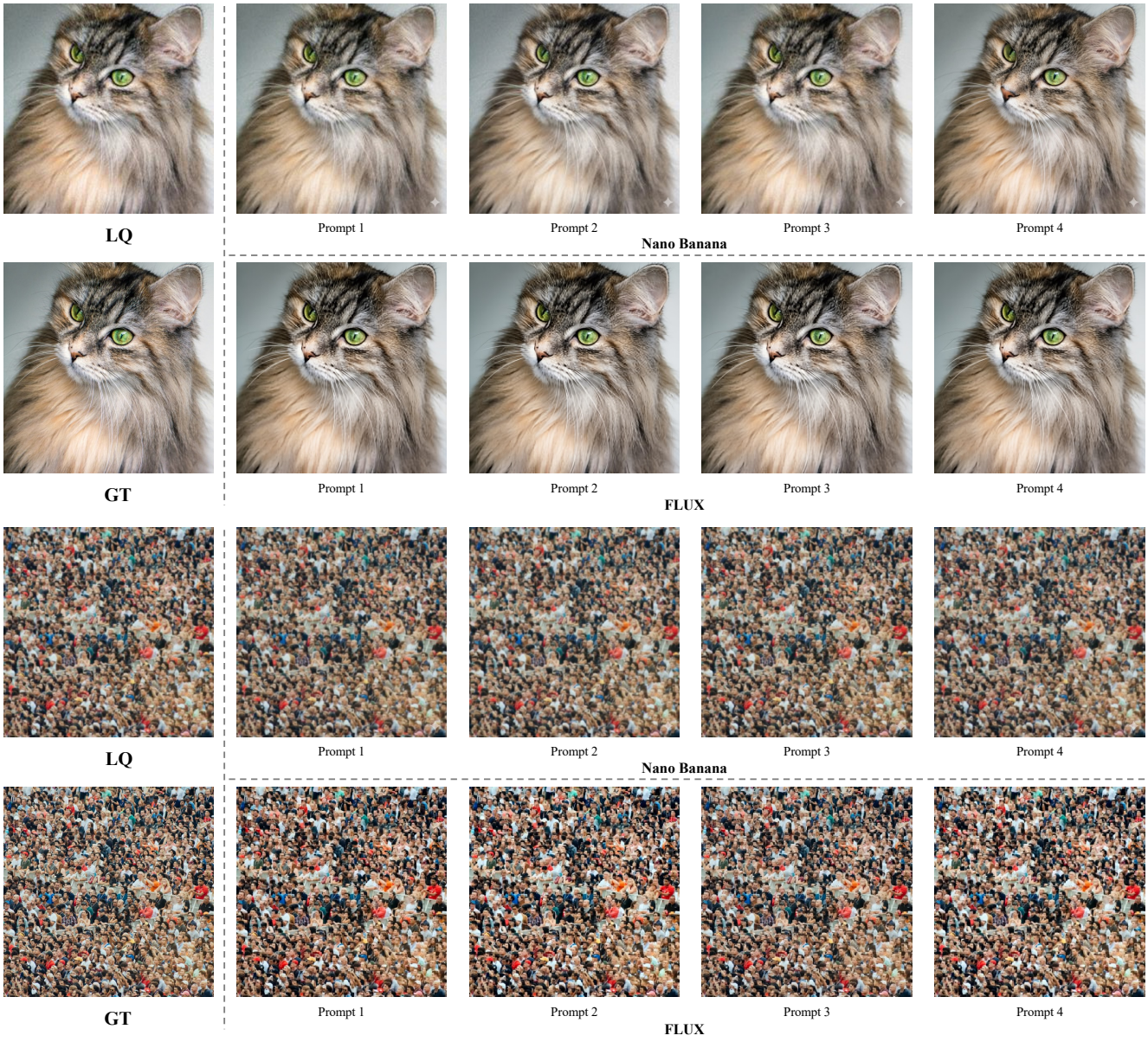


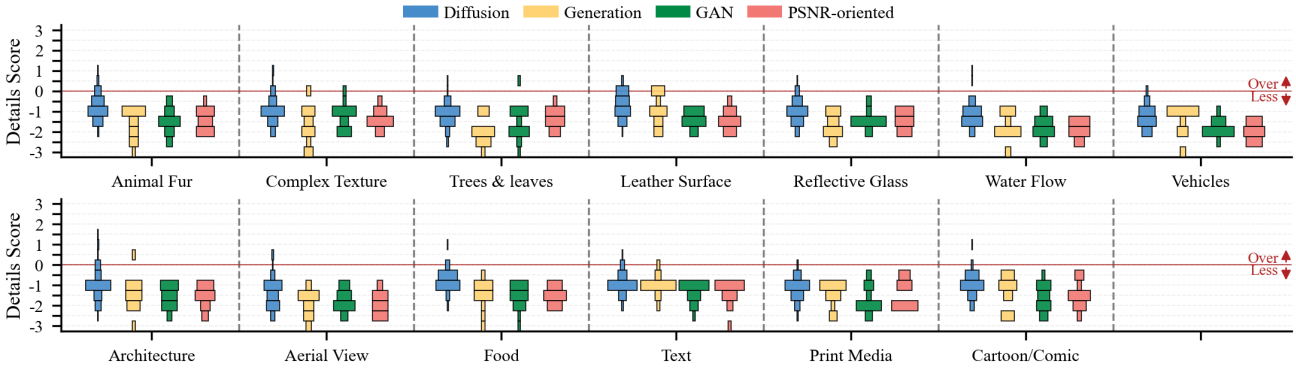
Figure 7. Results of different prompts used in generation models. Zoom in for a better view.

Prompt 1: `Perform high-quality image restoration and super-resolution in this image: enhance clarity, remove noise and scratches, reduce blur, and sharpen fine details|while preserving a natural, realistic appearance without over-processing or artificial artifacts.`

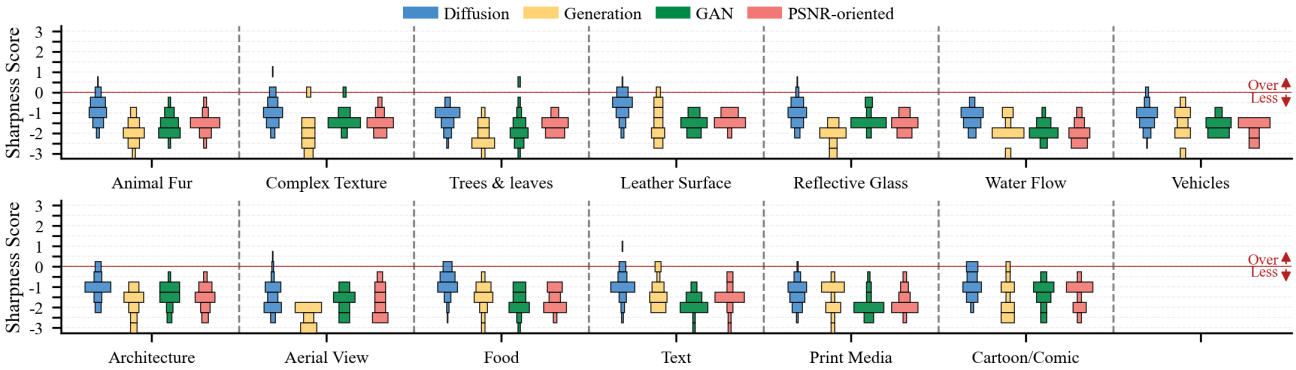
Prompt 2: `Restore this image with enhanced clarity and sharpness. Reduce noise, fix compression artifacts, and improve detail while strictly preserving the original content, style, and semantics.`

Prompt 3: `Clean and restore the image by removing artifacts, noise, and blur. Maintain full fidelity to the original composition and avoid creating new visual elements.`

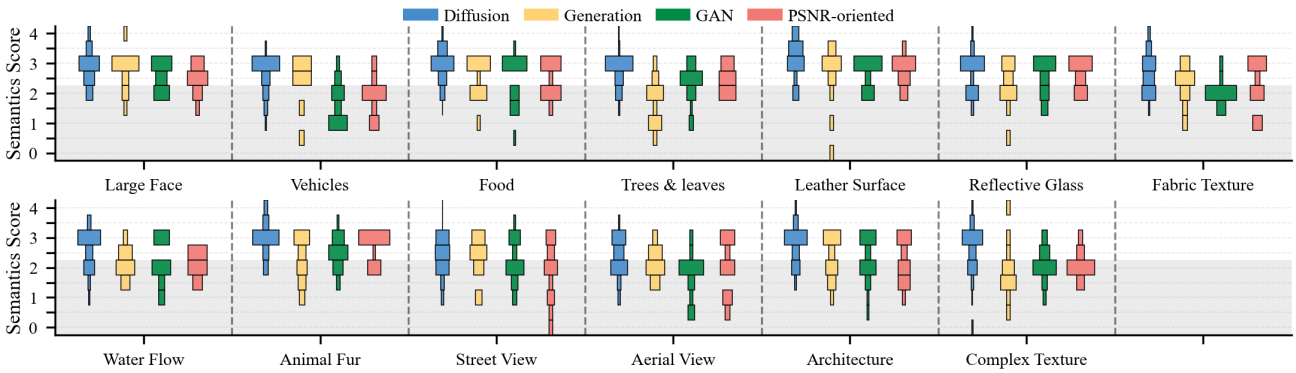
Prompt 4: `Improve this image's perceptual quality: refine edges, correct degradation, and enhance fine details. Do not alter identity, expression, or any semantic attributes.`



(a) Distribution of detail scores across semantic scene groups.

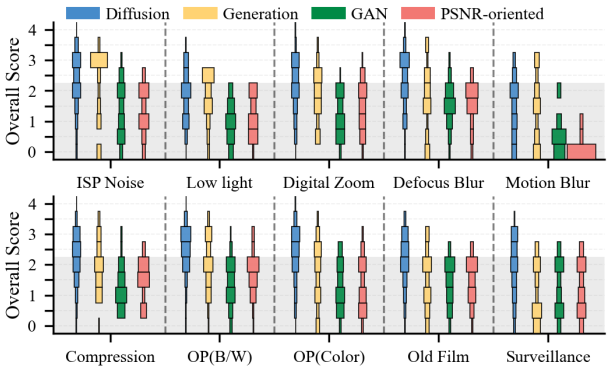


(b) Distribution of sharpness scores across semantic scene groups.

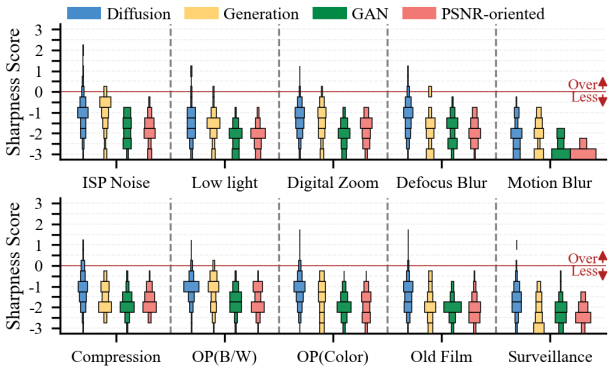


(c) Distribution of semantic scores across semantic scene groups.

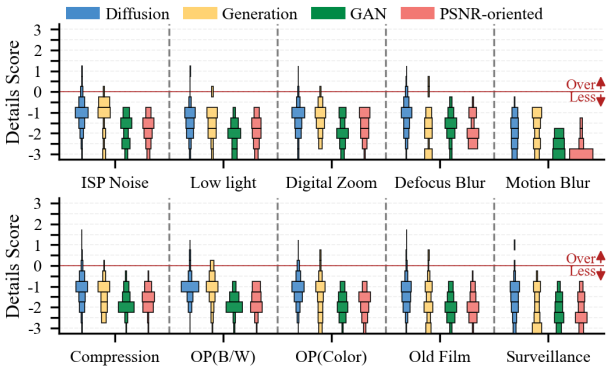
Figure 8. Distribution of different scores across various semantic scene groups. The horizontal width of each box indicates the percentage of samples within each score interval. The red line indicates the balance point; scores above it (Over ↑) denote over-generation, and scores below it (Less ↓) denote under-generation. The light gray region indicates low overall scores, representing generally unacceptable results.



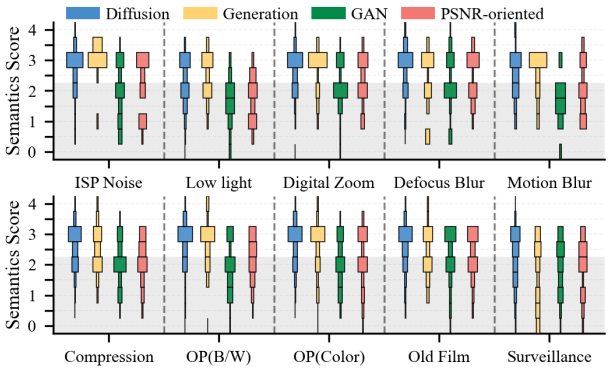
(a) Distribution of overall scores across degradations.



(b) Distribution of semantic scores across degradations.



(c) Distribution of detail scores across degradations.



(d) Distribution of sharpness scores across degradations.

Figure 9. Distribution of different scores across various degradation types. The horizontal width of each box indicates the percentage of samples within each score interval. The light gray region indicates low overall scores, representing generally unacceptable results. The red line indicates the balance point; scores above it (Over ↑) denote over-generation, and scores below it (Less ↓) denote under-generation.

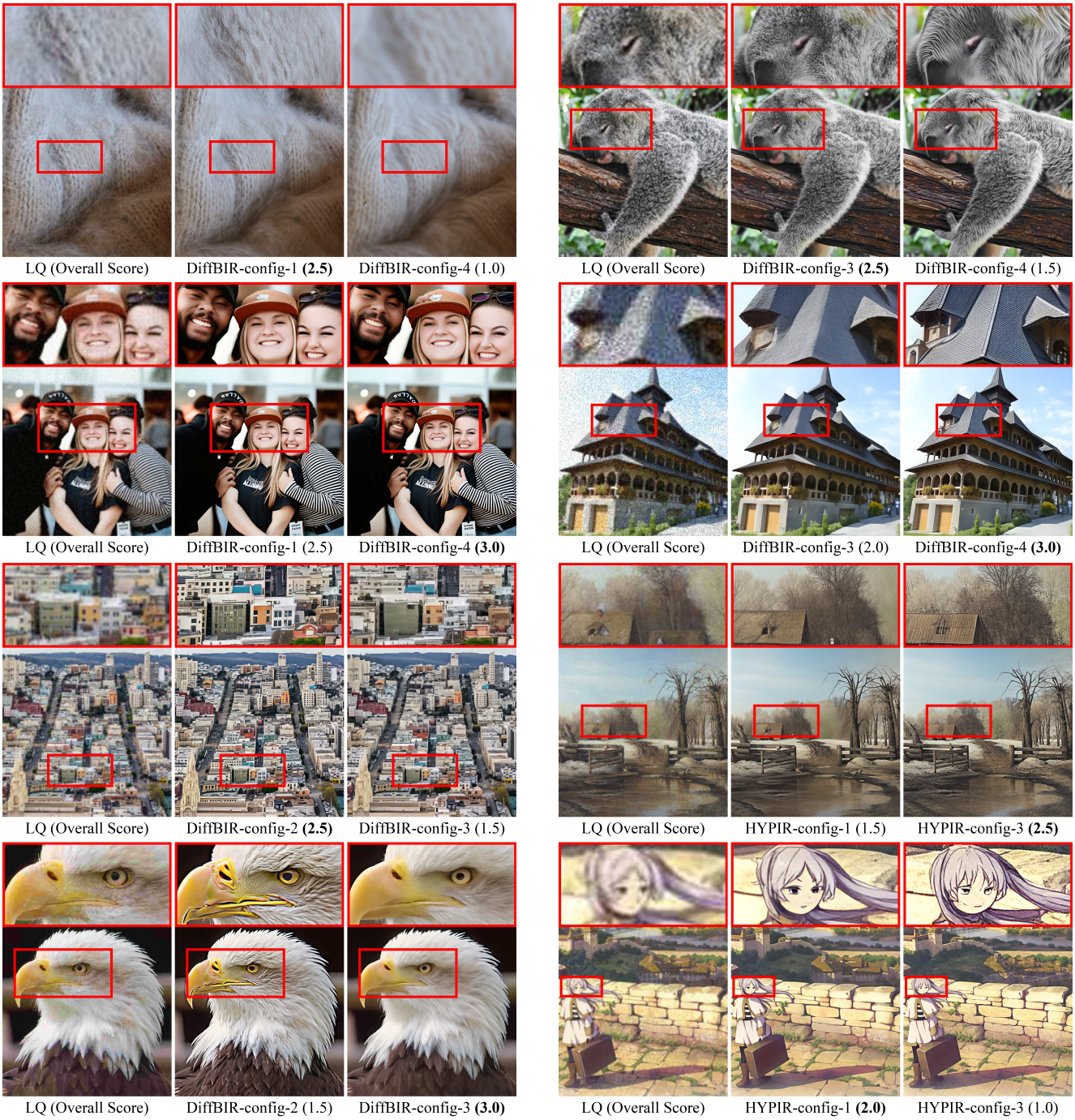


Figure 10. Effect of parameter configurations on restoration behavior in diffusion-based models. Zoom in for a better view.

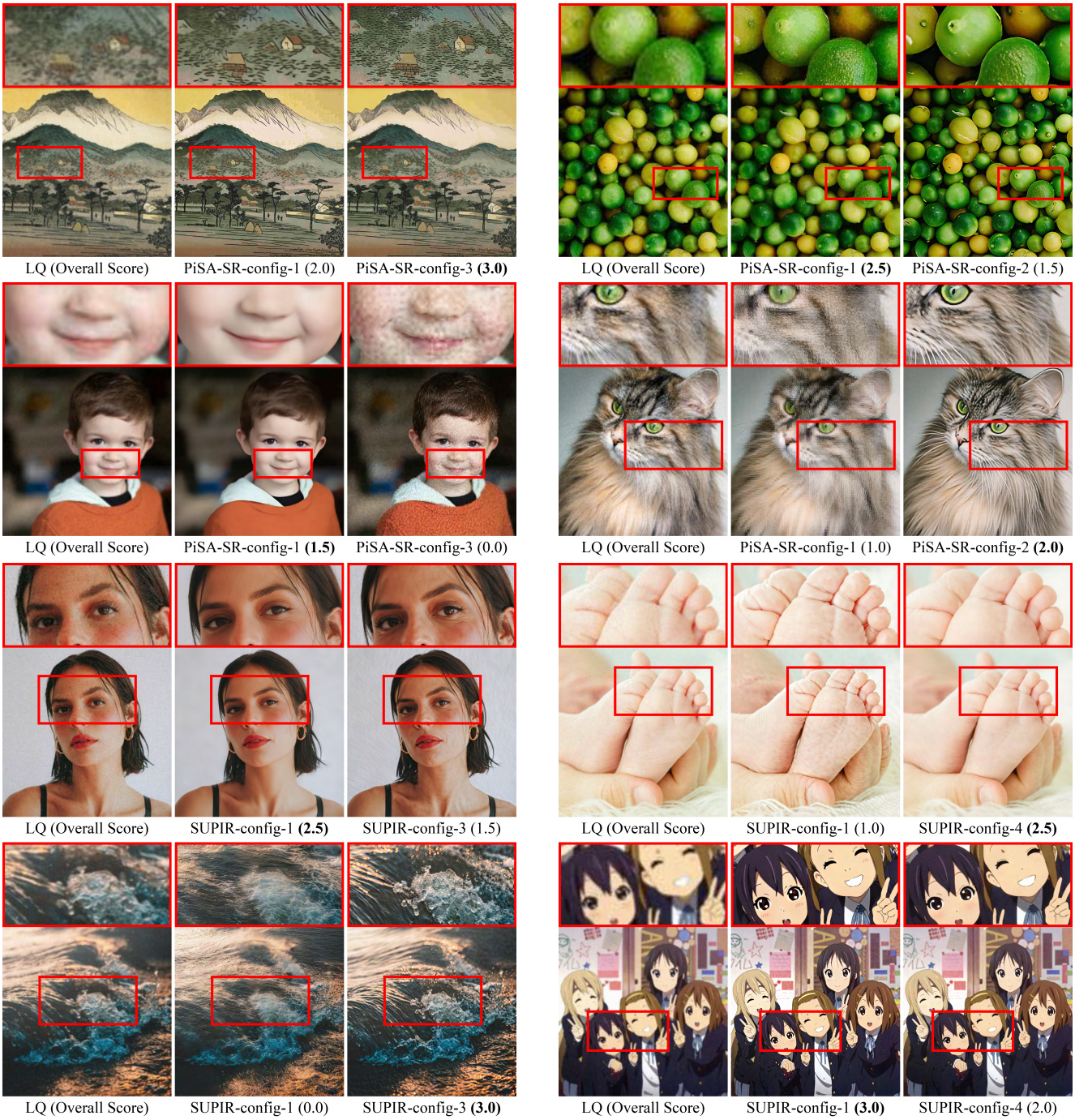


Figure 11. Effect of parameter configurations on restoration behavior in diffusion-based models. Zoom in for a better view.

Image A Image B

	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	3.61	3.46	0.214	0.527	69.7	4.11	4.29	3.5
Image B	2.78	2.93	0.262	0.635	71.3	4.27	3.62	2
Better	B	B	B	B	B	B	A	A

Image A Image B

	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	3.38	3.45	0.161	0.463	67.2	4.00	3.01	2.5
Image B	3.20	3.20	0.250	0.621	71.8	4.00	2.29	1.5
Better	B	B	B	B	B	-	A	A

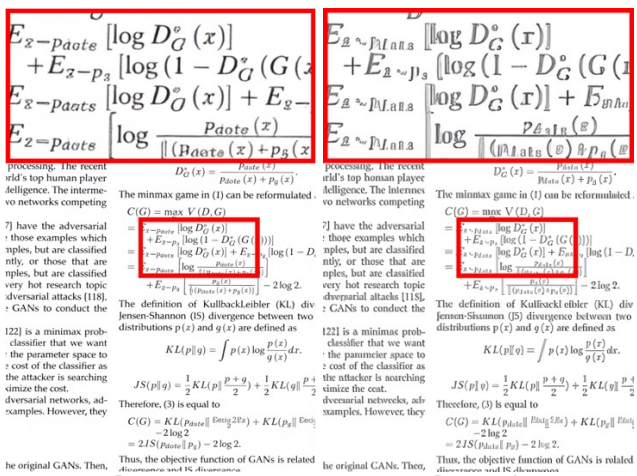


Image A Image B

	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	13.2	8.70	0.290	0.644	58.8	3.70	3.63	3.5
Image B	9.09	6.56	0.338	0.606	60.4	3.56	3.41	2
Better	B	B	B	B	B	A	A	A

Image A Image B

	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	3.72	3.73	0.184	0.513	64.6	3.32	3.33	2.5
Image B	3.68	3.37	0.206	0.676	63.4	3.39	2.25	2
Better	B	B	B	B	A	B	A	A

Figure 12. Qualitative analysis of image quality assessment methods. Zoom in for a better view.

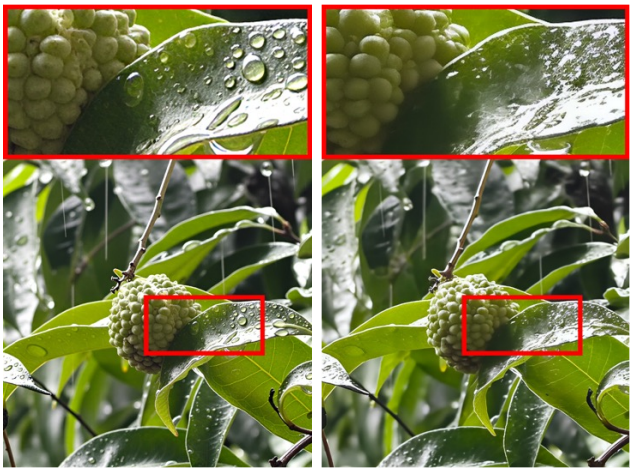


Image A Image B

	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	3.98	3.80	0.244	0.826	74.6	4.51	4.01	3
Image B	3.99	3.73	0.254	0.837	76.3	4.55	3.68	1.5
Better	A	B	B	B	B	B	A	A

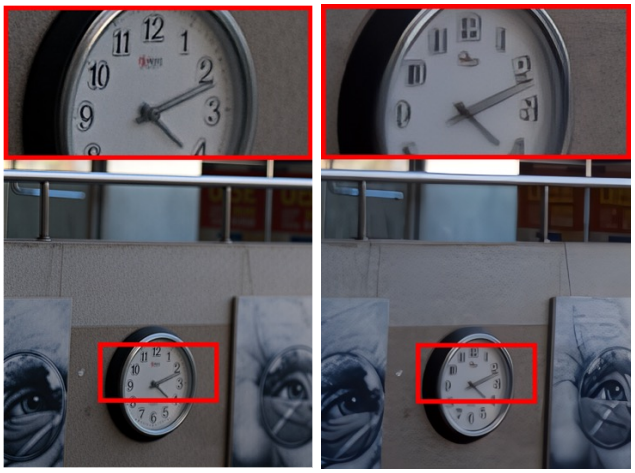


Image A Image B

	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	4.76	3.71	0.149	0.442	52.0	3.68	3.92	3
Image B	4.52	4.09	0.163	0.486	57.7	3.91	3.12	1.5
Rank	B	A	B	B	B	B	A	A

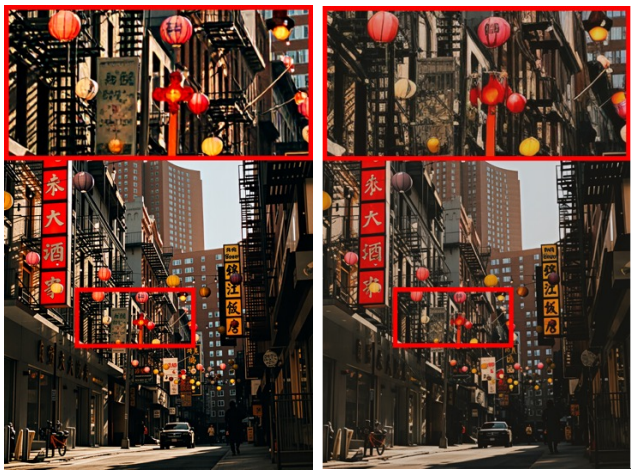


Image A Image B

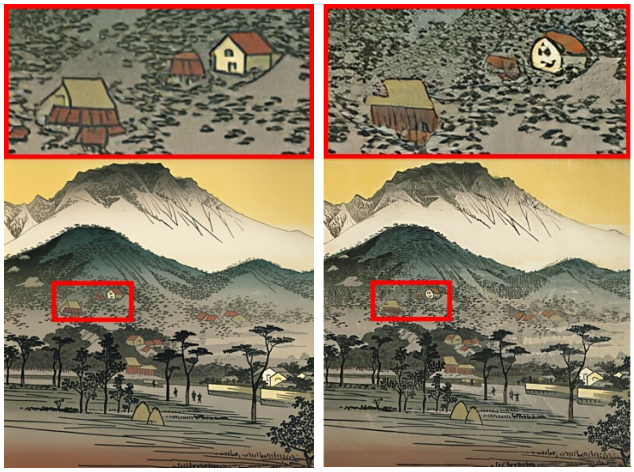
	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	4.36	3.95	0.266	0.453	70.3	3.98	3.69	3
Image B	3.11	3.08	0.272	0.524	70.9	4.00	3.28	2
Better	B	B	B	B	B	B	A	A

Image A Image B

	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	3.37	3.30	0.298	0.461	74.2	4.40	3.10	2.5
Image B	2.27	2.71	0.277	0.637	76.3	4.48	2.94	2
Better	B	B	A	B	B	B	A	A

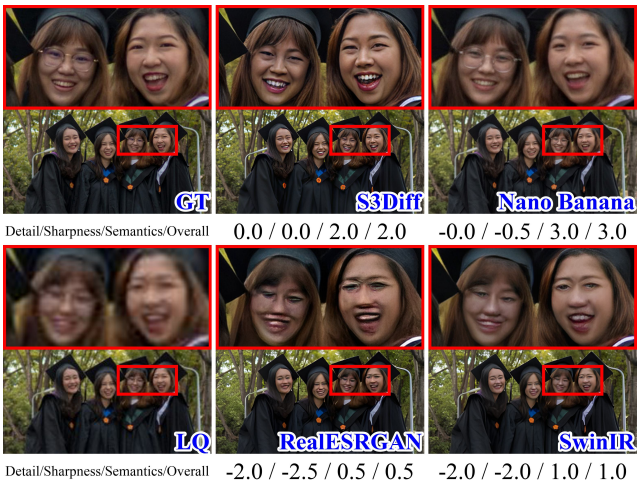
Figure 13. Qualitative analysis of image quality assessment methods. Zoom in for a better view.

	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	4.25	3.85	0.169	0.589	62.3	3.70	3.14	2.5
Image B	3.76	3.46	0.215	0.801	71.3	4.03	3.07	2
Better	B	B	B	B	B	B	A	A



	NIQE↓	PI↓	MANIQA	CLIP-IQA	MUSIQ	DeQA-Score	**Ours**	**Human**
Image A	4.53	3.74	0.204	0.762	68.9	3.37	3.41	3.5
Image B	3.11	3.04	0.346	0.833	73.3	3.53	3.27	2
Better	B	B	B	B	B	B	A	A

Figure 14. Qualitative analysis of image quality assessment methods. Zoom in for a better view.



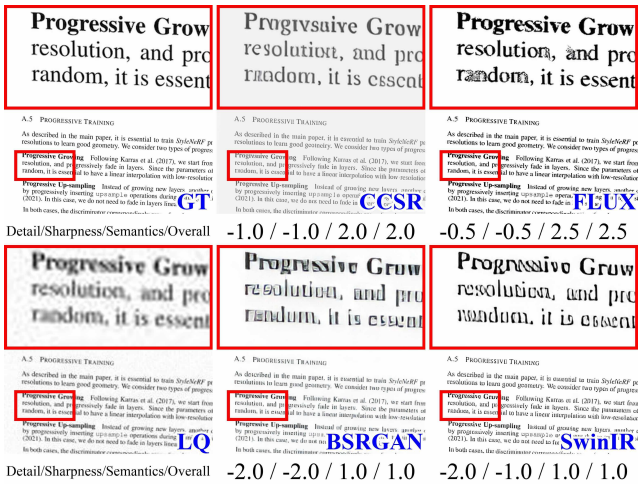


Figure 15. Failure cases of semantic errors. Zoom in for a better view.

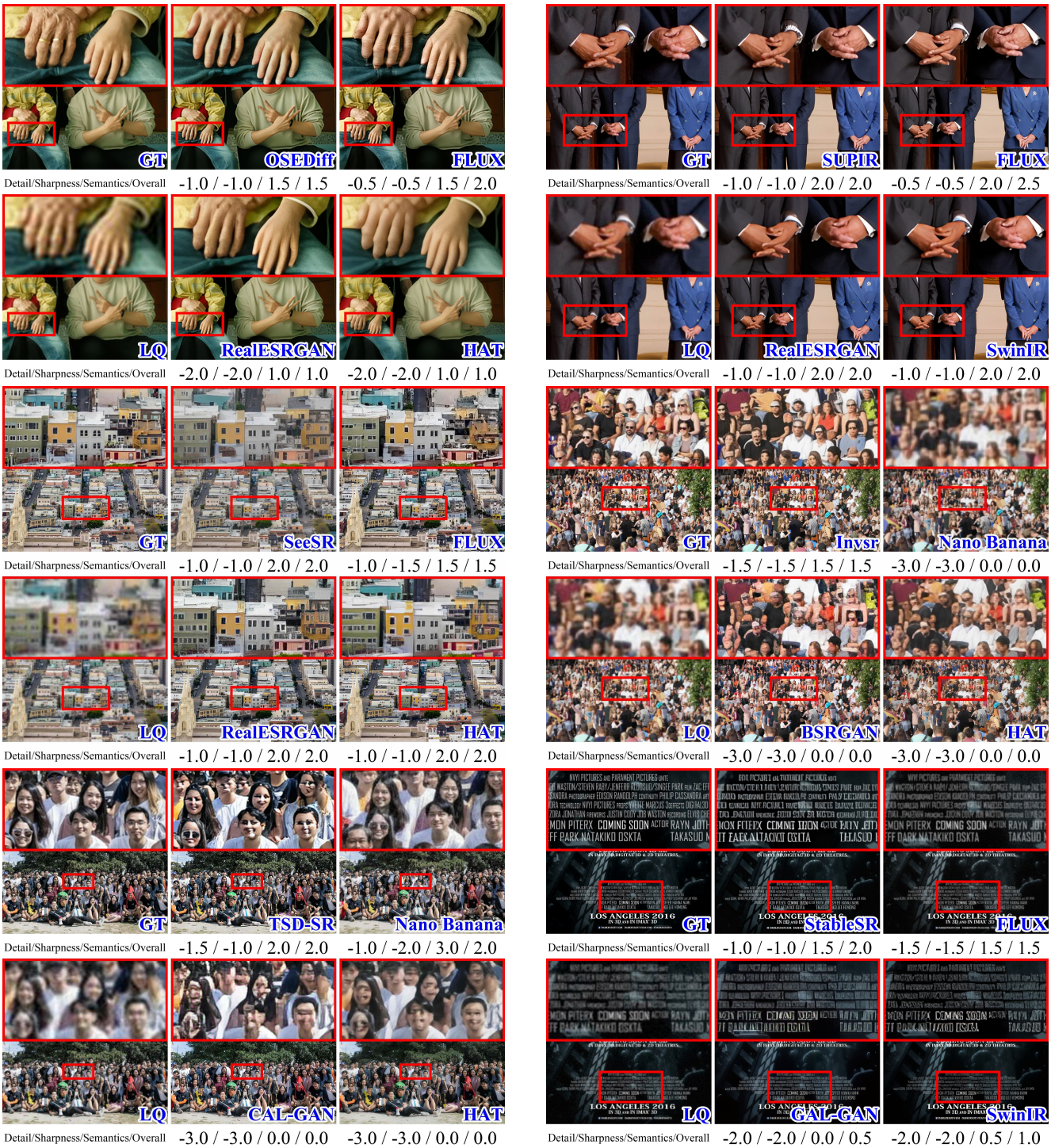


Figure 16. Failure cases of semantic errors. Zoom in for a better view.

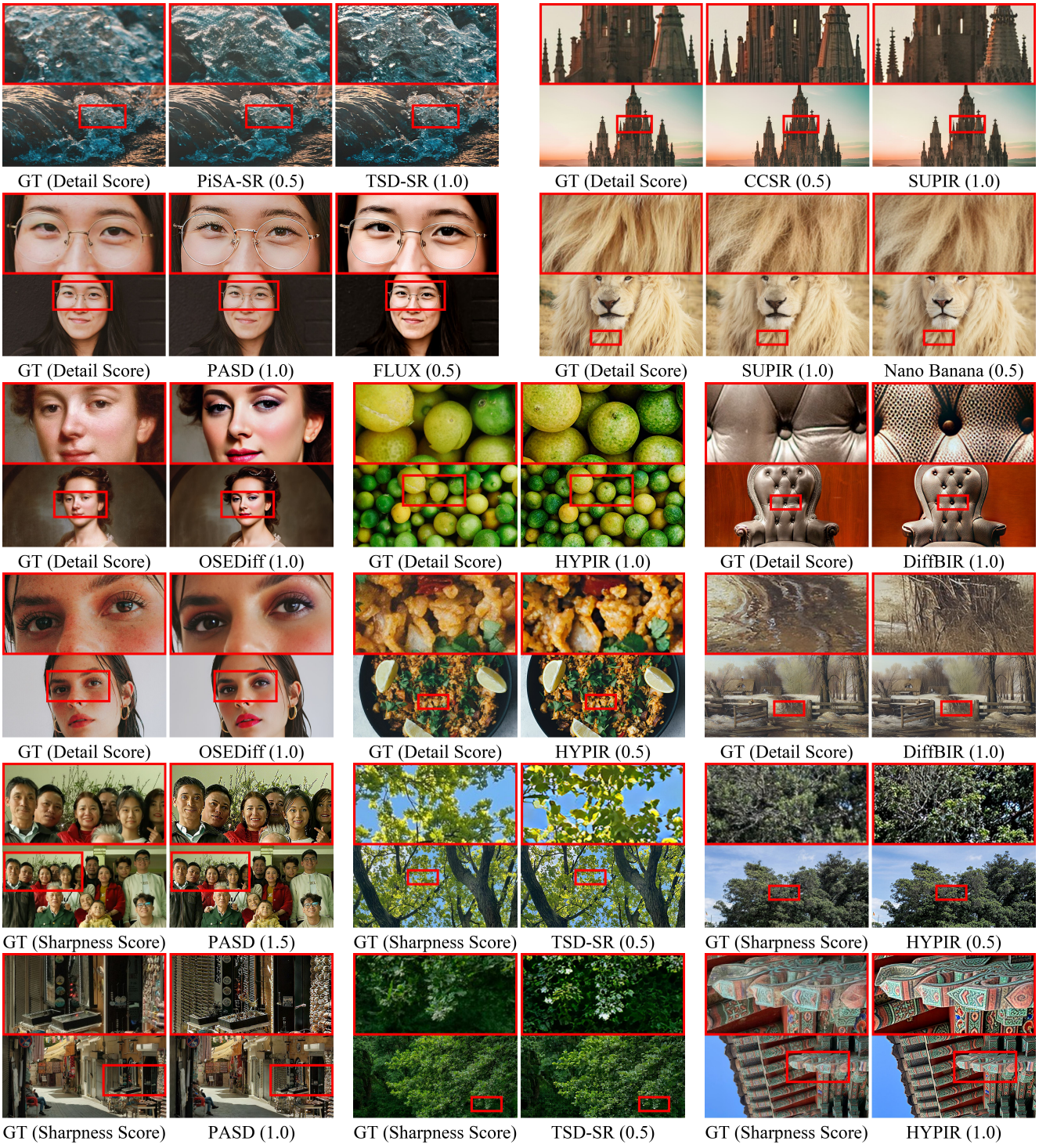


Figure 17. Failure cases of over-generation. Zoom in for a better view.

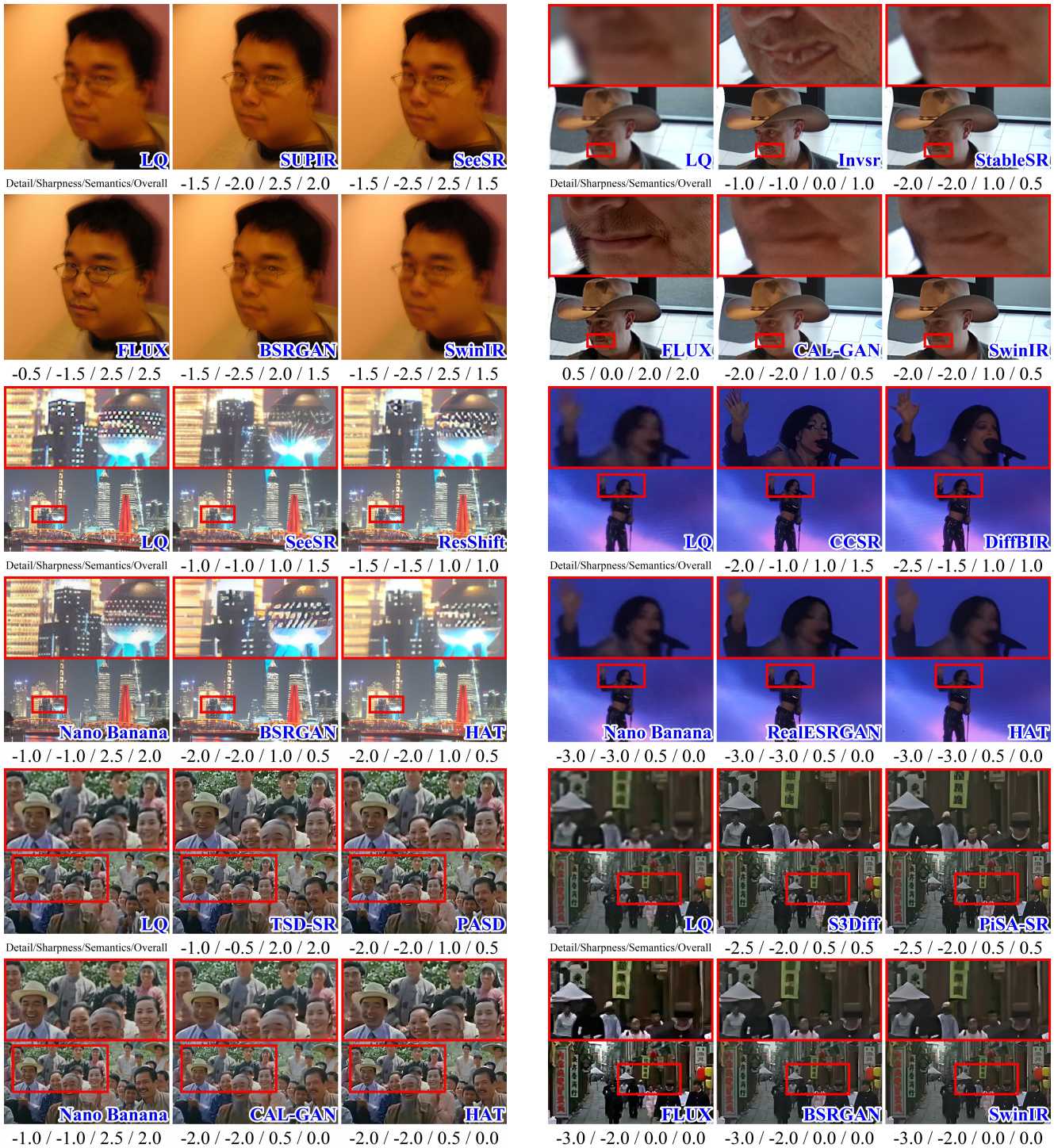


Figure 18. Failure cases of hard degradations. Zoom in for a better view.

Model	Crowd				Hands/Feet				Animal Fur			
	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑
HYPIR	2.40	1.30	1.50	2.40	3.05	0.50	0.80	3.15	2.94	0.50	0.67	3.11
SUPIR	1.80	1.70	1.80	1.70	2.55	1.25	1.10	2.75	2.89	1.00	0.78	3.11
PiSA-SR	1.90	1.60	1.60	1.70	2.40	1.15	1.15	2.60	3.06	0.72	0.56	3.22
SeeSR	1.70	1.60	1.80	1.40	2.30	1.40	1.35	2.45	2.50	1.17	1.17	2.83
OSEDiff	1.20	1.80	1.90	1.50	2.40	1.15	1.25	2.50	2.72	1.06	0.83	3.06
CCSR	0.90	1.80	2.00	1.40	2.25	1.30	1.60	2.35	2.83	0.72	0.61	3.00
DiffBIR	1.60	1.70	1.70	1.60	2.40	0.80	1.20	2.60	3.06	0.50	0.72	3.11
StableSR	1.30	1.80	1.90	1.20	1.70	1.75	1.70	2.20	2.28	1.67	1.39	2.61
PASD	1.70	1.40	1.50	1.90	2.25	1.10	1.30	2.50	2.78	0.89	0.94	2.94
Invsr	1.80	1.80	1.70	1.60	2.40	1.10	1.20	2.60	3.00	0.56	0.72	3.00
S3Diff	2.00	1.40	1.60	1.80	2.80	0.95	0.85	2.90	3.17	0.67	0.67	3.22
TSD-SR	1.80	1.70	1.80	1.40	2.30	1.25	1.25	2.25	2.67	0.89	0.83	2.78
ResShift	1.10	1.80	2.00	1.30	1.55	1.75	1.75	2.00	2.11	1.44	1.28	2.61
FLUX	1.40	1.60	1.60	1.50	1.95	1.45	1.25	2.30	1.67	1.94	1.67	2.22
Nano Banana	0.20	2.70	2.50	1.30	1.05	2.40	2.15	1.90	1.78	1.94	1.56	2.22
BSRGAN	1.00	2.00	2.00	0.90	1.45	1.60	1.85	1.95	1.78	1.67	1.72	2.44
CAL-GAN	0.80	2.20	2.10	1.20	1.35	1.80	1.90	1.90	1.83	1.67	1.28	2.44
RealESRGAN	1.30	2.00	2.20	1.40	1.40	1.70	2.05	1.85	2.11	1.50	1.50	2.56
HAT	0.80	2.10	2.20	0.90	1.60	1.85	1.80	1.95	2.28	1.72	1.50	2.72
SwinIR	0.90	1.40	1.80	1.00	1.50	2.00	2.00	2.05	2.28	1.39	1.39	2.67

Table 6. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Crowd, Crowd and Animal Fur. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	Complex Texture				Trees & leaves				Fabric Texture			
	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑
HYPIR	3.00	0.92	0.75	3.17	2.79	1.07	1.14	2.86	2.70	0.30	0.70	2.70
SUPIR	2.58	0.67	0.67	2.67	2.86	1.07	0.93	2.86	3.10	0.50	0.10	3.20
PiSA-SR	2.58	1.00	1.08	2.83	2.57	1.00	1.07	2.71	2.50	0.60	1.10	2.40
SeeSR	2.25	1.17	1.25	2.58	2.43	1.57	1.43	2.86	2.80	0.90	0.70	3.00
OSEDiff	2.67	1.08	1.17	2.75	2.36	1.21	1.14	2.57	2.60	0.60	0.60	2.60
CCSR	2.25	1.33	1.33	2.33	2.36	1.43	1.14	2.79	2.40	0.90	1.40	2.40
DiffBIR	2.75	0.67	0.92	2.92	2.50	1.43	1.21	2.64	2.60	0.80	0.80	3.00
StableSR	2.17	1.58	1.33	2.58	1.86	2.07	1.86	2.29	2.00	1.80	1.10	2.50
PASD	2.33	0.58	0.58	2.50	2.57	1.29	1.07	2.93	2.70	1.00	0.80	2.90
Invsr	2.00	0.75	1.17	2.08	2.64	0.93	1.07	2.71	2.10	0.50	1.30	2.20
S3Diff	2.83	0.83	0.83	2.92	3.00	0.86	1.07	3.07	2.70	0.50	1.00	2.70
TSD-SR	2.58	0.67	0.75	2.50	2.71	1.07	0.93	3.07	2.50	0.70	0.70	2.50
ResShift	1.67	1.33	1.42	2.08	1.50	1.50	1.50	2.29	1.60	1.40	1.50	2.20
FLUX	1.33	1.75	1.42	1.67	0.79	2.07	2.00	1.71	0.60	2.40	1.80	1.90
Nano Banana	0.83	2.25	2.00	2.17	1.21	2.14	2.00	1.71	1.50	1.70	1.50	2.60
BSRGAN	1.92	1.58	1.33	2.17	1.36	2.14	1.86	2.36	1.70	1.50	1.80	2.20
CAL-GAN	1.58	1.42	1.17	2.00	1.14	1.71	1.79	2.07	1.60	1.30	1.50	1.90
RealESRGAN	1.92	1.08	1.33	2.17	1.79	1.50	1.29	2.50	2.10	1.40	2.00	1.90
HAT	1.75	1.67	1.42	2.17	2.07	1.71	1.57	2.50	1.80	2.00	1.80	2.20
SwinIR	2.00	1.25	1.33	2.00	2.29	1.50	1.14	2.43	2.10	0.90	1.50	2.30

Table 7. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Complex Texture, Complex Texture and Fabric Texture. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	Leather Surface				Reflective Glass				Water Flow			
	Overall↑	Sharpness ↓	Detail ↓	Semantics↑	Overall↑	Sharpness ↓	Detail ↓	Semantics↑	Overall↑	Sharpness ↓	Detail ↓	Semantics↑
HYPIR	3.50	0.33	0.17	3.58	3.21	0.64	0.64	3.36	2.40	1.30	1.30	2.80
SUPIR	3.00	0.92	0.83	3.17	2.50	0.93	0.86	2.86	2.30	1.30	1.00	2.60
PiSA-SR	3.25	0.50	0.58	3.25	2.43	1.14	1.07	2.64	2.50	1.20	1.30	2.70
SeeSR	2.42	0.92	1.25	2.67	2.43	1.14	1.21	2.71	2.40	1.30	1.30	2.70
OSEDiff	3.42	0.50	0.58	3.50	2.50	1.14	1.14	2.86	2.40	1.50	1.50	2.70
CCSR	2.75	0.42	0.75	2.83	1.93	1.43	1.50	2.36	2.00	1.40	1.60	2.50
DiffBIR	3.33	0.25	0.42	3.58	2.57	0.79	0.86	2.64	2.60	1.10	1.20	2.90
StableSR	2.25	1.67	1.33	2.75	1.86	1.64	1.57	2.29	1.80	2.00	1.60	2.10
PASD	3.42	0.33	0.42	3.50	3.07	0.50	0.36	3.00	3.00	0.90	0.80	3.20
Invsr	2.92	0.75	0.58	2.83	2.57	0.93	1.07	2.64	2.50	1.00	1.20	2.50
S3Diff	3.00	0.42	0.50	3.17	2.64	0.79	0.79	2.86	2.50	1.00	1.00	2.50
TSD-SR	3.08	0.58	0.42	3.17	2.64	0.71	0.86	2.57	2.40	0.90	1.10	2.40
ResShift	2.25	1.50	1.42	2.75	1.79	1.57	1.50	2.07	2.00	1.50	1.10	2.60
FLUX	1.75	1.75	1.33	2.75	1.86	1.79	1.43	2.50	0.90	1.90	1.90	1.90
Nano Banana	1.75	1.08	0.58	2.25	1.21	2.43	1.93	2.00	1.80	1.80	1.60	2.40
BSRGAN	2.00	1.58	1.50	2.58	2.29	1.21	1.36	2.71	1.30	1.90	2.00	1.90
CAL-GAN	2.25	1.67	1.42	2.67	1.64	1.50	1.64	2.00	1.20	1.90	1.70	2.20
RealESRGAN	2.33	1.33	1.25	2.75	2.36	1.36	1.29	2.79	1.40	1.80	1.60	2.30
HAT	2.25	1.33	1.58	2.83	1.93	1.71	1.71	2.64	1.60	2.00	1.80	2.00
SwinIR	2.33	1.42	1.25	2.67	2.14	1.36	1.21	2.57	1.80	1.90	2.00	2.20

Table 8. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Leather Surface, Leather Surface and Water Flow. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	Vehicles				Street View				Aerial View			
	Overall↑	Sharpness ↓	Detail ↓	Semantics↑	Overall↑	Sharpness ↓	Detail ↓	Semantics↑	Overall↑	Sharpness ↓	Detail ↓	Semantics↑
HYPIR	2.80	1.00	1.00	2.80	2.64	1.07	1.36	2.93	2.50	1.30	1.40	2.80
SUPIR	2.70	1.00	1.00	2.60	2.43	1.07	1.07	2.43	2.40	1.60	1.00	2.40
PiSA-SR	2.70	1.00	1.60	3.00	2.36	1.36	1.29	2.43	1.40	1.60	1.70	2.10
SeeSR	2.20	1.00	1.40	2.70	1.93	1.64	1.71	2.07	1.70	1.70	1.40	2.40
OSEDiff	2.50	1.50	1.60	2.90	1.79	1.64	1.71	2.00	1.80	1.90	1.70	2.20
CCSR	2.50	1.30	1.20	2.60	2.21	1.43	1.36	2.71	1.70	2.20	1.80	1.90
DiffBIR	2.40	1.40	1.40	2.60	2.14	1.36	1.36	2.43	2.10	1.10	1.60	2.20
StableSR	2.00	1.70	1.60	2.50	1.64	2.00	1.71	2.14	1.80	1.90	1.50	2.50
PASD	3.10	0.60	0.70	2.90	2.21	1.00	0.93	2.29	2.90	0.70	0.60	3.00
Invsr	2.70	1.20	1.30	2.80	2.57	1.14	1.14	2.64	2.30	1.30	1.30	2.50
S3Diff	2.60	0.70	0.70	2.50	2.21	1.07	1.36	2.57	2.30	1.20	1.20	2.50
TSD-SR	2.90	1.10	1.40	2.80	2.43	0.64	1.07	2.50	2.30	1.00	1.00	2.30
ResShift	1.00	1.90	1.90	1.60	1.29	2.14	2.00	1.93	1.10	1.80	2.00	1.60
FLUX	2.40	1.00	1.00	2.60	1.50	1.86	1.57	2.14	1.30	2.10	1.60	2.10
Nano Banana	1.50	2.10	1.90	2.20	0.79	2.50	2.36	2.50	0.70	2.70	2.20	2.20
BSRGAN	1.10	1.80	1.90	1.40	1.57	2.21	2.07	2.07	1.40	1.70	1.60	2.10
CAL-GAN	1.50	1.50	1.70	1.50	1.50	1.79	1.64	2.14	1.10	1.70	1.70	1.70
RealESRGAN	1.70	1.60	1.90	2.10	2.07	1.64	1.64	2.36	0.80	1.70	1.90	1.50
HAT	1.20	2.00	2.10	1.80	1.29	1.86	1.86	1.50	1.00	1.90	2.00	1.80
SwinIR	1.60	1.60	1.90	2.00	1.57	1.43	1.71	2.21	1.70	1.50	1.80	2.20

Table 9. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Vehicles, Vehicles and Aerial View. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	Architecture				Food				Text			
	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑
HYPIR	2.88	0.75	0.88	3.06	2.93	0.57	0.57	3.00	2.56	0.78	0.94	2.56
SUPIR	2.62	0.44	0.81	2.69	3.00	0.71	0.86	3.00	2.50	0.94	0.61	2.39
PiSA-SR	2.44	1.06	1.19	2.69	3.07	1.07	0.79	3.14	2.33	0.94	0.94	2.61
SeeSR	2.38	1.12	1.25	2.69	2.64	0.79	0.86	2.79	1.78	1.00	0.89	1.89
OSEDiff	2.81	1.06	0.94	2.81	2.57	1.14	0.93	2.79	1.94	1.28	1.28	2.06
CCSR	2.38	0.94	1.06	2.50	2.93	0.93	0.86	3.00	2.22	0.94	1.22	2.33
DiffBIR	2.62	0.81	1.00	2.75	2.93	0.71	0.79	3.00	2.22	0.72	0.72	2.22
StableSR	1.62	1.88	1.69	2.38	2.43	1.29	1.14	2.86	1.61	1.39	1.22	2.11
PASD	2.75	0.81	0.62	3.00	3.14	0.21	0.43	2.93	2.11	0.78	0.83	2.33
Invsr	2.88	1.00	1.06	3.00	2.93	0.64	0.64	2.79	2.06	0.89	0.67	2.06
S3Diff	3.00	0.62	0.50	3.06	2.86	0.50	0.71	3.00	2.33	0.78	0.67	2.28
TSD-SR	2.88	1.06	1.06	2.94	3.07	0.36	0.43	3.07	2.33	0.94	0.83	2.33
ResShift	2.00	1.44	1.44	2.38	1.93	1.79	1.21	2.57	1.28	1.56	1.33	1.61
FLUX	1.75	1.56	1.12	2.19	2.14	1.43	1.36	2.43	1.50	1.22	1.06	1.89
Nano Banana	1.44	2.06	1.88	2.38	2.00	1.64	1.57	2.50	2.00	1.39	0.94	2.33
BSRGAN	1.62	1.75	1.56	2.12	2.07	1.71	1.43	3.07	1.33	2.00	1.39	1.83
CAL-GAN	1.69	1.25	1.62	2.12	1.36	1.86	1.86	1.86	1.22	1.78	1.33	1.50
RealESRGAN	2.12	1.25	1.50	2.44	1.79	1.64	1.43	2.36	1.72	1.61	1.28	1.83
HAT	1.75	1.62	1.56	2.12	1.64	1.93	1.79	2.36	1.61	1.44	1.44	1.83
SwinIR	1.81	1.44	1.56	2.25	2.21	1.36	1.21	2.50	1.56	1.61	1.28	1.89

Table 10. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Architecture, Architecture and Text. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	Hand-drawn				Print Media				Cartoon/Comic			
	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑
HYPIR	3.00	0.79	0.64	2.93	2.64	0.93	0.86	2.50	3.17	0.50	0.67	3.17
SUPIR	2.00	1.57	1.14	2.64	2.29	1.21	1.21	2.07	2.58	1.08	1.00	2.75
PiSA-SR	2.71	1.21	1.29	2.79	2.29	1.21	0.93	2.29	2.83	0.67	1.08	2.92
SeeSR	2.50	1.36	1.21	2.57	2.21	1.43	1.14	2.50	2.25	0.92	1.33	2.42
OSEDiff	2.21	1.64	1.57	2.21	1.93	1.71	1.71	1.86	3.00	0.83	1.00	2.92
CCSR	2.07	0.93	1.21	2.07	2.07	1.00	1.07	2.29	2.83	0.83	0.92	2.83
DiffBIR	2.29	1.29	1.43	2.36	2.14	0.93	0.86	2.14	2.00	0.75	1.33	2.08
StableSR	1.71	2.14	1.43	2.14	1.71	1.79	1.36	2.00	2.08	1.67	1.17	2.58
PASD	2.50	0.86	0.86	2.71	2.86	0.79	0.86	2.71	2.75	0.75	0.75	2.83
Invsr	2.36	1.07	1.21	2.43	2.00	1.71	1.43	2.14	2.67	0.67	0.83	2.67
S3Diff	2.71	0.93	0.93	2.64	2.07	1.43	1.29	2.21	2.92	0.50	0.75	3.25
TSD-SR	2.29	1.14	1.07	2.36	2.57	0.79	0.93	2.71	2.92	0.50	0.67	2.92
ResShift	1.43	1.93	1.79	1.71	1.43	1.79	1.57	1.64	1.58	1.42	1.42	2.00
FLUX	1.79	1.36	1.29	2.43	2.43	0.93	1.14	2.50	2.50	1.08	1.08	2.67
Nano Banana	1.43	1.86	1.50	2.00	1.43	2.14	1.71	1.93	1.58	2.00	1.58	2.33
BSRGAN	1.86	1.86	1.71	2.36	1.21	2.14	2.00	1.64	2.00	1.33	1.50	2.17
CAL-GAN	1.64	1.93	1.86	2.29	1.29	1.71	1.50	1.14	1.58	1.50	1.67	2.17
RealESRGAN	1.93	1.57	1.50	2.00	1.29	1.79	1.57	1.50	2.08	1.25	1.67	2.42
HAT	1.36	1.93	2.07	1.64	1.57	1.57	1.29	1.86	1.58	1.67	1.58	1.92
SwinIR	1.93	1.43	1.43	2.07	1.43	1.79	1.50	1.64	2.17	1.08	1.42	2.50

Table 11. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Hand-drawn, Hand-drawn and Cartoon/Comic. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	Compression				Defocus Blur				Digital Zoom			
	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑
HYPIR	2.76	0.63	0.78	3.00	2.95	0.95	0.77	3.27	2.59	0.98	1.00	2.98
SUPIR	2.26	1.13	1.20	2.54	2.09	1.55	1.41	2.45	2.39	1.28	1.13	2.96
PiSA-SR	2.52	1.02	1.02	2.85	2.77	0.86	0.91	3.18	2.35	1.26	1.09	2.78
SeeSR	2.17	1.28	1.43	2.70	2.18	1.23	1.27	2.55	2.09	1.57	1.65	2.65
OSEDiff	2.41	1.02	1.04	2.74	2.73	0.82	0.95	3.05	2.04	1.37	1.30	2.67
CCSR	2.24	1.04	1.22	2.50	2.36	0.91	1.05	2.64	2.15	1.17	1.28	2.61
DiffBIR	2.22	1.22	1.33	2.50	2.59	0.82	1.00	2.73	2.17	1.28	1.13	2.59
StableSR	2.11	1.26	1.17	2.63	1.77	1.68	1.41	2.68	2.33	1.33	1.11	2.80
PASD	2.37	1.17	1.09	2.74	1.95	1.45	1.32	2.64	2.28	1.33	1.20	2.74
Invsr	2.30	1.04	1.09	2.63	2.32	1.09	1.05	2.73	2.09	1.41	1.33	2.76
S3Diff	1.54	1.80	1.54	2.37	2.23	1.45	1.41	2.77	1.28	2.17	2.11	2.39
TSD-SR	2.33	1.04	1.11	2.76	2.18	1.23	1.36	2.73	2.63	0.96	0.89	2.83
ResShift	1.39	1.78	1.65	2.20	1.36	2.00	1.91	2.23	1.26	1.89	1.85	2.11
FLUX	1.91	1.63	1.50	2.57	2.27	1.32	1.27	2.73	2.26	1.28	1.20	2.83
Nano Banana	2.00	1.52	1.41	2.43	1.14	2.41	2.32	1.91	1.52	1.85	1.52	2.59
BSRGAN	1.15	1.87	1.74	2.00	1.59	1.64	1.41	2.32	1.20	1.96	1.80	2.26
CAL-GAN	1.15	1.87	1.74	1.91	1.27	2.05	1.91	2.09	0.93	2.28	2.11	2.04
RealESRGAN	1.48	1.70	1.74	2.24	1.73	1.82	1.64	2.41	1.39	1.98	1.85	2.24
HAT	1.52	1.80	1.74	1.98	1.55	2.09	1.95	2.41	1.33	1.98	1.65	2.17
SwinIR	1.57	1.57	1.48	2.28	1.59	1.73	1.64	2.50	1.57	1.76	1.67	2.52

Table 12. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Compression, Compression and Digital Zoom. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	ISP Noise				Low light				Motion Blur			
	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑
HYPIR	3.12	0.54	0.71	3.25	2.39	1.13	1.05	2.61	2.43	1.50	1.64	2.79
SUPIR	2.08	1.71	1.42	2.71	1.66	1.97	1.66	2.39	0.86	2.50	2.43	2.50
PiSA-SR	2.75	0.96	0.88	2.88	2.45	1.21	1.08	2.66	1.43	2.14	1.57	2.21
SeeSR	1.92	1.33	1.38	2.50	1.61	1.71	1.63	2.05	1.57	2.07	2.07	2.21
OSEDiff	2.67	0.71	0.92	2.83	1.97	1.34	1.21	2.26	2.00	1.86	1.64	2.71
CCSR	2.29	1.29	1.08	2.58	1.74	1.42	1.24	2.21	1.00	2.21	2.14	2.43
DiffBIR	2.67	1.00	0.92	2.96	2.05	1.42	1.37	2.45	2.14	1.57	1.50	2.79
StableSR	2.54	0.96	0.83	3.04	1.71	1.76	1.84	2.37	0.71	2.43	2.57	2.36
PASD	2.33	1.17	1.08	2.79	1.97	1.47	1.42	2.26	1.57	2.21	1.71	2.71
Invsr	2.33	1.12	1.21	2.50	1.71	1.68	1.63	2.34	1.50	2.07	2.00	2.64
S3Diff	1.46	1.79	1.50	2.33	1.24	2.16	1.76	2.05	1.14	2.29	1.86	2.43
TSD-SR	2.71	0.79	0.79	2.96	2.18	1.39	1.45	2.50	1.50	2.00	1.79	2.50
ResShift	1.46	2.00	1.88	2.33	1.13	2.03	1.92	2.05	0.64	2.64	2.50	1.86
FLUX	2.75	0.79	0.88	3.17	1.79	1.55	1.53	2.68	2.14	1.50	1.43	2.86
Nano Banana	1.62	1.50	1.38	2.92	1.66	1.84	1.55	2.24	0.57	2.64	2.29	2.21
BSRGAN	1.38	1.83	1.92	1.88	1.03	2.16	1.97	1.45	0.64	2.71	2.64	1.64
CAL-GAN	1.17	1.96	1.83	1.88	0.82	2.26	2.16	1.50	0.43	2.79	2.50	1.79
RealESRGAN	1.54	1.71	1.79	2.17	1.08	2.13	2.03	1.87	0.57	2.50	2.50	1.50
HAT	1.38	1.92	1.83	2.08	0.89	2.13	1.97	1.68	0.00	3.00	2.86	1.79
SwinIR	1.29	1.75	1.83	2.25	1.39	2.16	1.82	1.82	0.21	2.79	2.64	2.36

Table 13. The result of average overall, sharpness, detail, and semantics scores for all restoration models in ISP Noise, ISP Noise and Motion Blur. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	Old Film				Old Photo (B/W)				Old Photo (Color)			
	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑	Overall↑	Sharpness |↓	Detail |↓	Semantics↑
HYPIR	2.68	0.96	0.99	2.85	2.79	0.88	0.95	3.03	2.90	0.77	0.77	3.00
SUPIR	1.69	1.90	1.65	2.59	2.12	1.38	1.19	2.60	2.16	1.47	1.37	2.66
PiSA-SR	2.46	1.25	1.32	2.79	2.67	0.86	1.02	2.88	2.61	0.84	0.89	2.91
SeeSR	2.09	1.60	1.59	2.51	2.57	1.07	1.19	2.81	2.33	1.26	1.17	2.74
OSEDiff	2.38	1.22	1.22	2.68	2.64	0.90	0.95	2.83	2.59	1.03	1.19	2.83
CCSR	2.07	1.38	1.46	2.35	2.41	1.07	1.10	2.83	2.33	1.14	1.21	2.66
DiffBIR	2.35	1.19	1.25	2.68	2.78	0.97	0.98	2.86	2.51	1.03	1.13	2.83
StableSR	1.63	1.90	1.74	2.47	2.50	1.22	1.21	2.76	1.89	1.53	1.50	2.63
PASD	1.81	1.59	1.47	2.54	2.64	0.98	0.93	2.76	2.49	1.03	1.01	2.71
Invsr	2.10	1.22	1.22	2.50	2.52	1.16	1.16	2.83	2.31	0.99	1.06	2.69
S3Diff	2.32	1.32	1.25	2.82	2.03	1.57	1.31	2.47	2.06	1.47	1.37	2.69
TSD-SR	2.43	1.19	1.24	2.60	2.76	0.86	0.98	2.91	2.49	1.10	1.10	2.79
ResShift	1.46	2.03	1.87	2.37	1.60	1.64	1.53	2.14	1.46	1.80	1.86	2.20
FLUX	1.97	1.54	1.41	2.44	2.21	0.81	0.78	2.83	2.19	1.23	1.13	2.69
Nano Banana	1.00	2.21	2.09	2.24	1.81	1.69	1.45	2.52	1.29	2.13	1.97	2.36
BSRGAN	1.46	2.00	1.94	2.22	1.47	1.71	1.97	1.86	1.29	1.83	1.89	2.06
CAL-GAN	1.31	2.03	1.99	2.15	1.16	1.74	1.81	1.76	1.20	2.03	1.89	2.01
RealESRGAN	1.38	2.10	1.94	2.35	1.52	1.62	1.69	2.03	1.39	1.87	1.90	2.11
HAT	1.53	2.12	1.85	2.29	1.57	1.83	1.81	2.10	1.14	1.96	1.80	2.07
SwinIR	1.38	2.01	1.93	2.21	1.88	1.55	1.55	2.26	1.51	1.80	1.83	2.27

Table 14. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Old Film, Old Film and Old Photo (Color). Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.

Model	Surveillance			
	Overall↑	Sharpness |↓	Detail |↓	Semantics↑
HYPIR	2.25	1.43	1.36	2.50
SUPIR	2.00	1.57	1.39	2.29
PiSA-SR	2.00	1.64	1.54	2.46
SeeSR	2.04	1.64	1.68	2.32
OSEDiff	2.00	1.57	1.57	2.29
CCSR	1.61	1.82	1.64	1.96
DiffBIR	1.96	1.61	1.68	2.25
StableSR	1.25	2.18	2.11	1.89
PASD	2.25	1.18	1.29	2.68
Invsr	1.57	1.64	1.64	1.79
S3Diff	2.21	1.43	1.61	2.39
TSD-SR	1.79	1.61	1.61	2.11
ResShift	1.21	2.43	2.21	1.71
FLUX	1.25	2.21	1.93	1.89
Nano Banana	0.89	2.25	2.14	1.46
BSRGAN	1.43	2.21	2.14	1.96
CAL-GAN	1.00	2.18	2.11	1.61
RealESRGAN	1.39	2.18	2.07	2.21
HAT	1.25	2.36	2.11	1.89
SwinIR	1.21	2.18	2.11	2.00

Table 15. The result of average overall, sharpness, detail, and semantics scores for all restoration models in Surveillance. Note that the Detail and Sharpness scores are calculated by taking the absolute value first, and then averaging the scores.