GGT-100K: GENERATIVE GROUND TRUTH FOR GENERALIZABLE REAL-WORLD IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

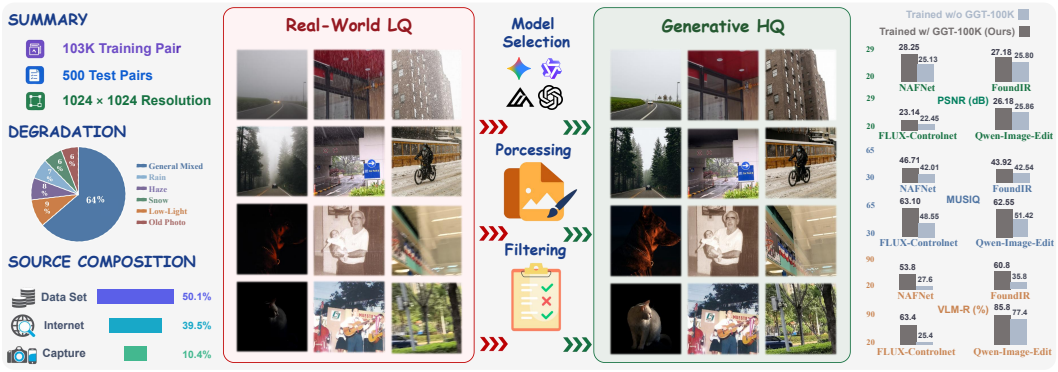


Figure 1: We construct GGT-100K through a carefully designed pipeline, aiming to improve the generalization performance of a wide range of image restoration models.

ABSTRACT

Real-world image restoration (IR) is bottlenecked by the scarcity of high-quality paired training data. Synthetic datasets are abundant but often fail to model real-world degradations, while real-world paired datasets are expensive and difficult to capture. As a result, IR models trained on these datasets show limited generalization in real-world scenarios. In this work, we propose Generative Ground Truth (GGT) by using generative multimodal foundation models (MFMs) to produce high-quality (HQ) targets from real-world low-quality (LQ) images. We first conduct a systematic evaluation of nine state-of-the-art MFMs, including Nano-Banana-2 and GPT-Image-2, on images of various scenes and degradation types. The results demonstrate that Nano-Banana-2 with VLM-based adaptive prompting shows the highest capability to synthesize perceptually realistic and content-faithful HQ targets, which can serve as the GGT for the LQ input. We then employ Nano-Banana-2 to build a GGT synthesis pipeline, which involves multi-stage quality control to ensure data reliability, and construct GGT-100K, an LQ-HQ paired dataset comprising 103,707 training pairs and covering diverse scenes and complex real-world degradations. A test set of 500 image pairs is also established. Extensive experiments show that GGT-100K consistently improves the real-world generalization of a wide range of IR models, with particularly strong benefits for finetuning generative models for IR tasks. Our results suggest that MFMs can serve as practical tools for restoration-oriented data generation, and GGT-100K is a useful resource to expand the generalization boundaries of real-world IR models. Our dataset will be publicly available.

1 INTRODUCTION

Classical image restoration (IR) (Zhang et al., 2017; Dong et al., 2014; Li et al., 2019; Cho et al., 2021) typically focuses on a specific predefined degradation, such as denoising, deblurring, super-resolution or dehazing. In contrast, real-world IR (Cai et al., 2019; Li et al., 2025) aims to recover images affected by complex, mixed, and often unknown degradations in practical environments. Despite substantial progress in IR architectures–from CNN and Transformer backbones (Liang et al., 2021;

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



Figure 2: Qualitative comparison of FoundIR (Li et al., 2025) and Qwen-Image-Edit (Wu et al., 2025) trained/finetuned on different datasets. FoundIR shows limited generalization to real rain and haze, even after finetuned on existing restoration datasets. Qwen-Image-Edit produces sharper details with its strong generative prior but may introduce hallucinations or color shifts, while finetuning on existing datasets weakens its detail generation ability. GGT-100K significantly improves the generalization capability of them to real-world degradations, enhancing visual details while preserving scene fidelity.

Zamir et al., 2021; Chen et al., 2022; 2024b) to all-in-one models (Li et al., 2022a; Potlapalli et al., 2023; Zamfir et al., 2024; Luo et al., 2024; Li et al., 2025) and recent generative frameworks (Lin et al., 2023; Yu et al., 2024; Wu et al., 2024)–robust generalization to real-world scenarios remains far from solved.

The lack of paired training data is one of the key bottlenecks for generalizable real-world IR models. Existing paired data construction mainly follows two routes: synthetic generation and real-world acquisition. Synthetic data are scalable, but simulated degradations often fail to model the complexity of real-world image formation process, leading to a substantial domain gap (Zhang et al., 2021a; Wang et al., 2021; Liu et al., 2022; 2023; Kong et al., 2024). Physically collected real-world image pairs provide more realistic supervision. Dedicated capture pipelines such as RealSR (Cai et al., 2019), SIDD (Abdelhamed et al., 2018), RealBlur (Rim et al., 2020) and FoundIR (Li et al., 2025) obtain well-aligned LQ-HQ pairs through controlled imaging or repeated acquisition, but such capture is expensive and the scenes and degradation types they can cover remain inherently restricted. More fundamentally, this route is only available when the pair is captured *on purpose*. For *the vast amount of real-world LQ images that have been captured*, the corresponding HQ reference images are nearly impossible to obtain because the instant scene and the associated degradation cannot be re-staged. As shown in Fig. 2, the FoundIR model (Li et al., 2025) trained on synthetic data or its officially released real-world dataset still produces noticeable artifacts for low-quality (LQ) input images. This persistent data bottleneck motivates us to develop a scalable paradigm to construct high-quality (HQ) restoration targets from diverse real-world LQ inputs.

Recent generative multimodal foundation models (MFMs) (Team et al., 2023; OpenAI, 2025a; Wu et al., 2025) offer a promising opportunity to achieve our goal. Modern MFMs can take image and instructions as inputs to produce the desired output, suggesting that they may generate restoration-oriented HQ targets for real-world LQ images. However, this task is nontrivial: current MFMs may distort image structures, hallucinate details, or behave inconsistently across images and prompts. This raises a key question: *can MFMs generate HQ targets with sufficient fidelity and stability for supervised real-world IR model training*?

In this work, we conduct a systematic investigation to answer this question. We first evaluate nine modern MFMs, including Nano-Banana-2 (Team et al., 2023) and GPT-Image-2 (OpenAI, 2025b), by prompting them to generate HQ counterparts of the input LQ images with various image scenes and

degradation types. Fixed and VLM-based adaptive prompting strategies are employed. We evaluate these MFMs in terms of image content fidelity, perceptual quality, VLM-based evaluation, and human preference, and find that Nano-Banana-2 with adaptive prompting can generate the most reliable HQ target image, which can serve as the Generative Ground Truth (GGT) to supervise the real-world IR model training.

Based on the above finding, we employ Nano-Banana-2 to construct GGT-100K, an LQ-HQ paired real-world IR dataset with 103,707 training pairs and a carefully established test set of 500 pairs. As shown in Fig. 3, we collect real-world LQ images from existing datasets, Internet sources, and our own captures, covering general mixed degradations, rain, haze, snow, low-light conditions, and old photos. These categories are not isolated single-degradation settings: images captured under rain are also coupled with blur, noise, and compression artifacts. We then generate restoration targets using Nano-Banana-2 with a multi-stage quality control process. To validate GGT-100K, we retrain representative CNN- (Zamir et al., 2021; Chen et al., 2022), transformer- (Liang et al., 2021; Chen et al., 2024b), all-in-one (Potlapalli et al., 2023; Zamfir et al., 2024; Luo et al., 2024; Li et al., 2025), T2I- (Labs, 2024) and TI2I-adapted (Wu et al., 2025) generative restoration models with and without it, and observe consistent generalization gains across multiple evaluation datasets and model families.

Our contributions are threefold. First, we propose GGT, a scalable paradigm for constructing real-world paired IR training data with MFMs, and systematically evaluate nine MFMs with multiple prompting strategies. Second, we build GGT-100K, a paired dataset with 103K high-quality LQ-HQ pairs for real-world IR training. Finally, we show that GGT-100K consistently improves the generalization of different restoration model families, especially modern generative models with strong priors and learning capacity. An overview is illustrated in Fig. 1.

2 Related Work

Real-world Image Restoration. Real-world IR aims to recover LQ images captured in complex real-world environments to an expected HQ one. Restoration architectures have evolved from CNN- (Zamir et al., 2021; Chen et al., 2022) and transformer-based backbones (Liang et al., 2021; Zamir et al., 2022) to all-in-one frameworks that handle multiple degradation types within a unified model (Li et al., 2022a; Potlapalli et al., 2023) and, more recently, generative approaches (Ho et al., 2020; Lin et al., 2023; Wu et al., 2024; Wei et al., 2025), which show stronger restoration capability and generalization but may sacrifice fidelity and introduce hallucinated details. Lightweight CNN backbones remain attractive for their efficiency, while heavier transformer-based models often achieve stronger restoration performance (Zamir et al., 2022; Liang et al., 2021). Despite this rapid progress in model design, robust generalization across diverse real-world degradations remains difficult, mainly due to the data bottleneck.

Most existing real-world IR methods rely on synthetic training data, where LQ inputs are generated from HQ images using hand-crafted degradation models (Zhang et al., 2021a; Wang et al., 2021). Such synthetic data are easy to scale but hard to capture the complexity of real-world image degradation, leading to a substantial domain gap (Liu et al., 2022; 2023). A few real-world paired datasets have been built through settings such as multiple acquisition, controlled imaging, etc., (Cai et al., 2019; Wei et al., 2020; Plotz & Roth, 2017; Li et al., 2019). These datasets provide more realistic supervision, but they are expensive and difficult to scale. Recent efforts such as FoundIR (Li et al., 2025) have advanced real-world data collection in this direction, but relying on real capture alone encounters practical difficulties since scenarios such as rain and haze still need synthetic data and the captured data are constrained in scene and degradation diversity. These pipelines can only produce pairs that are captured on purpose and cannot supply GT for the far larger pool of real LQ images that have already been recorded. As a result, existing real-world paired datasets remain limited in scalability, scene diversity, and degradation coverage.

Multimodal Foundation Models. Recent generative MFMs have rapidly advanced visual understanding, instruction-based editing, and image generation (OpenAI, 2025a; Team et al., 2023; Wu et al., 2025), making them a promising tool for generating HQ targets from LQ inputs. However, real-world IR poses stricter requirements than general editing, demanding both perceptual quality improvement and faithful content preservation, and current MFMs may still hallucinate details and behave inconsistently across prompts and degradation types. Their effectiveness in IR also remains insufficiently studied: recent works (Sun et al., 2026) evaluate MFMs only on limited degradation categories, while RealRestorer (Yang et al., 2026) relies on simple fixed prompts without studying

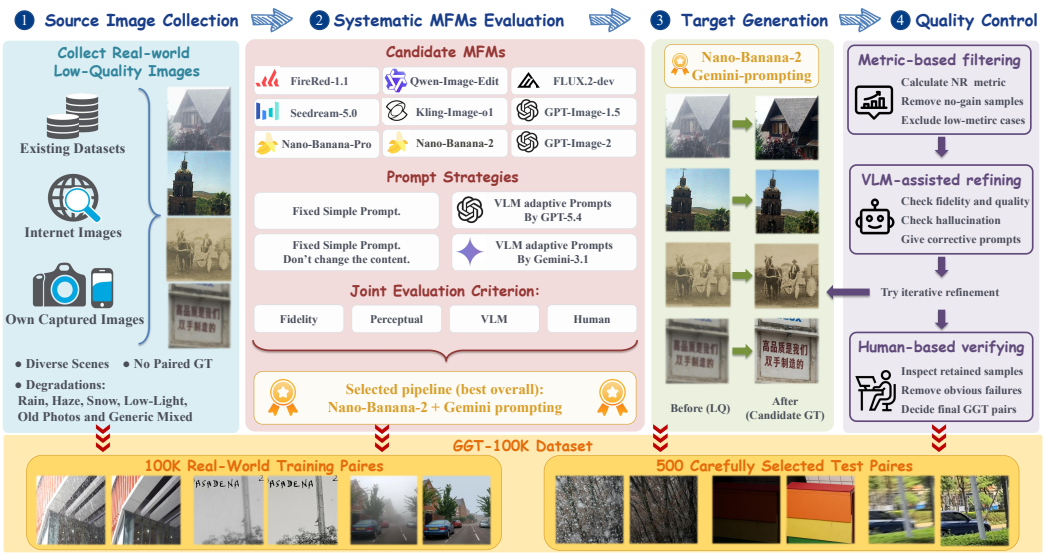


Figure 3: Overview of the GGT-100K construction pipeline. We collect diverse real-world LQ images, evaluate MFMs for HQ target generation, and apply multi-stage quality control to build the dataset.

fidelity preservation. More importantly, no complete pipeline has yet been established for using MFMs to build IR training data, covering model selection, prompting design and data screening. We provide such a systematic evaluation and pipeline.

3 GGT-100K: Dataset Construction

This section presents our pipeline for constructing real-world LQ-HQ paired data using MFMs. As illustrated in Fig. 3, we begin by collecting real-world LQ images from various sources. Given these inputs, we evaluate candidate MFMs and prompting strategies to identify the generation setting that best balances perceptual quality and fidelity preservation. We then use the selected MFM to generate candidate HQ targets at scale, followed by a multi-stage screening process to retain samples with both strong perceptual quality and high content fidelity, resulting in our GGT-100K.

3.1 Source Image Collection

To expand the generalization boundary of real-world IR models, we collect degraded images that lack HQ references and are not covered by existing paired datasets, from three main sources. The first is restoration datasets without ground-truth references (Li et al., 2019; Zhang et al., 2021b; Chen et al., 2018) and broader vision datasets containing low-quality or bad-weather images (Deng et al., 2009; Loh & Chan, 2019; Yang et al., 2020; Sakaridis et al., 2021). The second is Internet images collected through web crawling (Unsplash, 2025; Pexels, 2025; Pixabay, 2025; Flickr, 2025), with proper usage rights ensured. The third is our own captures with different cameras and mobile phones, covering blur, noise, low-light and related practical artifacts. All candidate images are normalized to 1024×1024. More details are in **Appendix A**.

3.2 Systematic MFMs Evaluation

Candidate MFMs and Prompts. We evaluate models and prompts jointly because they are tightly coupled. For models, we evaluate 3 open-source MFMs (FireRed-1.1 (Team et al., 2026), Qwen-Image-Edit-2511 (Wu et al., 2025) and FLUX.2-dev (Labs, 2025)) and 6 closed-source MFMs (Kling-Image-O1 (Team & MiraclePlus, 2025), Seedream-5.0 (ByteDance, 2025), GPT-Image-1.5 (OpenAI, 2025a), GPT-Image-2 (OpenAI, 2025b), Nano-Banana-Pro (Team et al., 2023) and Nano-Banana-2 (Team et al., 2023)). For prompts, we consider both fixed and adaptive ones. Fixed prompts share a restoration instruction for each input category, while a prompt variant (denoted as "Fix-NC" in Tab. 1) explicitly requires the image content to remain unchanged. For adaptive prompts, we first use a VLM (GPT-5.4-Pro (OpenAI, 2026) or Gemini-3.1-Pro (Team et al., 2023)) to analyze

Table 1: Comparison of different MFMs and prompting strategies, including fidelity metrics, perceptual metrics, VLM-based success rate (VLM-R), average score (Avg.), and human preference. The **best**, **second-best**, and **third-best** results are highlighted in red, blue, and green, respectively.

Model	Prompt	Fidelity results of DIV2K-Val				Perceptual results of 200 Real LQ images					VLM-R ↑	Avg.↑	Human ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓			
FireRed-1.1 (Open-source)(Team et al., 2026)	Fix	23.9340	0.7520	0.2165	0.1089	4.8022	64.6538	0.5928	0.5013	-0.9917	57.0%	0.7224	6.0%
	Fix-NC	23.7276	0.7493	0.2189	0.1096	4.7554	65.0938	0.5936	0.5041	**-0.9922**	58.0%	0.7297	
	GPT	19.9648	0.6870	0.2730	0.1272	**4.0810**	68.2131	0.6058	0.5452	**-0.9957**	23.0%	0.4787	
	Gemini	21.2432	0.7233	0.2603	0.1237	4.5822	62.8360	0.5798	0.4802	-0.9582	52.0%	0.6322	
Qwen-Image-Edit (Open-source) (Wu et al., 2025)	Fix	**27.5471**	**0.8170**	0.1710	0.0876	5.8545	51.7165	0.5280	0.3966	-0.9278	57.0%	0.6602	5.5%
	Fix-NC	**27.2219**	**0.8101**	0.1761	0.0906	6.0303	51.4212	0.5263	0.3977	-0.9285	56.5%	0.6433	
	GPT	26.1490	0.7930	0.1830	0.0925	5.5224	54.9141	0.5400	0.4120	-0.9250	58.5%	0.6755	
	Gemini	26.1282	0.7816	0.2082	0.1056	5.9969	51.0975	0.5172	0.3861	-0.9171	59.5%	0.6245	
FLUX.2-dev (Open-source) (Labs, 2025)	Fix	19.9945	0.7314	0.2251	0.1324	5.4224	57.4157	0.5617	0.4777	-0.8987	58.0%	0.6058	7.0%
	Fix-NC	20.2668	0.7371	0.2173	0.1288	5.5124	56.4839	0.5574	0.4700	-0.8894	57.5%	0.5994	
	GPT	21.3079	0.7378	0.2029	0.1258	5.0138	60.5277	0.5867	0.4971	-0.9105	68.0%	0.7285	
	Gemini	21.1881	0.7054	0.2351	0.1384	5.4573	55.6259	0.5605	0.4491	-0.8947	68.5%	0.6597	
Kling-Image-o1 (Closed-source) (Team & MiraclePlus, 2025)	Fix	24.5655	0.7422	0.2239	0.1115	5.5488	44.0313	0.5097	0.3233	-0.7539	55.5%	0.5187	5.5%
	Fix-NC	24.3646	0.7334	0.2177	0.1104	5.6452	42.2917	0.5018	0.3114	-0.7367	62.0%	0.5451	
	GPT	25.9992	0.7765	0.1824	0.1026	5.9970	39.3803	0.4810	0.2928	-0.7291	64.5%	0.5626	
	Gemini	25.1805	0.7665	0.1880	0.1053	5.8738	39.4746	0.4876	0.2956	-0.7172	63.5%	0.5501	
Seedream-5.0 (Closed-source) (ByteDance, 2025)	Fix	22.4237	0.7361	0.1889	0.1086	5.1858	54.5541	0.5339	0.4477	-0.8574	69.0%	0.6977	7.0%
	Fix-NC	22.7089	0.7362	0.1894	0.1089	5.2306	52.9141	0.5278	0.4344	-0.8469	64.5%	0.6568	
	GPT	23.3330	0.7515	0.1909	0.1088	5.0132	54.0857	0.5370	0.4597	-0.8566	61.0%	0.6607	
	Gemini	23.7293	0.7477	0.1856	0.1103	5.3304	53.0703	0.5231	0.4380	-0.8510	68.0%	0.6855	
GPT-Image-1.5 (Closed-source) (OpenAI, 2025a)	Fix	14.5390	0.3559	0.4902	0.1849	4.4747	**69.6380**	0.6226	**0.5842**	-0.9675	31.0%	0.3412	0%
	Fix-NC	14.6410	0.3567	0.4854	0.1822	4.6371	67.6234	0.6114	0.5549	-0.9427	41.0%	0.3860	
	GPT	15.0862	0.3665	0.4706	0.1801	4.8494	66.1326	0.6083	0.5449	-0.9112	39.5%	0.3646	
	Gemini	14.8436	0.3560	0.4719	0.1814	4.7243	66.4074	0.6151	0.5519	-0.9216	34.0%	0.3344	
GPT-Image-2 (Closed-source) (OpenAI, 2025b)	Fix	16.7439	0.4410	0.3374	0.1381	4.4268	**70.8689**	**0.6651**	**0.6166**	**-0.9984**	57.0%	0.6479	**21.5%**
	Fix-NC	18.8703	0.5157	0.2503	0.1149	4.5807	**68.3071**	**0.6503**	**0.5741**	-0.9672	65.0%	0.7336	
	GPT	17.8234	0.4830	0.2977	0.1143	4.9140	63.9640	0.6275	0.5020	-0.9304	**73.0%**	0.7115	
	Gemini	18.2098	0.4995	0.2722	0.1137	4.7421	67.1697	**0.6429**	0.5558	-0.9532	67.5%	0.7213	
Nano-Banana-Pro (Closed-source) (Team et al., 2023)	Fix	25.5366	0.7671	0.1588	0.0845	4.8273	57.1012	0.5739	0.4373	-0.8739	64.5%	0.7522	**15.0%**
	Fix-NC	26.5977	0.7865	0.1375	0.0780	5.0341	54.0572	0.5573	0.4063	-0.8473	68.5%	0.7667	
	GPT	27.0899	0.7917	0.1327	0.0753	5.3172	49.5661	0.5361	0.3681	-0.8234	**72.0%**	0.7569	
	Gemini	26.8784	0.7816	0.1429	0.0786	5.2612	51.0284	0.5440	0.3793	-0.8324	**70.0%**	0.7479	
Nano-Banana-2 (Closed-source) (Team et al., 2023)	Fix	25.6879	0.7758	0.1557	0.0856	**4.2132**	64.4820	0.6185	0.5189	-0.9393	52.0%	0.7564	**32.5%**
	Fix-NC	26.1795	0.7824	**0.1418**	**0.0814**	**4.4061**	62.7827	0.6089	0.4958	-0.9186	63.0%	**0.8171**	
	GPT	26.7364	0.7923	**0.1300**	**0.0757**	4.8221	56.6527	0.5765	0.4272	-0.8825	69.0%	**0.8078**	
	Gemini	**27.1701**	**0.7949**	**0.1280**	**0.0738**	4.6881	59.2831	0.5872	0.4536	-0.8980	**70.0%**	**0.8427**	

the input image and then generate an image-specific instruction based on its content and degradation. Detailed prompt designs are provided in **Appendix B**.

Evaluation Criterion. We consider four complementary aspects: *fidelity preservation*, *perceptual quality*, *VLM-based assessment*, and *human preference*. For fidelity, we use 100 DIV2K validation images (Agustsson & Timofte, 2017), degrade them with the Real-ESRGAN pipeline (Wang et al., 2021), and compare generated results against the original clean images with four full-reference metrics; we exclude tasks such as low-light enhancement and dehazing here, since their acceptable outputs may differ in global brightness and bias full-reference evaluation. For perceptual quality, we use 200 collected real-world images (20 each for rain, haze, snow, low-light, and old photo, plus 100 general mixed) with five no-reference metrics. To capture problems that IQA metrics miss, such as hallucinated objects or unreasonable edits, we further report the VLM-estimated restoration success rate (VLM-R); details are in **Appendix C**.

All metrics are first converted to the same direction (higher is better) and min–max normalized to $\tilde{m}_{i,j}$, where $m_{i,j}$ is metric j for the i-th model-prompt setting. Normalized values are averaged within each aspect $a \in \{\text{fid}, \text{per}, \text{vlm}\}$ and then across the three aspects:

$$\tilde{m}_{i,j} = \frac{m_{i,j} - m_j^{\min}}{m_j^{\max} - m_j^{\min}}, \qquad s_i^a = \frac{1}{|\mathcal{M}_a|} \sum_{j \in \mathcal{M}_a} \tilde{m}_{i,j}, \qquad \text{Avg.}_i = \frac{1}{3}\left(s_i^{\text{fid}} + s_i^{\text{per}} + s_i^{\text{vlm}}\right). \tag{1}$$

We also conduct a user study (details are in **Appendix D**) on these 200 real-world samples: for each of the nine MFMs, we select its best-performing prompt setting, present the LQ input together with nine anonymous results, and record how often each method is chosen as the best ("Human" in Tab. 1).

Evaluation Results. The results are reported in Tab. 1. First, we see that both model choice and prompting strategy have a substantial impact on restoration behavior, and the performance gap across different model–prompt combinations can be large. Even for the same MFM, different prompts may lead to noticeably different outcomes. For example, Nano-Banana-2 improves from an Avg. score of 0.76 under fixed prompting to 0.84 under Gemini-based adaptive prompting. This indicates that prompting sensitivity should be taken seriously when evaluating and using MFMs for restoration.

Second, different MFMs exhibit clear preference biases. As shown in Fig. 4 and Tab. 1, some models preserve input content better but produce conservative restorations, while others generate stronger visual enhancement at the cost of content inconsistency. For example, Qwen-Image-Edit and Kling-Image-o1 achieve high fidelity scores, but their perceptual scores and visual results are



Figure 4: Typical outputs of the MFMs on real-world LQ inputs. For each model, we show its best-performing prompt setting. The examples highlight the different restoration preferences of these MFMs, especially the trade-off between perceptual enhancement and fidelity preservation.

relatively modest. In contrast, GPT-Image family and the FireRed-1.1 obtain better perceptual metrics, but their fidelity is much weaker and they often significantly change the content of the image.

Overall, **Nano-Banana-2 with Gemini-based adaptive prompting** is the most well-balanced setting among all candidates. It achieves the best Avg. score of 0.84 and the highest human preference of 32.5%. Although it does not achieve the best score on every perceptual metric, it achieves highly competitive perceptual results. Rather than excelling in only one aspect, it performs strongly and consistently across all the four aspects, maintaining a favorable balance between perceptual quality and content faithfulness. We therefore choose it as the MFM for GGT-100K construction.

3.3 Multi-stage Quality Control

Even with the best-performing MFM and adaptive prompting strategy, we cannot guarantee that the generated outputs are always ideal HQ targets. Therefore, we further apply a multi-stage quality control process to improve the reliability of the generated HQ targets.

Metric-based Filtering. We first conduct coarse automatic filtering by no-reference perceptual quality metrics. Specifically, for each LQ image and its HQ counterpart generated by Nano-Banana-2, we compare their metric scores. If the score shows little improvement or even worse, we regard this sample as a failure case and exclude it from the dataset.

VLM-assisted Refinement. Metric-based filtering is effective for coarse selection but cannot reliably capture subtler MFM failures such as local structure distortion, semantic inconsistency, or unreasonable edits. We thus add a VLM-assisted refinement step, in which the VLM (Team et al., 2023) assesses every generated result along five aspects important for an HQ target: *restoration quality*, *object consistency*, *geometry alignment*, *content reasonableness*, and *color consistency*. Unsatisfactory samples are not discarded; instead we use the VLM feedback to regenerate them by appending prompt instructions tailored to the identified issues, then re-evaluate. The VLM thus serves as both quality controller and feedback source. Details are in **Appendix C**.

Manual Verification. Finally, we perform manual verification as the last safeguard of the pipeline. Human annotators review the automatically retained samples and remove those that still contain noticeable artifacts, structural inconsistency or unreasonable content changes.

By combining metric filtering, VLM-assisted refinement, and manual verification, we obtain a more reliable set of restoration-oriented targets. As shown in Fig. 1, GGT-100K contains 103K training pairs and 500 test pairs at a unified resolution of 1024×1024; the test pairs are jointly reviewed by multiple researchers to ensure high fidelity, strong restoration quality, and no obvious hallucinated content. GGT-100K thus covers diverse real degradations arising naturally in real photography, storage, and transmission, making it particularly suitable for improving the generalization performance of real-world IR models.

Table 2: Comparison of representative restoration models "w/o" and "w/" GGT-100K on our test set. For some models whose official releases can also handle real-world degradations, we additionally report their **official results**. "Improvement" indicates the performance gain brought by GGT-100K. The positive improvements are highlighted in **red**.

Model	GGT-100K	Full-reference fidelity metrics				No-reference perceptual metrics					VLM-R ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓	
MPRNet (Zamir et al., 2021)	w/o	24.7919	0.7637	0.3779	0.2390	6.9774	42.1039	0.4641	0.3120	-0.7140	22.2%
	w/	27.3044	0.8189	0.3377	0.2212	7.0035	44.9597	0.4431	0.3075	-0.7794	33.2%
	Improvement	**+2.5125**	**+0.0552**	**-0.0402**	**-0.0178**	+0.0261	**+2.8558**	-0.0210	-0.0045	**-0.0654**	**+11.0%**
NAFNet (Chen et al., 2022)	w/o	25.1255	0.7708	0.3653	0.2298	6.6654	42.0124	0.4635	0.3118	-0.7012	27.6%
	w/	28.2461	0.8349	0.3110	0.2043	6.7983	46.7094	0.4330	0.3097	-0.7881	53.8%
	Improvement	**+3.1206**	**+0.0641**	**-0.0543**	**-0.0255**	+0.1329	**+4.6970**	-0.0305	-0.0021	**-0.0869**	**+26.2%**
SwinIR (Liang et al., 2021)	w/o	23.9225	0.7590	0.3878	0.2343	6.4979	41.0176	0.4488	0.3082	-0.6940	18.6%
	w/	27.0781	0.8150	0.3369	0.2131	6.7366	43.6689	0.4308	0.2966	-0.7437	37.6%
	Improvement	**+3.1556**	**+0.0560**	**-0.0509**	**-0.0212**	+0.2387	**+2.6513**	-0.0180	-0.0116	**-0.0497**	**+19.0%**
XRestormer (Chen et al., 2024b)	w/o	24.6901	0.7705	0.3587	0.2234	6.6564	43.1437	0.4643	0.3192	-0.7185	30.4%
	w/	28.2298	0.8362	0.3130	0.2065	6.9218	46.8069	0.4295	0.3109	-0.8126	54.6%
	Improvement	**+3.5397**	**+0.0657**	**-0.0457**	**-0.0169**	+0.2654	**+3.6632**	-0.0348	-0.0083	**-0.0941**	**+24.2%**
PromptIR (Potlapalli et al., 2023)	w/o	24.5733	0.7564	0.3630	0.2312	6.5942	43.2318	0.4695	0.3202	-0.7176	24.8%
	w/	28.1775	0.8344	0.3113	0.2051	6.8029	46.7198	0.4369	0.3124	-0.7968	49.6%
	Improvement	**+3.6042**	**+0.0780**	**-0.0517**	**-0.0261**	+0.2087	**+3.4880**	-0.0326	-0.0078	**-0.0792**	**+24.8%**
MoCE-IR (Zamfir et al., 2024)	w/o	24.8641	0.7575	0.3654	0.2291	6.6248	43.0279	0.4671	0.3200	-0.7116	25.4%
	w/	28.2140	0.8392	0.3097	0.2093	6.9402	48.4846	0.4344	0.3242	-0.8270	55.2%
	Improvement	**+3.3499**	**+0.0817**	**-0.0557**	**-0.0198**	+0.3154	**+5.4567**	-0.0327	**+0.0042**	**-0.1154**	**+29.8%**
DA-CLIP (Luo et al., 2024)	w/o	25.8809	0.7726	0.3427	0.2180	6.3011	39.3370	0.4846	0.3080	-0.6853	31.6%
	w/	26.6914	0.7910	0.2954	0.1938	6.3039	41.8839	0.4834	0.3073	-0.7277	51.0%
	Improvement	**+0.8105**	**+0.0184**	**-0.0473**	**-0.0242**	+0.0028	**+2.5469**	-0.0012	-0.0007	**-0.0424**	**+19.4%**
FoundIR (Li et al., 2025)	**official**	26.0398	0.7866	0.3486	0.2199	6.5311	39.3646	0.4793	0.3048	-0.6971	28.8%
	w/o	25.8048	0.7844	0.3508	0.2220	6.9023	42.5388	0.4701	0.3153	-0.7365	35.8%
	w/	27.1777	0.8213	0.3351	0.2209	7.4717	43.9238	0.4402	0.3019	-0.8087	60.8%
	Improvement	**+1.3729**	**+0.0369**	**-0.0157**	**-0.0011**	+0.5694	**+1.3850**	-0.0299	-0.0134	**-0.0722**	**+25.0%**
FLUX-Controlnet (Labs, 2024)	w/o	22.4486	0.6901	0.3773	0.2129	5.5636	48.5454	0.5157	0.3798	-0.6912	25.4%
	w/	23.1413	0.7325	0.2625	0.1520	4.8504	63.0910	0.5854	0.5013	-0.9280	63.4%
	Improvement	**+0.6927**	**+0.0424**	**-0.1148**	**-0.0609**	**-0.7132**	**+14.5456**	**+0.0697**	**+0.1215**	**-0.2368**	**+38.0%**
Qwen-Image-Edit (Wu et al., 2025)	**official**	22.3141	0.7479	0.3042	0.1693	5.4776	60.8628	0.5576	0.4709	-0.9565	68.0%
	w/o	25.8559	0.7787	0.2813	0.1625	6.0198	51.4215	0.5333	0.3765	-0.8401	77.4%
	w/	26.1811	0.7828	0.2155	0.1183	5.4648	62.5519	0.5811	0.4707	-0.9611	87.6%
	Improvement	**+0.3252**	**+0.0041**	**-0.0658**	**-0.0442**	**-0.5550**	**+11.1304**	**+0.0478**	**+0.0942**	**-0.1210**	**+10.2%**

4 Expanding Real-World IR Boundaries using GGT-100K

4.1 Experimental Settings

Representative IR Models. We employ a series of representative restoration models, including L_1-loss optimized CNN and transformer backbones (MPRNet (Zamir et al., 2021), NAFNet (Chen et al., 2022), SwinIR (Liang et al., 2021), X-Restormer (Chen et al., 2024b)) and all-in-one models (PromptIR (Potlapalli et al., 2023), MoCE-IR (Zamfir et al., 2024), DA-CLIP (Luo et al., 2024), FoundIR (Li et al., 2025)). In addition, since finetuning large-scale generative models has shown impressive real-world IR performance (Wu et al., 2024; Yu et al., 2024), we also finetune the T2I model FLUX.1-dev (Labs, 2024) and the TI2I model Qwen-Image-Edit-2511 (Wu et al., 2025) on GGT-100K in the experiments.

Dataset Settings. We train IR models under two settings. **(1) Existing training data ("w/o" in Tab. 2)**, containing 200K training pairs from 15 existing synthetic and real-world datasets; details are in **Appendix E**. **(2) Existing training data + GGT-100K ("w/" in Tab. 2)**, which adds GGT-100K to the training pool with a **1:1 sampling ratio**. This reflects the intended use of GGT-100K in practice: *as a complementary real-world data source to expand the generalization boundary of IR models rather than a replacement for existing datasets*. For testing sets, we use two benchmark groups: our GGT-100K test set with paired GGT references, and public real-world test sets without GT (social media and old photo of RealDeg (Chen et al., 2025b), and OpenReal80K (Lin et al., 2025) subsets covering haze, rain, snow, and night scenarios).

Evaluation. As in Sec. 3.2, we measure content fidelity with the full-reference PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2020), perceptual quality with the no-reference NIQE (Zhang et al., 2015), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), TOPIQ (Chen et al., 2024a), and AFINE-NR (Chen et al., 2025a), and further report VLM-R for a multimodal semantic perspective. It sould be noted that although the reference images in the GGT test set are generated rather than physically captured, they are carefully filtered and verified so that they can serve as plausible clean references. The resulting full-reference metrics therefore provide a useful, albeit approximate, assessment of degradation removal and reconstruction quality.

Table 3: Comparison of representative restoration models "w/o" and "w/" GGT-100K on public RealLQ test sets without GT, including social media (Chen et al., 2025b), old photos (Chen et al., 2025b), rain (Lin et al., 2025), haze (Lin et al., 2025), snow (Lin et al., 2025), and low light (Lin et al., 2025). The average results over the six test sets are reported. "Improvement" indicates the performance gain brought by using GGT-100K. Positive improvements are highlighted in **red**.

Model	GGT-100K	NIQE ↓		MUSIQ ↑		MANIQA ↑		TOPIQ ↑		AFINE-NR ↓		VLM-R ↑	
		Value	Imp.	Value	Imp.	Value	Imp.	Value	Imp.	Value	Imp.	Value	Imp.
MPRNet (Zamir et al., 2021)	w/o	5.1502		56.2493		0.5928		0.4427		-0.8439		38.0%	
	w/	5.0973	**-0.0529**	58.6289	**+2.3796**	0.5937	**+0.0009**	0.4540	**+0.0113**	-0.8762	**-0.0323**	45.0%	**+7.0%**
NAFNet (Chen et al., 2022)	w/o	4.9061		57.9680		0.5986		0.4576		-0.8559		36.0%	
	w/	5.0503	+0.1442	57.5578	-0.4102	0.5841	-0.0145	0.4348	-0.0228	-0.8665	**-0.0106**	45.0%	**+9.0%**
SwinIR (Liang et al., 2021)	w/o	4.8474		56.3785		0.5791		0.4500		-0.8176		28.0%	
	w/	5.0991	+0.2517	57.4882	**+1.1097**	0.5832	**+0.0041**	0.4473	-0.0027	-0.8539	**-0.0363**	38.7%	**+10.7%**
XRestormer (Chen et al., 2024b)	w/o	4.9505		57.2285		0.6030		0.4589		-0.8625		43.3%	
	w/	5.0702	+0.1197	58.0245	**+0.7960**	0.5828	-0.0202	0.4411	-0.0178	-0.8801	**-0.0176**	51.3%	**+8.0%**
PromptIR (Potlapalli et al., 2023)	w/o	4.9273		57.1754		0.6029		0.4556		-0.8507		41.3%	
	w/	5.0520	+0.1247	58.1513	**+0.9759**	0.5901	-0.0128	0.4409	-0.0147	-0.8753	**-0.0246**	48.3%	**+7.0%**
MoCE-IR (Zamfir et al., 2024)	w/o	4.9853		57.9966		0.6053		0.4611		-0.8622		39.0%	
	w/	5.2014	+0.2161	58.3756	**+0.3790**	0.5837	-0.0216	0.4411	-0.0200	-0.8845	**-0.0223**	45.7%	**+6.7%**
DA-CLIP (Luo et al., 2024)	w/o	4.7681		57.3662		0.6067		0.4637		-0.8592		46.7%	
	w/	4.8087	+0.0406	58.1019	**+0.7357**	0.6032	-0.0035	0.4568	-0.0069	-0.8804	**-0.0212**	55.0%	**+8.3%**
FoundIR (Li et al., 2025)	**official**	5.1570		58.0616		0.6131		0.4556		-0.8610		52.3%	
	w/o	5.0551	+0.4790	59.2164	-1.1196	0.6143	-0.0237	0.4803	-0.0390	-0.8837	**-0.0079**	48.7%	**+7.7%**
	w/	5.5341		58.0968		0.5906		0.4413		-0.8916		56.3%	
FLUX-Controlnet (Labs, 2024)	w/o	4.4951		60.9375		0.6431		0.4795		-0.8779		47.7%	
	w/	4.0684	**-0.4267**	67.5600	**+6.6225**	0.6828	**+0.0397**	0.5848	**+0.1053**	-1.0784	**-0.2005**	59.0%	**+11.3%**
Qwen-Image-Edit (Wu et al., 2025)	**official**	4.3440		67.6115		0.6647		0.6069		-1.2154		56.7%	
	w/o	4.6176	**-0.6793**	63.1324	**+7.2735**	0.6524	**+0.0383**	0.4945	**+0.1340**	-1.0204	**-0.1825**	60.7%	**+10.0%**
	w/	3.9383		70.4059		0.6907		0.6285		-1.2029		70.7%	

Implementation Details. Models trained from scratch follow their original settings. For FLUX.1-dev (Labs, 2024) and Qwen-Image-Edit-2511 (Wu et al., 2025), we finetune with a simple prompt at 512×512 input resolution, randomly cropped from GGT-100K: FLUX.1-dev with a ControlNet-style setup (4 double blocks, 0 single block), batch size 16, learning rate 5×10^{-5}, 200K iterations; Qwen-Image-Edit-2511 with rank-8 LoRA (Hu et al., 2021), batch size 64, learning rate 1×10^{-4}, 50K iterations. All experiments use PyTorch (Paszke et al., 2019) on 32 NVIDIA A800 GPUs.

4.2 Experimental Results

Results on GGT-100K Test set. As shown in Tab. 2, incorporating GGT-100K generally benefits real-world restoration across diverse model families. Training with GGT-100K improves all fidelity metrics, as well as MUSIQ, AFINE-NR and VLM-R, for all evaluated IR models. Notably, AFINE-NR is a more recent no-reference metric designed for the era of generative models, making it better aligned with human preference. Since many no-reference metrics are not well-suited for weather-related scenarios such as deraining and desnowing, VLM-R provides an important complementary metirc, and adding GGT-100K in training yields clear VLM-R gains across different models.

A closer look at model types reveals two findings. First, the fidelity gains are large for fidelity-oriented methods. X-Restormer (Chen et al., 2024b) and PromptIR (Potlapalli et al., 2023) both achieve more than 3.5 dB PSNR gains, along with clear improvements in SSIM, LPIPS, and DISTS. This suggests that GGT-100K provides effective supervision for faithful reconstruction. In contrast, for generative models, such as FLUX-Controlnet (Labs, 2024) and Qwen-Image-Edit (Wu et al., 2025), GGT-100K substantially improves all metrics, with much larger perceptual gains. This indicates that GGT-100K is particularly suitable for generative models with strong priors and learning capacity.

The comparison with officially released models further supports this. For both FoundIR and Qwen-Image-Edit, training with GGT-100K brings broad improvements over their official versions. Qwen-Image-Edit is representative: its official version has strong generative ability but limited fidelity and over-generation; training only on existing data improves fidelity while degrading perceptual quality, whereas GGT-100K improves both (see Fig. F.2 in **Appendix F**).

Results on Public RealLQ Test sets. The average results on the existing RealLQ test sets are listed in Tab. 3, which shows a similar trend to Tab. 2. Adding GGT-100K consistently improves AFINE-NR and VLM-R for all evaluated IR models, and MUSIQ for most models, with especially notable gains for generative IR models. The VLM-R improvements are smaller than that on GGT-100K test set, probably because these test sets contain samples with only mild or light degradations so that models trained without GGT-100K can also work.

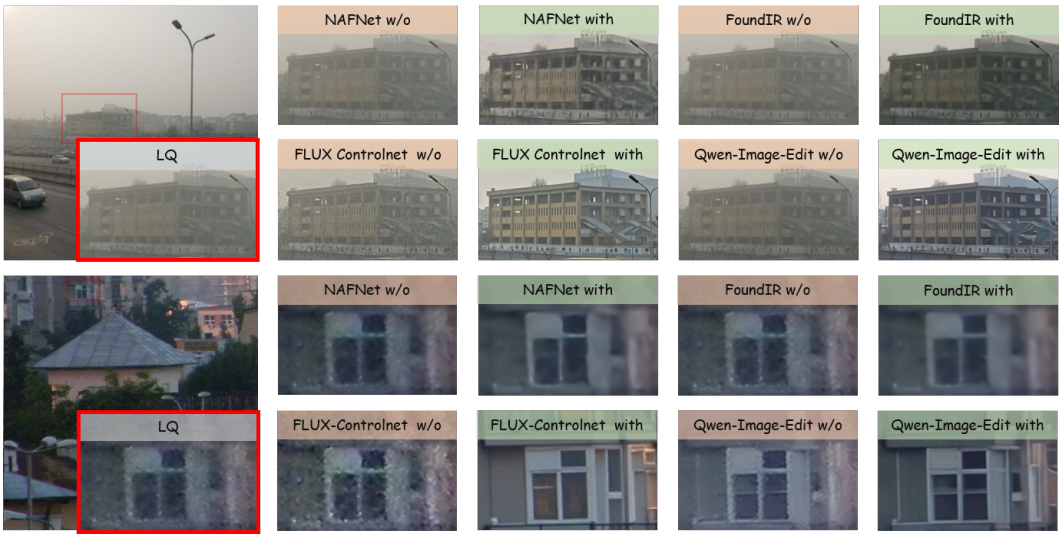


Figure 5: Qualitative comparison of restoration methods trained without and with GGT-100K. Training with GGT-100K yields clearer details and more faithful structures under real-world degradations.

Visual Results. Fig. 5 compares four methods covering distinct paradigms: L_1-loss-based NAFNet (Chen et al., 2022), train-from-scratch diffusion model FoundIR (Li et al., 2025), pre-trained T2I-based FLUX-Controlnet (Labs, 2024), and editing-model-based Qwen-Image-Edit (Wu et al., 2025). In the first case with real haze, all four models trained only on existing datasets leave noticeable residual haze; adding GGT-100K clearly improves haze removal for all of them. The slight color differences among the restored results are expected, since the true scene colors under real haze are inherently unknown and each method makes a reasonable estimation based on its own learned prior. NAFNet and FoundIR, lacking strong generative priors, struggle to recover fine building details, whereas FLUX-Controlnet and Qwen-Image-Edit reconstruct sharper and more faithful textures. The second case with real noise shows a similar trend: without GGT-100K, all four models leave residual noise of unknown real-world distribution, and FLUX-Controlnet even erroneously amplifies the input noise, while adding GGT-100K substantially improves their denoising capability. More visual results are in **Appendix F**.

Ablation Study We conduct a comparison of GGT data before (directly generated by MFMs) and after quality control (final GGT-100K) to demonstrate the effectiveness of our pipeline in **Appendix G**.

5 Conclusion

We proposed GGT, a practical paradigm for constructing real-world paired training data for image restoration using generative MFMs. Based on a systematic evaluation of nine MFMs and multiple prompting strategies, we built GGT-100K, a large-scale real-world LQ-HQ paired dataset with 103K training pairs and a curated test set of 500 pairs. Extensive experiments showed that GGT-100K consistently improved the generalization performance, in terms of both content fidelity and perceptual quality, of various restoration models, particularly those finetuned generative models, whose strong priors and high learning capacity enable them to better exploit the advantages of GGT-100K. These results validated the use of MFMs for restoration-oriented data generation and demonstrated the value of GGT-100K as a useful resource for advancing generalizable real-world image restoration.

Limitations. GGT-100K should be regarded as a scalable and high-quality approximation of the real-world ground-truth. Its scope is deliberately disjoint from degradations that can be synthesized or for which real pairs can be easily captured. GGT-100K is a complement to, not a substitute for, these pipelines. It targets the far larger pool of already-captured real-world LQ images, for which the corresponding physically collected GT images are nearly impossible to obtain retrospectively. Although with multi-stage quality control, some of the images in GGT-100K can still have subtle visual artifacts hallucinated by MFMs. Nevertheless, compared with synthetic degradation pipelines or costly real-world paired acquisition, our approach offers a more practical way to build diverse real-world LQ-HQ pairs.

Ethics Statement

This work introduces GGT-100K, a dataset of real-world low-quality images paired with high-quality targets synthesized by generative multimodal foundation models (MFMs). The source low-quality images are collected from Internet sources with permissive usage rights (including CC0-licensed sources), from publicly released restoration datasets used under their original licenses, and from our own captures; all public datasets and models we build upon are cited in the paper. We did not collect personally identifiable information, and images containing identifiable individuals were removed during the filtering stage described in Appendix A. The user study reported in Sec. 3.2 and Appendix D was minimal risk: participants only compared pairs of restored images and judged perceptual quality and fidelity, with no collection of sensitive personal data; participants were informed of the purpose of the study and participated voluntarily. No IRB approval was sought, as the study falls within the minimal-risk category. We note one intrinsic limitation with ethical relevance: because the high-quality targets are *generated* rather than physically captured, they may contain hallucinated content, and models trained on GGT-100K should therefore not be used in forensic, medical, or evidentiary settings where content fidelity must be guaranteed. We state this explicitly in the Limitations paragraph and will document it in the dataset card released with the data.

Reproducibility Statement

We have made the following efforts to ensure that our results are reproducible. The full construction pipeline for GGT-100K — source collection and filtering, the MFM evaluation protocol, the VLM-based adaptive prompting strategy, and the multi-stage quality control — is described in Sec. 3 and detailed in Appendices A–C, including the exact prompt templates used for each MFM and for the VLM evaluator. The training configuration for every restoration model we finetune (data composition, optimizer, learning rate, schedule, and number of iterations) is given in Appendix E, and the evaluation protocol and metrics are given in Sec. 4.1. An ablation of each quality-control stage is reported in Appendix G, and per-degradation results for all evaluated models are reported in Appendix H. The GGT-100K training set, the 500-pair test set, and the accompanying documentation will be made publicly available upon publication.

AI Use Statement

Generative AI tools are a core object of study in this work rather than an incidental writing aid, and we disclose their use as follows. Generative multimodal foundation models (Nano-Banana-2, GPT-Image-2, and the other seven MFMs listed in Sec. 3.2) were used to synthesize every high-quality target image in GGT-100K; this is the central methodological contribution of the paper and is described in full in Sec. 3. Vision-language models were additionally used for adaptive prompt generation and as automatic evaluators within our multi-stage quality control pipeline, as described in Sec. 3.3 and Appendix C. All automatic quality judgments were validated against human preferences through the user study reported in Appendix D. We also used a large language model for grammar and phrasing improvements on the manuscript text; all research ideas, experimental designs, analyses, and conclusions are the authors' own, and every AI-assisted passage was reviewed by the authors. We take full responsibility for the final content of this work.

References

Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1692–1700, 2018.

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017.

Anthropic. Claude opus 4.8. `https://www.anthropic.com/claude`, 2026.

ByteDance. Seedream 4.0-5.0 tutorial. `https://docs.byteplus.com/zh-CN/docs/ModelArk/1824121`, 2025.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3086–3095, 2019.

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

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12742–12752, 2025a.

Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, and He Li. Robust video content alignment and compensation for rain removal in a cnn framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6341–6349, 2018. doi: 10.1109/CVPR.2018.00658.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. FaithDiff: Unleashing diffusion priors for faithful image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025b.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pp. 17–33. Springer, 2022.

Xiang Chen, Hao Li, Jiangxin Dong, Jinshan Pan, Xin Li, Xin He, Naiwei Chen, Shengyuan Li, Fengning Liu, Haoyi Lv, et al. Lovif 2026 challenge on real-world all-in-one image restoration: Methods and results. *arXiv preprint arXiv:2604.19445*, 2026.

Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024b.

Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4641–4650, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pp. 184–199. Springer, 2014.

Flickr. Website. `https://www.flickr.com`, 2025.

Qiyuan Guan, Qianfeng Yang, Xiang Chen, Tianyu Song, Guiyue Jin, and Jiyu Jin. Weatherbench: A real-world benchmark dataset for all-in-one adverse weather image restoration. In *Proceedings of the 33rd ACM international conference on multimedia*, pp. 12607–12613, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

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*, pp. 5148–5157, 2021.

Xiangtao Kong, Jinjin Gu, Yihao Liu, Wenlong Zhang, Xiangyu Chen, Yu Qiao, and Chao Dong. A preliminary exploration towards general image restoration. *arXiv preprint arXiv:2408.15143*, 2024.

Black Forest Labs. Flux. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2019.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-In-One Image Restoration for Unknown Corruption. In *IEEE Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, June 2022a.

Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. In *ICLR*, 2023.

Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12626–12636, 2025.

Wei Li, Qiming Zhang, Jing Zhang, Zhen Huang, Xinmei Tian, and Dacheng Tao. Toward real-world single image deraining: A new benchmark and beyond. *arXiv preprint arXiv:2206.05514*, 2022b.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

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

Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Wen Kairun, Yeying Jin, Wenbo Li, and Xinghao Ding. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yihao Liu, Hengyuan Zhao, Jinjin Gu, Yu Qiao, and Chao Dong. Evaluating the generalization ability of super-resolution networks. *arXiv preprint arXiv:2205.07019*, 2022.

Yihao Liu, Hengyuan Zhao, Jinjin Gu, Yu Qiao, and Chao Dong. Evaluating the generalization ability of super-resolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. doi: https://doi.org/10.1016/j.cviu.2018.10.010.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Photo-realistic image restoration in the wild with controlled vision-language models. *arXiv preprint arXiv:2404.09732*, 2024.

OpenAI. Gpt-image-1.5 model documentation. https://platform.openai.com/docs/models/gpt-image-1.5, 2025a.

OpenAI. Gpt-image-2 model documentation. https://platform.openai.com/docs/models/gpt-image-2, 2025b.

OpenAI. ChatGPT GPT-5.4 Release Notes. https://help.openai.com/en/articles/6825453-chatgpt-release-notes, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pexels. Website. https://www.pexels.com, 2025.

Pixabay. Website. https://pixabay.com, 2025.

Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1586–1595, 2017.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9147–9156, 2021.

Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European conference on computer vision*, pp. 184–201. Springer, 2020.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021.

Weixiong Sun, Xiang Yin, and Chao Dong. Can nano banana 2 replace traditional image restoration models? an evaluation of its performance on image restoration tasks. *arXiv preprint arXiv:2604.03061*, 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Kling Team and MiraclePlus. Kling-image-o1: Technical report on high-fidelity video generation. https://klingai.com, 2025.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 114–125, 2017.

Unsplash. Website. https://unsplash.com, 2025.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

Hongyang Wei, Shuaizheng Liu, Chun Yuan, and Lei Zhang. Perceive, understand and restore: Real-world image super-resolution with autoregressive multimodal generative models. *arXiv preprint arXiv:2503.11073*, 2025.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 101–117. Springer, 2020.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25456–25467, 2024.

J Xu, H Li, Z Liang, D Zhang, and L Zhang. Real-world noisy image denoising: A new benchmark. arxiv 2018. *arXiv preprint arXiv:1804.02603*, 2018.

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*, pp. 1191–1200, 2022.

Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1357–1366, 2017.

Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

Yufeng Yang, Xianfang Zeng, Zhangqi Jiang, Fukun Yin, Jianzhuang Liu, Wei Cheng, Shiyu Liu, Yuqi Peng, Gang YU, Shifeng Chen, et al. Realrestorer: Towards generalizable real-world image restoration with large-scale image editing models. *arXiv preprint arXiv:2603.25502*, 2026.

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*, pp. 25669–25680, 2024.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration, 2024.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, 2018.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7): 3142–3155, 2017.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4791–4800, 2021a.

Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and geometric priors. *IEEE Transactions on Image Processing*, 2021b.

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Appendix

In this appendix, we provide the following materials:

- **A.** More details of source image collection for GGT-100K (referring to Sec. 3.1).
- **B.** Detailed prompt designs for MFM evaluation (referring to Sec. 3.2).
- **C.** The usage of VLM as evaluator and quality controller (referring to Sec. 3.2 and Sec. 3.3).
- **D.** Details of the user study (referring to Sec. 3.2).
- **E.** Details of training data composition (referring to Sec. 4.1).
- **F.** Additional qualitative comparisons and visual results (referring to Sec. 4.2).
- **G.** Ablation study of the multi-stage quality control pipeline (referring to Sec. 4.2).
- **H.** Detailed results on specific degradation categories (referring to Sec. 4.2).

A Details of Source Image Collection

In this section, we provide more details on the source image collection process for GGT-100K. We collect real-world LQ images from three main sources: Internet sources, public datasets, and our own captures. The collected images are organized into six categories: *General Mixed*, *Low-Light*, *Haze*, *Rain*, *Snow*, and *Old Photo*. Because the raw images from these sources vary substantially in content, degradation type, and perceptual quality, we further apply source-specific filtering, cropping, and quality-control procedures before constructing the final dataset.

For web-collected images, we adopt category-specific collection and filtering strategies for different degradation categories, and all collected images from the Internet are restricted to those released under CC0 licenses. For *Rain* and *Snow*, we search online image platforms (Unsplash, 2025; Pexels, 2025; Pixabay, 2025; Flickr, 2025) using keywords associated with visible rain and snow phenomena (e.g., rain, rain streak, heavy rain, snow, and snowstorm). After removing duplicate results, we manually inspect the full-resolution images and discard those without clearly visible artifacts, and then resize and crop the remaining images into patches, yielding over 30K candidate patches before filtering. We further apply VLM-based filtering using Gemini to remove patches in which the target artifacts are not sufficiently apparent, resulting in more than 10K retained patches from the *Rain* and *Snow* categories combined. For *Old Photo*, we retrieve over 10K historical photographs under CC0 licenses shared by Flickr (Flickr, 2025) contributors who specifically curate old-photo content. For *General Mixed*, we collect over 20K CC0-licensed user-uploaded images from online image platforms (Unsplash, 2025; Pexels, 2025; Pixabay, 2025; Flickr, 2025). To preserve the original noise characteristics of *Old Photo* and *General Mixed* types, we do not resize these images and instead crop them directly into patches, producing over 100K candidate patches. We then use Gemini to remove content-poor patches (e.g., regions with little structural information) and overly degraded patches that have already lost most useful content. Finally, we divide the remaining patches into high-, medium-, and low-quality levels based on subjective criteria, and sample them with a ratio of 1:2:3 to increase the proportion of moderately and severely degraded samples.

For dataset-sourced images, we first gather over 10K images from existing real-world restoration datasets that do not provide ground-truth references (Li et al., 2019; Zhang et al., 2021b; Chen et al., 2018), as well as over 1,000K images from other image datasets containing low-quality or bad-weather conditions (Deng et al., 2009; Loh & Chan, 2019; Yang et al., 2020; Sakaridis et al., 2021). We then retain only images with clear and suitable degradations through metric-based and VLM-based screening and subjective filtering. In addition, we include a self-captured subset, where we manually adjust camera and smartphone parameters such as exposure time and ISO to capture low-light and noisy images, and also vary the focus or introduce physical camera shake during capture to obtain realistic blurry images. These images are further cropped into over 20K patches and filtered using VLMs to keep only suitable samples.

Overall, we screen more than 1,100K image patches through the above procedures and retain about 120K low-quality images. We then use Nano-Banana-2 to generate corresponding high-quality counterparts and further filter out image pairs through the multi-stage quality control described in Sec. 3.3. Owing to the strong generation quality of Nano-Banana-2, we retain 103,707 image pairs

Table B.1: Fixed prompt and fixed-prompt-no-change used for different degradation types.

Degradation Type	Fixed-Prompt	Fixed-Prompt-No-Change
GGT General Mixed	Restore this low-quality image to a clear, clean and natural state.	Restore this low-quality image to a clear, clean and natural state. But do not change the image content.
GGT Rain	Please remove the rain from the image and restore its clarity.	Please remove the rain from the image and restore its clarity. But do not change the image content.
GGT Haze	Please dehaze the image.	Please dehaze the image. But do not change the image content.
GGT Snow	Please remove the snow from the image and restore its clarity.	Please remove the snow from the image and restore its clarity. But do not change the image content.
GGT Low-Light	Please restore this low-quality image, recovering its normal brightness and clarity.	Please restore this low-quality image, recovering its normal brightness and clarity. But do not change the image content.
GGT Old Photo	Please restore the old photo to a clear, clean and natural state.	Please restore the old photo to a clear, clean and natural state. But do not change the image content.

after this multi-stage quality control process. Among them, *General Mixed* contains 66,058 pairs, *Low-Light* contains 9,786 pairs, *Haze* contains 7,822 pairs, *Rain* contains 7,177 pairs, *Snow* contains 6,759 pairs, and *Old Photo* contains 6,105 pairs.

General Mixed. The *General Mixed* category contains 66,058 pairs collected from three sources: Internet images, public datasets, and our own captures. Specifically, it includes 37,463 pairs from Internet images mainly collected from Flickr (Flickr, 2025), 19,438 pairs from public dataset images mainly from ImageNet (Deng et al., 2009), where we select images with relatively poor visual quality, and 9,157 pairs from our own captures.

Rain. The *Rain* category contains 7,177 pairs collected from both public datasets and Internet sources, including 40 pairs from NTURain (Chen et al., 2018), 2,222 pairs from OpenReal80K (Lin et al., 2025), 1,590 pairs from Flickr (Flickr, 2025), 1,806 pairs from Unsplash (Unsplash, 2025), and 1,517 pairs from Pexels (Pexels, 2025).

Haze. The *Haze* category contains 7,822 pairs, all collected from public datasets, including 2,000 pairs from ACDC (Sakaridis et al., 2021) and 5,822 pairs from RTTS (Li et al., 2019).

Snow. The *Snow* category contains 6,759 pairs, including 5,202 pairs from Unsplash (Unsplash, 2025) and 1,557 pairs from Snow100K-Real (Zhang et al., 2021b).

Low-Light. The *Low-Light* category contains 9,786 pairs, including 5,702 pairs from DarkFace (Yang et al., 2020), 2,858 pairs from ExDark (Loh & Chan, 2019), and 1,226 pairs from our own captures.

Old Photo. The *Old Photo* category contains 6,105 pairs, all collected from Flickr (Flickr, 2025) from different providers, ensuring source diversity within this category.

B Detailed Prompt Designs for MFMs Evaluation

In Sec. 3.2 and Tab. 1 of the main paper, we evaluate 9 candidate MFMs under 4 prompting strategies. Here we provide the detailed designs of these prompts.

As shown in Tab. B.1, similar to (Yang et al., 2026), our fixed prompt and fixed-prompt-no-change adopt the simplest instruction form. A simple "do not change the image content" instruction consistently improves content fidelity for most MFMs.

We show an example of a haze image to illustrate the generation of our VLM-based prompt in Fig. B.1. An input image is first fed into a VLM (GPT-5.4-Pro (OpenAI, 2026) or Gemini-3.1-Pro (Team et al., 2023)) together with a meta prompt including system prompt and user prompt. The VLM then returns a detailed description of the image content and degradations, along with explicit instructions for removing those degradations. As can be seen, the generated prompt is highly detailed and specific. Tab. 1 of the main paper demonstrate that most models benefit from VLM-based

An Example of Obtaining VLM Prompt (Haze)

SYSTEM PROMPT



Input Image

You are an expert in image quality analysis for haze removal. The user will provide one degraded image.

Analyze:
1) Scene content and depth structure (foreground/background, sky)
2) Haze characteristics (global veil, local dense haze, low contrast, color shift, bright-airlight effect).
3) Severity and where haze is strongest.

Then output one detailed English restoration prompt. It must:
- Set fidelity as the first priority: preserve scene content, geometry, and structure.
- In all regions where content is identifiable (including mid/background elements that are still visible), aim for the best possible haze removal and visibility recovery rather than a minimal change, while staying physically plausible and avoiding artifacts.
- Remove thin/light haze cleanly and noticeably on buildings/roads/edges/textures that are identifiable.
- Ensure visible-content regions look clean and transparent rather than slightly washed-out.
- Perform denoising and clarity/sharpness recovery for fine structures and edges.
- Preserve global color and overall tone as much as possible; do not intentionally shift hue, white balance, or global brightness. Allow only physically plausible changes caused by dehazing, while avoiding any stylized recoloring.
- Recover local contrast and true colors only within natural range, avoiding any color cast or stylization.
- Keep geometry/content unchanged; no hallucination, no object insertion/removal.
- Keep object count, object size, and object position unchanged; no translation, warping, deformation, or shape distortion.
- If there are human faces in the image, restore them clearly with special attention to skin texture, eyes, and hair. However, do not change the person's identity or distinctive facial features so that they become a different person.

USER PROMPT

Analyze this hazy image and produce a detailed restoration prompt in English. You must include all hard constraints in system_prompt without omission. Fidelity is the top priority: for mild/moderate haze with visible details, enforce clearly noticeable haze removal and remove residual whitish haze film on visible content; for dense uncertain haze, avoid hallucination and it may remain mostly unchanged. Also perform denoising + clarity recovery while preserving global color tone/white balance/brightness as much as possible, except physically plausible dehazing-induced visibility and moderate local contrast/color recovery (no stylized recoloring). Keep geometry/content unchanged, and keep object count, size, and position unchanged (no translation, warping, deformation, or shape distortion).. Output only the prompt:

VLM OUTPUT (Our VLM-Based Prompt)

Restore this outdoor scene with fidelity as the first priority, preserving the exact scene content, geometry, perspective, and structural layout of the temple-like building, road/pathway, side railings, foreground people, and all visible objects. The image contains a heavy gray haze/smog veil across the whole scene, strongest in the distant background and upper areas, with a bright-airlight whitish film reducing contrast on the building silhouette, road surface, railings, and clothing. Apply clearly noticeable haze removal in all mild-to-moderate haze regions where content is already identifiable, especially on the foreground person, the visible portions of the second person, the road edges, pavement texture, railing posts, and the temple roofline and walls. Recover the best possible visibility for all identifiable midground and background elements that are still visible, including the building contours and architectural edges, while staying physically plausible and avoiding artifacts. In the most densely obscured and uncertain distant haze regions, do not hallucinate missing architectural or environmental details; Perform denoising and clarity recovery, improving edge definition and fine structure on clothing folds, hoodie text, mask edges, railing shapes, roof contours, and pavement boundaries, but avoid over-sharpening halos, crunchy texture, or unnatural HDR appearance. Preserve the global color tone, overall brightness, and white balance as much as possible; do not intentionally recolor or stylize the image. Keep all content unchanged: no hallucination, no object insertion or removal, no translation, no warping, no deformation, no shape distortion. Keep object count, object size, and object position unchanged.

Figure B.1: A processing example of VLM prompts. Given an input image and designed meta prompt, the VLM produces a detailed, image-specific restoration instruction that describes both the image content and the degradations to be removed.

prompts, which provides detailed, image-specific guidance, and this effect is particularly strong for the Nano-Banana-2 that we finally adopt. Between GPT-based and Gemini-based VLM prompts, we choose the latter because it achieves better quantitative performance, and GPT-generated prompts often blur or anonymize facial regions. Therefore, we use the Gemini-based VLM prompt in our final setting. The other meta prompts are given below.

General Mixed:

> **System prompt:** You are an expert in image restoration. The user will provide one degraded image. Analyze: (1) scene content and key objects/structures; (2) degradation types such as sensor/read noise, compression artifacts, ringing/blocking, defocus blur, motion blur, and their combinations; and (3) severity distribution by region, especially where blur, noise, or compression is strongest. Then output one detailed English restoration prompt. It must set fidelity as the first priority and preserve the exact scene content, geometry, position, and structure; remove noise and compression artifacts cleanly while recovering detail and clarity for visible structures; keep the global color and overall tone unchanged as much as possible, avoiding stylized recoloring and hue-category shifts; keep object count, object size, and object position unchanged, with no translation, warping, deformation, or shape distortion; keep geometry and content unchanged, with no hallucination and no object insertion or removal; and, if human faces are present, restore them clearly with special attention to skin texture, eyes, and hair, without changing identity or distinctive facial features. Output only the prompt text, with no explanation.
>
> **User prompt:** Analyze this image and produce a detailed restoration prompt in English. You must include all hard constraints in the system prompt without omission. Focus on strong denoising, deartifacting, and clarity restoration for realistic degradations, including noise, compression artifacts, defocus blur, motion blur, and their combinations. Do not specifically prioritize foreground or background; instead require balanced full-image clarity improvement. Keep content, geometry, and color identity unchanged, avoid hallucination in uncertain regions. Output only the prompt.

Rain:

> **System prompt:** You are an expert in rainy-scene restoration. The user will provide one degraded image. Analyze: (1) scene content, depth layers, and moving/static structures; (2) rain artifacts, including streaks, raindrops, misty veil, contrast reduction, and windshield-like droplets; and (3) severity and affected regions. Then output one detailed English restoration prompt. It must remove rain artifacts cleanly and as completely as reliably possible, including visible rain streaks, raindrop artifacts, and rain haze veil; preserve the rainy weather ambiance and lighting mood while not leaving obvious residual rain artifacts on visible content; actively restore background sharpness and visibility when rain streaks or droplets reduce clarity, rather than merely removing rain while leaving the background blurred; recover contrast, visibility, textures, and overall clarity without making the scene appear dry or sunny; perform denoising and sharpness recovery for all regions, not only rain removal; keep content strictly unchanged, including scene/object identity, count, size, position, and geometry; keep color and overall tone unchanged, with no hue shift, saturation shift, white-balance change, or recoloring; avoid ghosting, smearing, and over-smoothing on fine edges; keep geometry and content unchanged with no hallucination; and, if human faces are present, restore them clearly with special attention to skin texture, eyes, and hair, without changing identity or distinctive facial features. Output only the prompt text, with no explanation.
>
> **User prompt:** Analyze this rainy image and produce a detailed restoration prompt in English. You must include all hard constraints in the system prompt without omission. Rain should be removed cleanly and as completely as possible, including rain streaks, droplets, and veil, and the result should not merely remove rain while keeping the background blurred. Preserve the rainy ambiance, keep content, geometry, and object layout unchanged, and keep global color and tone unchanged. Do not output any instruction related to watermark, logo, or text removal. Output only the prompt.

Snow:

> **System prompt:** You are an expert in snowy-scene restoration. The user will provide one degraded image. Analyze: (1) scene content and key objects; (2) snow-related degradations, including falling-flake occlusion, low contrast, haze-like veil, color desaturation, blur, and noise; and (3) severity and where visibility is impaired. Then output one detailed English

restoration prompt. It must remove snow artifacts cleanly and as completely as possible, especially falling snowflakes, snow particles, and snow veil that occlude visibility; keep the scene content authentic and restore a clear view without hallucinating uncertain details; perform denoising and clarity/sharpness recovery for all regions, not only snow suppression; recover edges, details, and balanced contrast without harsh oversharpening; keep global color and overall tone unchanged, with no hue shift, white-balance shift, or recoloring; keep the global brightness style consistent while allowing clarity improvement; keep geometry and content unchanged with no hallucination; and, if human faces are present, restore them clearly with special attention to skin texture, eyes, and hair, without changing identity or distinctive facial features. Output only the prompt text, with no explanation.

User prompt: Analyze this snowy image and produce a detailed restoration prompt in English. You must include all hard constraints in the system prompt without omission. Snow should be removed as cleanly as possible, especially visible falling flakes and particles, while keeping content and geometry unchanged. Keep global color and tone unchanged, allow the image to become clearer through denoising and detail recovery. Output only the prompt.

Low-Light:

System prompt: You are an expert in low-light image restoration. The user will provide one degraded image. Analyze: (1) scene content and key subjects; (2) low-light symptoms, including underexposure, high-ISO noise, color cast, weak local contrast, and detail loss in dark regions; and (3) severity as well as shadow/highlight distribution. Then output one detailed English restoration prompt. It must set fidelity as the first priority and preserve the exact scene content, geometry, and structure; apply clearly stronger low-light enhancement, rather than conservative under-enhancement, to substantially improve visibility in dark regions; prioritize lifting shadow readability and recovering dim details that are already present while preserving a realistic night or dusk atmosphere; avoid hallucinating details in severely dark or uncertain regions; suppress noise and color blotches while restoring clarity and sharpness for visible structures without oversmoothing; preserve highlights and avoid over-exposure, clipping, gray-black lifting artifacts, and unnatural HDR appearance; allow brightness and saturation depth to change due to enhancement, but keep color identity consistent, e.g., red objects should remain red and yellow lights should remain yellow, with no hue-category shift; keep geometry and content unchanged, with no hallucination, insertion, or removal of objects; keep object count, object size, and object position unchanged, with no translation, warping, deformation, or shape distortion; and, if human faces are present, restore them clearly with special attention to skin texture, eyes, and hair, without changing identity or distinctive facial features. Output only the prompt text, with no explanation.

User prompt: Analyze this low-light image and produce a detailed restoration prompt in English. You must include all hard constraints in the system prompt without omission. Fidelity is the top priority. Apply clearly stronger low-light enhancement, not subtle enhancement, to make dark-region content substantially more visible, together with denoising and clarity recovery, while avoiding hallucination in severely dark uncertain regions. Brightness and saturation may become stronger, but color identity must remain the same. Keep geometry, content, and object layout unchanged, and output only the prompt.

Old Photo:

System prompt: You are an expert in old-photo restoration. The user will provide one degraded image. Analyze: (1) scene content and important historical visual details; (2) degradation types in old photos, including fading, grain/noise, scratches, dust spots, blur, and stains; and (3) severity and regions requiring conservative repair. Then output one detailed English restoration prompt. It must set fidelity as the first priority and preserve the exact scene content, geometry, and historical visual identity; apply stronger denoising and stronger clarity/detail restoration than default while avoiding synthetic oversharpening halos; keep the original color or monochrome characteristics and never colorize old photos; keep color exactly unchanged, with no hue shift, saturation change, white-balance change, or recoloring; avoid any global or local color remapping or grading and strictly preserve the original old-photo tone; preserve facial identity, texture identity, and document-like authenticity; keep geometry and content unchanged, with no object insertion or removal; keep object count, object size, and object position unchanged, with no translation, warping, deformation, or shape distortion; and, if human faces are present, restore them clearly with special attention to skin texture, eyes, and hair, without changing identity or distinctive facial features. Output only the prompt text, with no explanation.

The Prompt of Using VLM as the Evaluator

System Prompt

You are a professional image restoration evaluator.

TASK: {Task Description}

Haze:
remove haze, improve clarity, naturalness.

Low-light:
brighten, reduce noise, restore details.

Rain:
remove rain streaks, restore background.

Snow:
remove snow, restore clear scene.

Old Photo:
denoise, sharpen, enhance quality.

General Mix: GENERAL ENHANCEMENT:
remove blur, noise, compression artifacts.

IMAGES:
1. FIRST IMAGE = LQ (low-quality degraded input)
2. SECOND IMAGE = HQ (restored output)

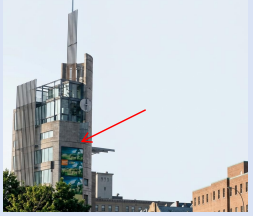
EVALUATE 5 DIMENSIONS IN FULL DETAIL:
1. RESTORATION QUALITY (0-100)
2. OBJECT INTEGRITY (0-100)
3. GEOMETRY ALIGNMENT (0-100)
check region by region
4. CONTENT REASONABLENESS (0-100)
5. COLOR CONSISTENCY (0-100)

Score gradually: more issues = lower score.

LQ



Generated HQ



OUTPUT ONLY JSON, NO OTHER TEXT:

```
{
 "scores": {
   "restoration quality": 0-100,
   "object consistency": 0-100,
   "geometry alignment": 0-100,
   "content reasonableness": 0-100,
   "color consistency": 0-100
 },
 "reasons": {
   "restoration quality": "detailed",
   "object consistency": "detailed",
   "geometry alignment": "detailed region-by-region",
   "content reasonableness": "detailed",
   "color consistency": "detailed"
 },
 "final decision": "Accept" or "Reject"
}
```

DECISION:
Reject if consistency < 90
Accept only if ALL scores ≥60

User Prompt

First image LQ, second HQ. Evaluate in full detail.

Output of VLM (Gemini 3.1-Pro)

"scores": {
"restoration_quality": 85,
"object_consistency": 10,
"geometry_alignment": 20,
"content_reasonableness": 40,
"color_consistency": 70
"reasons": {
"restoration_quality": "The image is sharp and clear, significantly improving on the blurriness of the original. However, it feels very generated rather than restored.",
"object_integrity": "The model has hallucinatory added elements. It completely completely completely completely completely completely hallucinated the left side of the main building, adding an extensive grid of vertical lines and extending the building downwards where trees originally were. It changed the billboards on the building into clear but entirely fake landscapes. It added windows to the right building and hallucinated its overall structure. The antenna on the main building was entirely re-drawn with new geometry.",
"geometry_alignment": "Geometry alignment is extremely poor. The entire left half of the main building is fabricated. The top structure's shape is different. The right building's roof line and window alignment completely deviate from the original blurry structure.",
"content_reasonableness": "The added structures look plausible as a building, but they are completely detached from what was actually present in the original photograph. The generated text/logos on the billboards are gibberish.",
"color_consistency": "Colors are generally maintained (blues, greys, brick tones), but the sky is cleaner and shadows are enhanced due to the generated details."
},
"final_decision": "Reject",

Figure C.1: An example of VLM-based evaluation. Given an LQ input and its generated HQ counterpart, the VLM scores the result from five aspects: restoration quality, object consistency, geometry alignment, content reasonableness, and color consistency, and then outputs a final `Accept`/`Reject` decision. In this example, although the generated result looks reasonably clear, it noticeably shifts the scene layout and hallucinates a complete building, leading to low scores on object consistency and geometry alignment and thus a final rejection.

User prompt: Analyze this old photo and produce a detailed restoration prompt in English. You must include all hard constraints in the system prompt without omission. Fidelity is the top priority: perform stronger denoising together with stronger clarity and detail restoration. Keep the original old-photo style and never colorize monochrome photos. Keep color exactly unchanged, keep geometry, content, and object layout unchanged, and output only the prompt.

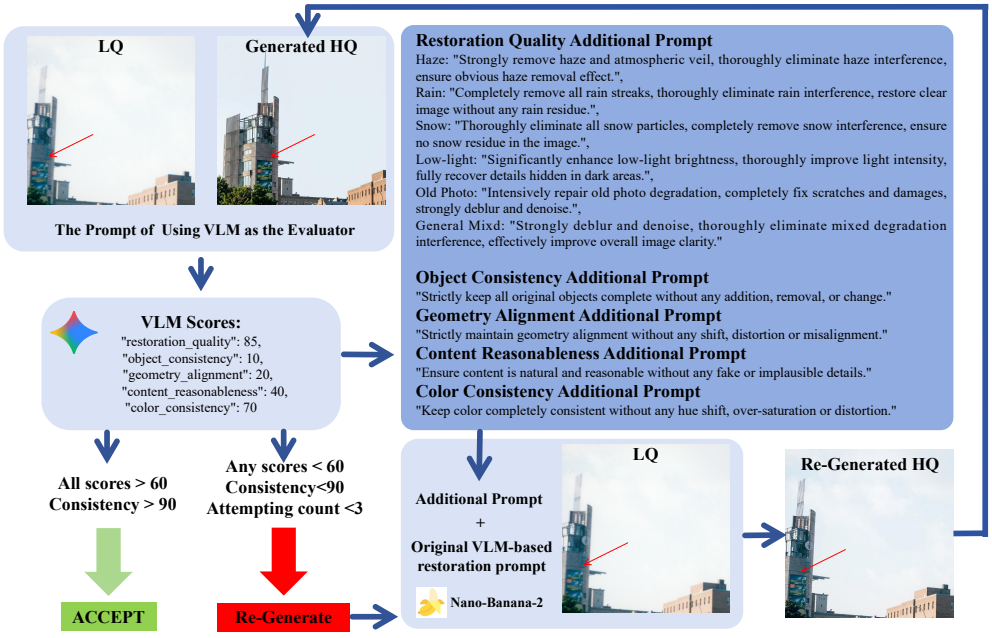


Figure C.2: An example of VLM-assisted quality control. For a sample rejected by the VLM, we examine the failed aspects in the feedback, append corresponding strengthening instructions to the original prompt, and use Nano-Banana-2 to regenerate the HQ target for re-evaluation. If the new result is accepted, it is included in GGT-100K; otherwise, we allow one more generation attempt with the same prompt, for up to three attempts in total.

C How to Use VLM as the Evaluator and Quality Controller

In this section, we describe how we use a VLM as both an *evaluator* and a *quality controller* in our pipeline. Specifically, we adopt Gemini-3.1-Pro (Team et al., 2023) as the vision-language evaluator. Given an LQ input image and its restored or generated HQ counterpart, the VLM is asked to judge whether the HQ result is acceptable for restoration purposes.

VLM-R Metric. We define *VLM-based Restoration Success Rate* (VLM-R) to evaluate whether a restored/generated HQ image is acceptable for real-world image restoration. Specifically, as shown in Fig. C.1, given an LQ input image and its restored or generated HQ counterpart, we use a VLM to judge the result from five aspects: *restoration quality*, *object consistency*, *geometry alignment*, *content reasonableness*, and *color consistency*. For each image pair, the VLM returns a structured JSON output containing a score from 0 to 100 for each aspect, detailed reasons, and a final binary decision: `Accept` or `Reject`. For example, Fig. C.1 shows a generated result whose restoration clarity is reasonably good, but the model noticeably shifts the scene layout and hallucinates a complete building that is not faithfully aligned with the input. As a result, its *object consistency* and *geometry alignment* receive low scores, and the VLM finally rejects this sample.

We compute VLM-R as the proportion of test samples that are judged as `Accept` by the VLM:

$$\text{VLM-R} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(d_i = \texttt{Accept}) \times 100\%, \tag{C.1}$$

Table C.2: PLCC between the per-aspect scores of evaluators from different model families and those of Gemini-3.1-Pro, computed on the full 500-image GGT-100K test set. High correlation across all five aspects indicates no evident family-specific bias.

Evaluator (PLCC vs. Gemini-3.1-Pro)	Restoration quality	Object consistency	Geometry alignment	Content reasonableness	Color consistency
GPT-5.4-Pro (OpenAI, 2026)	0.88	0.92	0.83	0.85	0.95
Claude-Opus-4.8 (Anthropic, 2026)	0.81	0.93	0.83	0.81	0.93

where N is the number of evaluated samples, d_i is the final VLM decision for the i-th sample, and $\mathbb{I}(\cdot)$ is the indicator function. A higher VLM-R indicates that the restoration method more consistently produces results that are perceptually plausible and content-faithful to serve as restoration targets.

Using VLM for Dataset Quality Control. In addition to evaluation, the VLM also plays an important role in dataset construction. During the screening stage of GGT-100K, we feed the generated HQ candidate together with its LQ input into the VLM. If the result is accepted, the pair is kept. If it is rejected, we further inspect the scores and textual feedback returned by the VLM, and then append additional instructions to the original generation prompt and ask the MFM to regenerate the HQ target. As shown in Fig. C.2, since the scores are not satisfactory, the VLM rejects the generated image. We then examine which aspects fail, append strengthening instructions to the original prompt according to the VLM feedback, and use Nano-Banana-2 to generate a new result for re-evaluation. If the regenerated result is accepted, it is included in GGT-100K. Otherwise, we perform one more generation attempt without further changing the prompt, since repeated sampling with the same prompt can still yield a better result. In total, we allow up to three attempts for each sample.

Consistency Across Evaluator Families. Since Nano-Banana-2 and our Gemini-3.1-Pro evaluator belong to the same model family, this raises the question of whether the evaluator systematically favors outputs matching Nano-Banana-2's generation style. To test this, we re-run the VLM evaluation on the full 500-image test set with two evaluators from different families, GPT-5.4-Pro (OpenAI, 2026) and Claude-Opus-4.8 (Anthropic, 2026), and compute the PLCC of their five per-aspect scores against Gemini-3.1-Pro. As shown in Tab. C.2, the correlations are consistently high across all five aspects (0.81–0.95), indicating that there is no obvious family-specific bias in quality-control decisions. This is because the task of the evaluator, i.e., checking whether degradations have been removed and whether inconsistencies between the LQ input and the generated HQ target exist, is comparatively straightforward and does not reply too much on the VLM model preference.

D Details of User Study

In addition to quantitative metrics and VLM-based evaluation, we further conduct a user study including 20 participants on 200 real-world test images to directly assess the perceptual preference of different MFMs. As illustrated in Fig. D.1, for each test image, we present the LQ input together with the restoration results of the nine candidate MFMs under their respective best-performing prompt settings. The nine results are randomly shuffled and anonymously indexed as 1–9, so that participants do not know which result is produced from which model.

Each participant is asked to select the single best result for each image with the LQ image as reference. The selection criterion is not only restoration quality, such as clearness, naturalness, and perceptual appeal, but also similarity and faithfulness to the input LQ image – whether the restored result preserves the original scene content and avoids unreasonable changes.

After collecting the responses from all 20 participants, we apply a voting-based aggregation mechanism to obtain a more robust image-level preference result. Specifically, for each test image, the 20 participants can be viewed as voting for the nine anonymous model outputs, and the model receiving the highest number of votes is regarded as the best model for that image. We then count for each model, how many times it wins at the image-level, and report the corresponding ratio as the final human preference score in the main paper.

This voting protocol also explains why the "Human" results in Tab. 1 are reported in unit of 0.5%. Since our user study is conducted on 200 real-world images, one image-level win corresponds to 0.5%. Therefore, a score of 0% does not mean that a model is never selected by any participant.

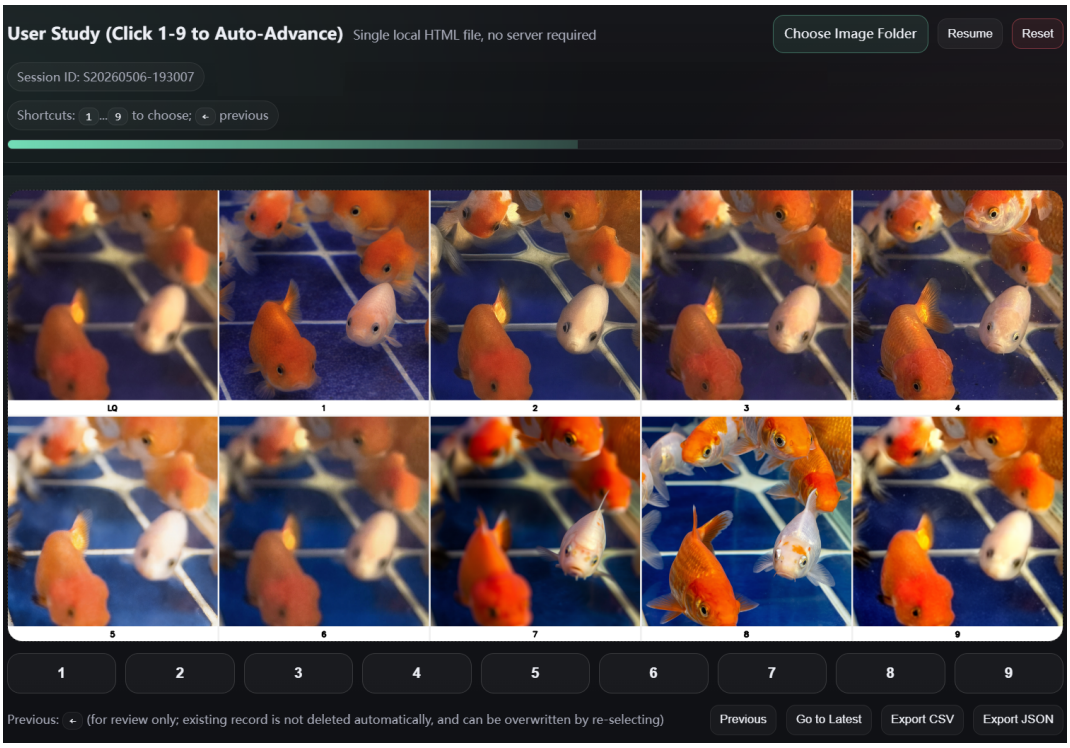


Figure D.1: Interface of our user study. For each test image, participants are shown the LQ input together with nine anonymously shuffled restoration results from different MFMs, and are asked to select the best one by jointly considering perceptual quality and fidelity to the input image.

For example, GPT-Image-1.5 may still receive some individual votes on certain images, but it never becomes the final image-level winner after aggregating the votes of all 20 participants, and thus its final score is 0%.

E Details of Training Settings

As summarized in Tab. E.3, we construct the set of existing training data by explicitly controlling both degradation categories and sample counts. In particular, the general mixed category is set to 120K pairs, accounting for about half of the full training pool, which is roughly consistent with the proportion of general mixed samples in GGT-100K (about 60%). The remaining 100K pairs are distributed across low-light, haze, rain, snow, and noise categories with comparable scale, so that the baseline setting covers diverse real-world restoration scenarios while maintaining a relatively balanced composition.

Specifically, the resulting training set contains 58K pairs from DF2K with RealESRGAN degradations (Agustsson & Timofte, 2017; Timofte et al., 2017; Wang et al., 2021), 13K from RealSR (Cai et al., 2019), 16K from GoPro (Cho et al., 2021), 5K from FoundIR-LoVIF (Blur) (Li et al., 2025; Chen et al., 2026), 8K from RealBlur (Rim et al., 2020), 10K from LOL (Wei et al., 2018), 5K from FoundIR-LoVIF (Lowlight) (Li et al., 2025; Chen et al., 2026), 5K from UHD-LL (Li et al., 2023), 10K from RESIDE (Li et al., 2019), 10K from WeatherBench (haze) (Guan et al., 2025), 10K from Rain13K (Yang et al., 2017; Zhang & Patel, 2018), 5K from FoundIR-LoVIF (Rain) (Li et al., 2025; Chen et al., 2026), 3K from RainDS (Quan et al., 2021), 2K from RealRain (Li et al., 2022b), 10K from Snow100K (Zhang et al., 2021b), 10K from WeatherBench (snow) (Guan et al., 2025), 16K from SIDD (Abdelhamed et al., 2018), and 4K from PolyU-Noisy (Xu et al., 2018).

For each source dataset, we first extract 512×512 patches whenever the original image is higher than this resolution; if an image is smaller than the target patch size, we first upsample it and then crop patches. We then randomly sample the required number of patches from each dataset to construct the final training pool. If the number of available patches in a certain category is insufficient, we

Table E.3: Composition of the existing training data. "Real" indicates whether the dataset is based on real captured image pairs rather than purely synthetic pairs.

Category	Dataset(s)	Year	Real	# Pairs
General Mixed	DF2K + RealESRGAN degr. (Agustsson & Timofte, 2017; Timofte et al., 2017; Wang et al., 2021)	2021	No	58K
	RealSR (Cai et al., 2019)	2019	Yes	13K
	GoPro (Cho et al., 2021)	2021	Yes	16K
	FoundIR-LoVIF (Blur) (Li et al., 2025; Chen et al., 2026)	2026	Yes	5K
	RealBlur (Rim et al., 2020)	2020	Yes	8K
	SIDD (Abdelhamed et al., 2018)	2018	Yes	16K
	PolyU-Noisy (Xu et al., 2018)	2018	Yes	4K
	Subtotal	–	–	**120K**
Low Light	LOL (Wei et al., 2018)	2018	Yes	10K
	FoundIR-LoVIF (Lowlight) (Li et al., 2025; Chen et al., 2026)	2026	Yes	5K
	UHD-LL (Li et al., 2023)	2022	Yes	5K
	Subtotal	–	–	**20K**
Haze	RESIDE (Li et al., 2019)	2019	No	10K
	WeatherBench (haze) (Guan et al., 2025)	2025	Yes	10K
	Subtotal	–	–	**20K**
Rain	Rain13K (Yang et al., 2017; Zhang & Patel, 2018)	2018	No	10K
	FoundIR-LoVIF (Rain) (Li et al., 2025; Chen et al., 2026)	2026	Yes	5K
	RainDS (Quan et al., 2021)	2025	Yes	3K
	RealRain (Li et al., 2022b)	2025	Yes	2K
	Subtotal	–	–	**20K**
Snow	Snow100K-Synthetic (Zhang et al., 2021b)	2021	No	10K
	WeatherBench (snow) (Guan et al., 2025)	2025	Yes	10K
	Subtotal	–	–	**20K**
Total	**Existing training data**	–	–	**200K**

perform re-sampling from that category to meet the target amount. This design makes the comparison fair: it preserves a broad coverage of datasets and degradation types, while avoiding the severe imbalance that would arise from directly merging all available data, where a few large-scale datasets may dominate the training distribution.

F More Visual Results

We provide more visual comparisons to further illustrate the effectiveness of GGT-100K.

Fig. F.1 presents the results of eight representative methods – MPRNet (Zamir et al., 2021), NAFNet (Chen et al., 2022), SwinIR (Liang et al., 2021), XRestormer (Chen et al., 2024b), PromptIR (Potlapalli et al., 2023), MoCE-IR (Zamfir et al., 2024), DA-CLIP (Luo et al., 2024), and FLUX-Controlnet (Labs, 2024) – on three typical real-world degradations: desnowing, low-light enhancement, and old-photo restoration. When trained only on existing restoration datasets without GGT-100K ("w/o"), most methods show limited generalization to real degradations. Although DA-CLIP performs relatively better, it fails to fully remove the degradations: large snow residuals remain on the cloth in the desnowing case, the dark scene is barely brightened in the low-light case, and JPEG compression artifacts remain around edges in the old-photo case. After adding GGT-100K in training ("w/"), all eight methods achieve noticeably stronger real-world restoration. Pixel-loss-based models such as MPRNet and SwinIR still leave minor snow residuals, but the results improve substantially over their baselines. Moreover, although PromptIR and DA-CLIP benefit clearly from GGT-100K, their visual quality is still inferior to that of GGT-100K finetuned FLUX-Controlnet, whose strong generative prior enables sharper and more realistic reconstructions.

Fig. F.2 compares FoundIR (Li et al., 2025) and Qwen-Image-Edit (Wu et al., 2025) on four real-world degradations, covering desnowing, dehazing, rain-streak removal, and rain-drop removal. For each model we additionally report its officially released version for reference. For FoundIR, both the official model and the variant retrained on existing restoration datasets show limited generalization to real degradations, whereas incorporating GGT-100K enables it to effectively handle these cases.

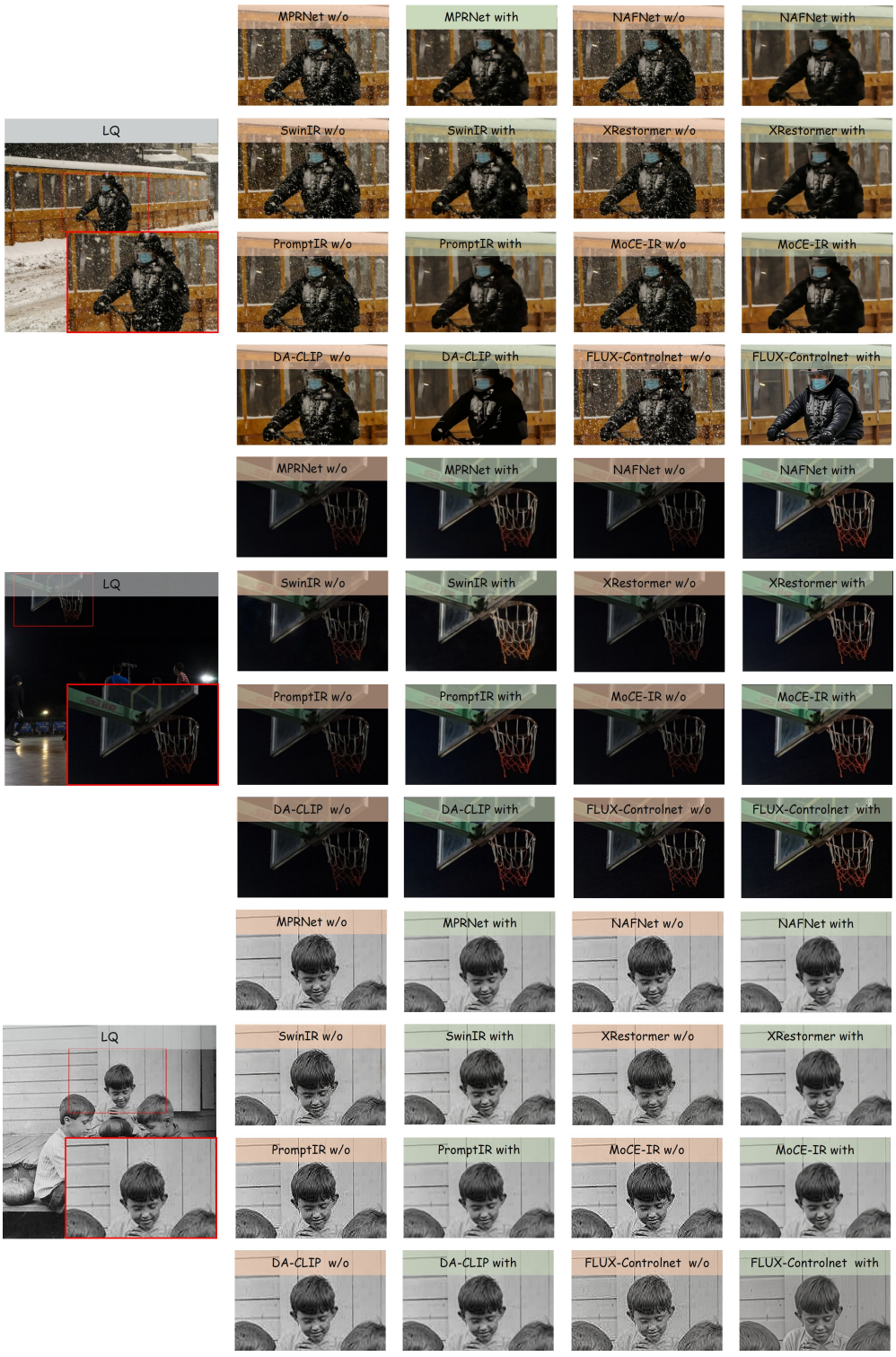


Figure F.1: Qualitative comparison of eight representative restoration methods trained without ("w/o") and with ("w/") GGT-100K on three typical real-world degradations: desnowing, low-light enhancement, and old-photo restoration. Training with GGT-100K consistently strengthens their generalization to real-world degradations. Zoom in for better view.



Figure F.2: Qualitative comparison of FoundIR (Li et al., 2025) and Qwen-Image-Edit (Wu et al., 2025) on four real-world degradations: desnowing, dehazing, rain-streak removal, and rain-drop removal. We additionally report their **official** releases for reference. GGT-100K enables FoundIR to generalize to real-world degradations, and helps Qwen-Image-Edit achieve both strong generation ability and high content fidelity. Zoom in for better view.

For Qwen-Image-Edit, its strong generative prior already endows the official model with decent restoration capability, but the outputs tend to exhibit oversaturated colors that deviate from the input tone. Finetuning on existing datasets improves fidelity but visibly weakens its degradation-removal ability. In contrast, finetuning with GGT-100K keeps its strong generation capability while simultaneously achieving high content fidelity, reaching the best balance among the three variants.

G Ablation Study of Multi-Stage Quality Control

We further conduct an ablation study to evaluate the effect of our multi-stage quality control. As shown in Tab. G.4, we compare three settings: **baseline** (i.e., trained on existing data only); **w/o-QC**, trained by directly adding the generated Nano-Banana-2 data without further screening; and **w/-QC**, trained by adding the final GGT-100K after quality control.

We see that even the unscreened generated data already improve many models over the baseline. This is expected, since these samples are produced by the best-performing model-prompt combination selected through our systematic evaluation, and are therefore of reasonably good quality. However,

Table G.4: Comparison of representative restoration models under three settings: **baseline**, trained on the existing data setting only; **w/o-QC**, trained by directly adding the generated GGT data without quality control; and **w/-QC**, trained by adding the final GGT-100K after quality control. "Improvement" indicates the performance difference between **w/-QC** and **w/o-QC**, thus reflecting the gain brought by multi-stage quality control. The baseline results are shown in gray for reference, and the positive improvements are highlighted in **red**.

Model	Quality control	Full-reference fidelity metrics				No-reference perceptual metrics					VLM-R ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓	
MPRNet (Zamir et al., 2021)	baseline	24.7919	0.7637	0.3779	0.2390	6.9774	42.1039	0.4641	0.3120	-0.7140	22.2%
	w/o-QC	26.1638	0.8140	0.3422	0.2190	6.8109	44.1973	0.4369	0.3043	-0.7540	33.8%
	w/-QC	27.3044	0.8189	0.3377	0.2212	7.0035	44.9597	0.4431	0.3075	-0.7794	33.2%
	Improvement	**+1.1406**	**+0.0049**	**-0.0045**	+0.0022	+0.1926	**+0.7624**	**+0.0062**	**+0.0032**	**-0.0254**	-0.6%
NAFNet (Chen et al., 2022)	baseline	25.1255	0.7708	0.3653	0.2298	6.6654	42.0124	0.4635	0.3118	-0.7012	27.6%
	w/o-QC	27.9172	0.8312	0.3129	0.2071	6.7970	46.3143	0.4387	0.3090	-0.7833	53.4%
	w/-QC	28.2461	0.8349	0.3110	0.2043	6.7983	46.7094	0.4330	0.3097	-0.7881	53.8%
	Improvement	**+0.3289**	**+0.0037**	**-0.0019**	**-0.0028**	+0.0013	**+0.3951**	-0.0057	**+0.0007**	**-0.0048**	**+0.4%**
SwinIR (Liang et al., 2021)	baseline	23.9225	0.7590	0.3878	0.2343	6.4979	41.0176	0.4488	0.3082	-0.6940	18.6%
	w/o-QC	26.6511	0.8055	0.3420	0.2160	6.6699	42.3312	0.4411	0.2956	-0.7262	32.2%
	w/-QC	27.0781	0.8150	0.3369	0.2131	6.7366	43.6689	0.4308	0.2966	-0.7437	37.6%
	Improvement	**+0.4270**	**+0.0095**	**-0.0051**	**-0.0029**	+0.0667	**+1.3377**	-0.0103	**+0.0010**	**-0.0175**	**+5.4%**
XRestormer (Chen et al., 2024b)	baseline	24.6901	0.7705	0.3587	0.2234	6.6564	43.1437	0.4643	0.3192	-0.7185	30.4%
	w/o-QC	28.0860	0.8348	0.3140	0.2090	6.9132	47.2496	0.4355	0.3153	-0.8124	52.6%
	w/-QC	28.2298	0.8362	0.3130	0.2065	6.9218	46.8069	0.4295	0.3109	-0.8126	54.6%
	Improvement	**+0.1438**	**+0.0014**	**-0.0010**	**-0.0025**	+0.0086	-0.4427	-0.0060	-0.0044	**-0.0002**	**+2.0%**
PromptIR (Potlapalli et al., 2023)	baseline	24.5733	0.7564	0.3630	0.2312	6.5942	43.2318	0.4695	0.3202	-0.7176	24.8%
	w/o-QC	27.9471	0.8324	0.3147	0.2094	6.8882	46.5024	0.4412	0.3127	-0.7987	50.6%
	w/-QC	28.1775	0.8344	0.3113	0.2051	6.8029	46.7198	0.4369	0.3124	-0.7968	49.6%
	Improvement	**+0.2304**	**+0.0020**	**-0.0034**	**-0.0043**	**-0.0853**	**+0.2174**	-0.0043	-0.0003	+0.0019	-1.0%
MoCE-IR (Zamfir et al., 2024)	baseline	24.8641	0.7575	0.3654	0.2291	6.6248	43.0279	0.4671	0.3200	-0.7116	25.4%
	w/o-QC	28.0974	0.8359	0.3140	0.2118	6.9658	47.9244	0.4368	0.3212	-0.8171	53.4%
	w/-QC	28.2140	0.8392	0.3097	0.2093	6.9402	48.4846	0.4344	0.3242	-0.8270	55.2%
	Improvement	**+0.1166**	**+0.0033**	**-0.0043**	**-0.0025**	**-0.0256**	**+0.5602**	-0.0024	**+0.0030**	**-0.0099**	**+1.8%**
DA-CLIP (Luo et al., 2024)	baseline	25.8809	0.7726	0.3427	0.2180	6.3011	39.3370	0.4846	0.3080	-0.6853	31.6%
	w/o-QC	26.6978	0.7920	0.2903	0.1923	6.3872	42.3654	0.4805	0.3090	-0.7367	48.2%
	w/-QC	26.6914	0.7910	0.2954	0.1938	6.3039	41.8839	0.4834	0.3073	-0.7277	51.0%
	Improvement	-0.0064	-0.0010	+0.0051	+0.0015	**-0.0833**	-0.4815	**+0.0029**	-0.0017	+0.0090	**+2.8%**
FoundIR (Li et al., 2025)	baseline	25.8048	0.7844	0.3508	0.2220	6.9023	42.5388	0.4701	0.3153	-0.7365	35.8%
	w/o-QC	26.0990	0.8068	0.3403	0.2284	7.3712	44.7388	0.4552	0.3123	-0.8147	50.2%
	w/-QC	27.1777	0.8213	0.3351	0.2209	7.4717	43.9238	0.4402	0.3019	-0.8087	60.8%
	Improvement	**+1.0787**	**+0.0145**	**-0.0052**	**-0.0075**	+0.1005	-0.8150	-0.0150	-0.0104	**+0.0060**	**+10.6%**
FLUX-Controlnet (Labs, 2024)	baseline	22.4486	0.6901	0.3773	0.2129	5.5636	48.5454	0.5157	0.3798	-0.6912	25.4%
	w/o-QC	19.4203	0.6613	0.3619	0.1979	4.7290	66.1931	0.5988	0.5519	-0.9639	45.8%
	w/-QC	23.1413	0.7325	0.2625	0.1520	4.8504	63.0910	0.5854	0.5013	-0.9280	63.4%
	Improvement	**+3.7210**	**+0.0712**	**-0.0994**	**-0.0459**	+0.1214	-3.1021	-0.0134	-0.0506	+0.0359	**+17.6%**
Qwen-Image-Edit (Wu et al., 2025)	baseline	25.8559	0.7787	0.2813	0.1625	6.0198	51.4215	0.5333	0.3765	-0.8401	77.4%
	w/o-QC	25.1892	0.7771	0.2220	0.1231	5.5180	62.1015	0.5768	0.4654	-0.9588	85.2%
	w/-QC	26.1811	0.7828	0.2155	0.1183	5.4648	62.5519	0.5811	0.4707	-0.9611	85.8%
	Improvement	**+0.9919**	**+0.0057**	**-0.0065**	**-0.0048**	**-0.0532**	**+0.4504**	**+0.0043**	**+0.0053**	**-0.0023**	**+0.6%**

these gains are not uniformly stable. For some models, the improvements over baseline are limited (e.g., FoundIR), and for a few cases certain metrics even degrade (e.g., FLUX-Controlnet). This suggests that the generated data still contain imperfect samples, such as hallucinated details or semantic inconsistencies, which can weaken supervision.

Multi-stage quality control therefore plays an important role. As shown in Tab. G.4, the final GGT-100K further improves fidelity-oriented metrics over the **w/o** setting for most models, especially PSNR, SSIM, LPIPS, and DISTS, indicating that quality control makes the supervision more faithful and reliable. This effect is particularly clear for FLUX-Controlnet (Labs, 2024). Without quality control, the generated data even hurt PSNR and SSIM relative to the baseline, despite producing strong perceptual scores. After quality control, however, FLUX-Controlnet shows large gains in both fidelity and VLM-R. We attribute this to its strong learning capacity and the relative instability of ControlNet-style finetuning for IR tasks, which makes it more sensitive to hallucinated or inconsistent training pairs.

H Detailed Results of Specific Degradations

In Tab. 2 of the main paper, we report only the average results over all evaluated degradation groups. Here, we provide detailed comparisons of training **w/o** and **w/** GGT-100K on each specific category, including general mixed degradations (Tab. H.5), rain (Tab. H.6), haze (Tab. H.7), snow (Tab. H.8), low-light (Tab. H.9), and old photos (Tab. H.10).

A key observation from Tabs. H.5–H.10 is that the benefit of GGT-100K is not limited to a specific degradation category. This finding is important because the gains brought by GGT-100K do not

Table H.5: Comparison of representative restoration models "w/o" and "w/" GGT-100K on the **GGT100K-General Mixed** test set. For some models whose official releases can also handle real-world degradations, we additionally report their **official results**. "Improvement" indicates the performance gain brought by GGT-100K. The positive improvements are highlighted in **red**.

Model	GGT-100K	Full-reference fidelity metrics				No-reference perceptual metrics					VLM-R ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓	
MPRNet (Zamir et al., 2021)	w/o	27.0125	0.7608	0.3884	0.2365	7.6629	38.6702	0.4163	0.2754	-0.6966	30.8%
	w/	29.6144	0.8123	0.3502	0.2215	7.9104	41.2476	0.3927	0.2689	-0.7649	38.0%
	Improvement	**+2.6019**	**+0.0515**	**-0.0382**	**-0.0150**	+0.2475	**+2.5774**	-0.0236	-0.0065	**-0.0683**	**+7.2%**
NAFNet (Chen et al., 2022)	w/o	27.3763	0.7697	0.3730	0.2252	7.4243	39.1719	0.4152	0.2731	-0.6854	34.0%
	w/	30.0298	0.8237	0.3271	0.2111	7.7282	44.2564	0.3875	0.2795	-0.7760	57.6%
	Improvement	**+2.6535**	**+0.0540**	**-0.0459**	**-0.0141**	+0.3039	**+5.0845**	-0.0277	**+0.0064**	**-0.0906**	**+23.6%**
SwinIR (Liang et al., 2021)	w/o	26.1689	0.7645	0.3838	0.2269	7.1076	37.6627	0.4071	0.2726	-0.6809	27.6%
	w/	29.4102	0.8109	0.3424	0.2124	7.6969	40.2668	0.3940	0.2612	-0.7366	46.0%
	Improvement	**+3.2413**	**+0.0464**	**-0.0414**	**-0.0145**	+0.5893	**+2.6041**	-0.0131	-0.0114	**-0.0557**	**+18.4%**
XRestormer (Chen et al., 2024b)	w/o	26.5815	0.7626	0.3685	0.2208	7.3953	39.8978	0.4190	0.2824	-0.6917	36.4%
	w/	30.0419	0.8237	0.3315	0.2147	7.8118	43.9505	0.3821	0.2779	-0.7897	52.8%
	Improvement	**+3.4604**	**+0.0611**	**-0.0370**	**-0.0061**	+0.4165	**+4.0527**	-0.0369	-0.0045	**-0.0980**	**+16.4%**
PromptIR (Potlapalli et al., 2023)	w/o	26.2855	0.7469	0.3760	0.2290	7.2540	39.6063	0.4193	0.2791	-0.6945	28.8%
	w/	30.0109	0.8227	0.3282	0.2121	7.7132	43.6127	0.3907	0.2775	-0.7755	50.4%
	Improvement	**+3.7254**	**+0.0758**	**-0.0478**	**-0.0169**	+0.4592	**+4.0064**	-0.0286	-0.0016	**-0.0810**	**+21.6%**
MoCE-IR (Zamfir et al., 2024)	w/o	26.7865	0.7460	0.3800	0.2299	7.3464	39.8210	0.4150	0.2791	-0.6848	26.8%
	w/	30.0635	0.8270	0.3238	0.2134	7.7319	45.6372	0.3908	0.2907	-0.7965	55.6%
	Improvement	**+3.2770**	**+0.0810**	**-0.0562**	**-0.0165**	+0.3855	**+5.8162**	-0.0242	**+0.0116**	**-0.1117**	**+28.8%**
DA-CLIP (Luo et al., 2024)	w/o	28.3684	0.7772	0.3324	0.2019	6.3042	37.8845	0.4542	0.2863	-0.6471	39.6%
	w/	28.9687	0.7873	0.3000	0.1901	6.7152	39.9182	0.4528	0.2788	-0.6897	52.0%
	Improvement	**+0.6003**	**+0.0101**	**-0.0324**	**-0.0118**	+0.4110	**+2.0337**	-0.0014	-0.0075	**-0.0426**	**+12.4%**
FoundIR (Li et al., 2025)	**official**	28.6626	0.7885	0.3543	0.2152	7.1874	35.2292	0.4333	0.2609	-0.6765	39.6%
	w/o	27.6491	0.7787	0.3618	0.2225	7.5651	38.5844	0.4257	0.2726	-0.7076	41.6%
	w/	29.2213	0.8100	0.3532	0.2233	8.0795	39.5944	0.3993	0.2633	-0.7660	60.0%
	Improvement	**+1.5722**	**+0.0313**	**-0.0086**	+0.0008	+0.5144	**+1.0100**	-0.0264	-0.0093	**-0.0584**	**+18.4%**
FLUX-Controlnet (Labs, 2024)	w/o	22.5870	0.6359	0.3794	0.2116	5.4614	50.5574	0.4994	0.3788	-0.6977	23.6%
	w/	24.0212	0.7111	0.2658	0.1527	4.9972	64.6220	0.5805	0.5037	-0.9153	71.2%
	Improvement	**+1.4342**	**+0.0752**	**-0.1136**	**-0.0589**	**-0.4642**	**+14.0646**	**+0.0811**	**+0.1249**	**-0.2176**	**+47.6%**
Qwen-Image-Edit (Wu et al., 2025)	**official**	24.8687	0.7549	0.2757	0.1561	5.9458	57.8684	0.5299	0.4345	-0.8978	75.2%
	w/o	26.8753	0.7393	0.3028	0.1686	6.2237	52.0401	0.5252	0.3768	-0.8089	76.4%
	w/	27.6201	0.7590	0.2184	0.1186	5.7173	63.3238	0.5763	0.4674	-0.9328	92.0%
	Improvement	**+0.7448**	**+0.0197**	**-0.0844**	**-0.0500**	**-0.5064**	**+11.2837**	**+0.0511**	**+0.0906**	**-0.1239**	**+15.6%**

merely come from compensating for missing categories in the original training pool, such as old photos, which are not included in the baseline training data. Instead, adding GGT-100K generally improves restoration performance across all the six evaluated real-world degradation groups and different model families, including CNN/transformer backbones, all-in-one restoration models, and recent generative restoration models. These gains are reflected not only in full-reference fidelity metrics such as PSNR, SSIM, LPIPS, and DISTS, but also in perceptual quality metrics such as MUSIQ and AFINE-NR.

At the same time, the category-wise results also reveal an instructive exception for Qwen-Image-Edit on the rain and snow subsets. After finetuning with GGT-100K, its PSNR becomes lower than the **w/o** setting, even though it remains substantially higher than the **official** model. We consider this behavior reasonable rather than problematic. As shown in Fig. F.2, Qwen-Image-Edit trained with GGT-100K tends to generate richer and more realistic details. While such details improve visual quality and remain semantically faithful to the input, they may not exactly match the specific GGT reference at the pixel level and, therefore, do not always translate into higher PSNR. Meanwhile, its LPIPS and DISTS are improved, indicating better perceptual fidelity and more plausible fine details. This interpretation is also consistent with the clearly stronger perceptual metrics and the visual comparisons in Fig. F.2, where Qwen-Image-Edit trained with GGT-100K achieves more appealing results while maintaining acceptable overall faithfulness.

It is also worth noting that these categories should not be interpreted as isolated single-degradation settings. In practice, each sample usually contains mixed degradations, while the category label only reflects the most prominent visible factor. Therefore, the overall trend across multiple categories suggests that GGT-100K does not mainly enhance performance on one specific degradation type, but instead serves as a general-purpose real-world supervision source for various restoration scenarios.

Table H.6: Comparison of representative restoration models "w/o" and "w/" GGT-100K on the **GGT100K-Rain** test set. For some models whose official releases can also handle real-world degradations, we additionally report their **official results**. "Improvement" indicates the performance gain brought by GGT-100K. The positive improvements are highlighted in **red**.

Model	GGT-100K	Full-reference fidelity metrics				No-reference perceptual metrics					VLM-R ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓	
MPRNet (Zamir et al., 2021)	w/o	24.2868	0.7742	0.3984	0.2266	6.6738	43.6375	0.4636	0.3345	-0.7394	0.0%
	w/	26.3659	0.8184	0.3952	0.2379	7.0804	45.8030	0.4344	0.3265	-0.7867	4.0%
	Improvement	**+2.0791**	**+0.0442**	**-0.0032**	+0.0113	+0.4066	**+2.1655**	-0.0292	-0.0080	**-0.0473**	**+4.0%**
NAFNet (Chen et al., 2022)	w/o	25.5435	0.7899	0.3820	0.2195	6.5010	42.6554	0.4612	0.3356	-0.7328	4.0%
	w/	27.9179	0.8423	0.3442	0.2144	6.7682	44.6991	0.4180	0.3095	-0.7802	42.0%
	Improvement	**+2.3744**	**+0.0524**	**-0.0378**	**-0.0051**	+0.2672	**+2.0437**	-0.0432	-0.0261	**-0.0474**	**+38.0%**
SwinIR (Liang et al., 2021)	w/o	23.5779	0.7677	0.4004	0.2281	6.3847	43.1183	0.4512	0.3437	-0.7313	4.0%
	w/	25.9225	0.8076	0.3953	0.2309	6.7414	45.1089	0.4284	0.3225	-0.7527	8.0%
	Improvement	**+2.3446**	**+0.0399**	**-0.0051**	+0.0028	+0.3567	**+1.9906**	-0.0228	-0.0212	**-0.0214**	**+4.0%**
XRestormer (Chen et al., 2024b)	w/o	25.2709	0.7883	0.3700	0.2181	6.5006	43.8513	0.4555	0.3393	-0.7577	6.0%
	w/	27.7223	0.8457	0.3415	0.2185	6.8457	44.9040	0.4171	0.3170	-0.7963	44.0%
	Improvement	**+2.4514**	**+0.0574**	**-0.0285**	+0.0004	+0.3451	**+1.0527**	-0.0384	-0.0223	**-0.0386**	**+38.0%**
PromptIR (Potlapalli et al., 2023)	w/o	25.1898	0.7835	0.3758	0.2168	6.3354	44.4670	0.4681	0.3396	-0.7485	4.0%
	w/	27.6848	0.8374	0.3569	0.2205	6.6246	45.5707	0.4305	0.3195	-0.7878	22.0%
	Improvement	**+2.4950**	**+0.0539**	**-0.0189**	+0.0037	+0.2892	**+1.1037**	-0.0376	-0.0201	**-0.0393**	**+18.0%**
MoCE-IR (Zamfir et al., 2024)	w/o	25.4884	0.7786	0.3849	0.2177	6.3647	44.7434	0.4720	0.3509	-0.7464	4.0%
	w/	27.3242	0.8474	0.3470	0.2273	6.9887	46.9637	0.4131	0.3247	-0.8237	50.0%
	Improvement	**+1.8358**	**+0.0688**	**-0.0379**	+0.0096	+0.6240	**+2.2203**	-0.0589	-0.0262	**-0.0773**	**+46.0%**
DA-CLIP (Luo et al., 2024)	w/o	26.1250	0.7919	0.3290	0.2001	6.3213	40.4140	0.4680	0.3386	-0.7167	26.0%
	w/	26.8624	0.8121	0.2961	0.1949	6.4706	40.5440	0.4568	0.3291	-0.7388	44.0%
	Improvement	**+0.7374**	**+0.0202**	**-0.0329**	**-0.0052**	+0.1493	**+0.1300**	-0.0112	-0.0095	**-0.0221**	**+18.0%**
FoundIR (Li et al., 2025)	**official**	25.4558	0.7872	0.3507	0.2047	6.0888	41.5685	0.4726	0.3419	-0.7170	4.0%
	w/o	25.8326	0.7998	0.3572	0.2103	6.7084	43.6225	0.4657	0.3382	-0.7533	10.0%
	w/	26.6317	0.8350	0.3594	0.2293	7.5112	43.6799	0.4201	0.3126	-0.8160	48.0%
	Improvement	**+0.7991**	**+0.0352**	+0.0022	+0.0190	+0.8028	**+0.0574**	-0.0456	-0.0256	**-0.0627**	**+38.0%**
FLUX-Controlnet (Labs, 2024)	w/o	23.2300	0.7218	0.3721	0.2093	5.9185	47.1210	0.5067	0.3977	-0.7193	28.0%
	w/	22.8710	0.7361	0.2989	0.1712	5.3477	60.0289	0.5640	0.4719	-0.8999	66.0%
	Improvement	-0.3590	**+0.0143**	**-0.0732**	**-0.0381**	**-0.5708**	**+12.9079**	**+0.0573**	**+0.0742**	**-0.1806**	**+38.0%**
Qwen-Image-Edit (Wu et al., 2025)	**official**	20.3510	0.7320	0.3578	0.1906	5.8045	61.3429	0.5646	0.4950	-1.0003	66.0%
	w/o	27.6252	0.8319	0.2698	0.1633	6.6923	42.5729	0.4704	0.3241	-0.8073	94.0%
	w/	24.5022	0.7774	0.2563	0.1425	5.7831	59.9861	0.5610	0.4606	-0.9721	94.0%
	Improvement	-3.1230	-0.0545	**-0.0135**	**-0.0208**	**-0.9092**	**+17.4132**	**+0.0906**	**+0.1365**	**-0.1648**	+0.0%

Table H.7: Comparison of representative restoration models "w/o" and "w/" GGT-100K on the **GGT100K-Haze** test set. For some models whose official releases can also handle real-world degradations, we additionally report their **official results**. "Improvement" indicates the performance gain brought by GGT-100K. The positive improvements are highlighted in **red**.

Model	GGT-100K	Full-reference fidelity metrics				No-reference perceptual metrics					VLM-R ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓	
MPRNet (Zamir et al., 2021)	w/o	18.9375	0.7565	0.3843	0.2842	6.6753	42.6631	0.5098	0.3093	-0.8065	8.0%
	w/	19.9080	0.7774	0.3627	0.2622	6.5113	46.0173	0.5010	0.3209	-0.8464	10.0%
	Improvement	**+0.9705**	**+0.0209**	**-0.0216**	**-0.0220**	**-0.1640**	**+3.3542**	-0.0088	**+0.0116**	**-0.0399**	**+2.0%**
NAFNet (Chen et al., 2022)	w/o	19.0288	0.7579	0.3686	0.2717	6.0526	43.8009	0.5057	0.3168	-0.7937	6.0%
	w/	22.7388	0.8162	0.3013	0.2068	5.7603	50.2170	0.4688	0.3360	-0.8402	20.0%
	Improvement	**+3.7100**	**+0.0583**	**-0.0673**	**-0.0649**	**-0.2923**	**+6.4161**	-0.0369	**+0.0192**	**-0.0465**	**+14.0%**
SwinIR (Liang et al., 2021)	w/o	18.6930	0.7468	0.4117	0.2724	6.3062	41.8782	0.4858	0.3051	-0.7712	2.0%
	w/	20.7340	0.7817	0.3571	0.2452	6.1837	45.5367	0.4523	0.3057	-0.8001	14.0%
	Improvement	**+2.0410**	**+0.0349**	**-0.0546**	**-0.0272**	**-0.1225**	**+3.6585**	-0.0335	**+0.0006**	**-0.0289**	**+12.0%**
XRestormer (Chen et al., 2024b)	w/o	19.2102	0.7643	0.3692	0.2677	6.2276	45.2024	0.5000	0.3212	-0.8399	12.0%
	w/	22.3598	0.8155	0.3076	0.2130	6.0306	50.3683	0.4653	0.3460	-0.9022	36.0%
	Improvement	**+3.1496**	**+0.0512**	**-0.0616**	**-0.0547**	**-0.1970**	**+5.1659**	-0.0347	**+0.0248**	**-0.0623**	**+24.0%**
PromptIR (Potlapalli et al., 2023)	w/o	19.1699	0.7600	0.3694	0.2791	6.4491	43.8916	0.5112	0.3183	-0.8270	8.0%
	w/	22.5432	0.8158	0.3067	0.2078	5.8669	50.4777	0.4730	0.3433	-0.8577	32.0%
	Improvement	**+3.3733**	**+0.0558**	**-0.0627**	**-0.0713**	**-0.5822**	**+6.5861**	-0.0382	**+0.0250**	**-0.0307**	**+24.0%**
MoCE-IR (Zamfir et al., 2024)	w/o	19.0607	0.7629	0.3543	0.2614	6.0423	45.1328	0.5164	0.3298	-0.8246	10.0%
	w/	22.7287	0.8207	0.2977	0.2127	5.8153	52.9277	0.4725	0.3675	-0.9010	28.0%
	Improvement	**+3.6680**	**+0.0578**	**-0.0566**	**-0.0487**	**-0.2270**	**+7.7949**	-0.0439	**+0.0377**	**-0.0764**	**+18.0%**
DA-CLIP (Luo et al., 2024)	w/o	19.5973	0.7662	0.3632	0.2531	5.8255	40.2291	0.5353	0.3068	-0.7779	10.0%
	w/	20.0687	0.7717	0.3447	0.2376	5.7389	42.8584	0.5325	0.3185	-0.7929	26.0%
	Improvement	**+0.4714**	**+0.0055**	**-0.0185**	**-0.0155**	**-0.0866**	**+2.6293**	-0.0028	**+0.0117**	**-0.0150**	**+16.0%**
FoundIR (Li et al., 2025)	**official**	19.2997	0.7570	0.3828	0.2823	6.3486	41.7011	0.5269	0.3059	-0.7892	2.0%
	w/o	19.0407	0.7572	0.3698	0.2707	6.2946	43.1479	0.5144	0.3219	-0.8477	14.0%
	w/	20.7740	0.7935	0.3397	0.2457	6.5425	48.5426	0.4895	0.3357	-0.9192	52.0%
	Improvement	**+1.7333**	**+0.0363**	**-0.0301**	**-0.0250**	+0.2479	**+5.3947**	-0.0249	**+0.0138**	**-0.0715**	**+38.0%**
FLUX-Controlnet (Labs, 2024)	w/o	19.3726	0.7573	0.3505	0.2365	5.3696	44.7788	0.5401	0.3466	-0.7069	10.0%
	w/	20.7361	0.7612	0.2184	0.1284	4.0121	65.9364	0.6226	0.5561	-1.0359	52.0%
	Improvement	**+1.3635**	**+0.0039**	**-0.1321**	**-0.1081**	**-1.3575**	**+21.1576**	**+0.0825**	**+0.2095**	**-0.3290**	**+42.0%**
Qwen-Image-Edit (Wu et al., 2025)	**official**	20.5206	0.7724	0.2445	0.1171	4.2007	67.3384	0.6219	0.5375	-1.0476	72.0%
	w/o	20.6342	0.7774	0.2920	0.1878	5.4593	49.6837	0.5796	0.3576	-0.9016	42.0%
	w/	21.8703	0.7803	0.1952	0.0997	4.6093	63.7163	0.6268	0.5074	-1.0201	64.0%
	Improvement	**+1.2361**	**+0.0029**	**-0.0968**	**-0.0881**	**-0.8500**	**+14.0326**	**+0.0472**	**+0.1498**	**-0.1185**	**+22.0%**

Table H.8: Comparison of representative restoration models "w/o" and "w/" GGT-100K on the **GGT100K-Snow** test set. For some models whose official releases can also handle real-world degradations, we additionally report their **official results**. "Improvement" indicates the performance gain brought by GGT-100K. The positive improvements are highlighted in **red**.

Model	GGT-100K	Full-reference fidelity metrics				No-reference perceptual metrics					VLM-R ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓	
MPRNet (Zamir et al., 2021)	w/o	22.6929	0.6970	0.3544	0.2089	4.8778	50.7305	0.5353	0.4197	-0.6816	12.0%
	w/	24.7152	0.7738	0.3144	0.2064	5.1178	49.8011	0.4927	0.3609	-0.7347	28.0%
	Improvement	**+2.0223**	**+0.0768**	**-0.0400**	**-0.0025**	+0.2400	-0.9294	-0.0426	-0.0588	**-0.0531**	**+16.0%**
NAFNet (Chen et al., 2022)	w/o	22.6229	0.6973	0.3633	0.2055	4.7855	48.0930	0.5307	0.4120	-0.6628	22.0%
	w/	26.4338	0.7983	0.3143	0.2047	5.5590	48.8828	0.4719	0.3371	-0.7594	54.0%
	Improvement	**+3.8109**	**+0.1010**	**-0.0490**	**-0.0008**	+0.7735	**+0.7898**	-0.0588	-0.0749	**-0.0966**	**+32.0%**
SwinIR (Liang et al., 2021)	w/o	21.3170	0.6695	0.3902	0.2213	4.7865	48.8828	0.5174	0.4254	-0.6632	0.0%
	w/	24.9210	0.7728	0.3183	0.1984	4.8716	48.7087	0.4809	0.3681	-0.7006	34.0%
	Improvement	**+3.6040**	**+0.1033**	**-0.0719**	**-0.0229**	+0.0851	-0.1741	-0.0365	-0.0573	**-0.0374**	**+34.0%**
XRestormer (Chen et al., 2024b)	w/o	22.7818	0.6951	0.3464	0.1983	4.6836	50.0319	0.5375	0.4172	-0.6860	14.0%
	w/	26.5405	0.7994	0.3113	0.2037	5.4127	49.0516	0.4727	0.3404	-0.7812	60.0%
	Improvement	**+3.7587**	**+0.1043**	**-0.0351**	+0.0054	+0.7291	-0.9803	-0.0648	-0.0768	**-0.0952**	**+46.0%**
PromptIR (Potlapalli et al., 2023)	w/o	23.3299	0.6993	0.3377	0.1966	4.7201	49.2773	0.5372	0.4202	-0.6737	14.0%
	w/	26.2458	0.7950	0.3081	0.2034	5.1493	48.5721	0.4764	0.3452	-0.7533	58.0%
	Improvement	**+2.9159**	**+0.0957**	**-0.0296**	+0.0068	+0.4292	-0.7052	-0.0608	-0.0750	**-0.0796**	**+44.0%**
MoCE-IR (Zamfir et al., 2024)	w/o	23.4365	0.7088	0.3396	0.1967	4.7384	50.5416	0.5346	0.4191	-0.6859	26.0%
	w/	26.3553	0.8067	0.3104	0.2128	5.7355	51.8140	0.4747	0.3609	-0.8247	56.0%
	Improvement	**+2.9188**	**+0.0979**	**-0.0292**	+0.0161	+0.9971	**+1.2724**	-0.0599	-0.0582	**-0.1388**	**+30.0%**
DA-CLIP (Luo et al., 2024)	w/o	23.6500	0.6909	0.3455	0.1987	4.9824	44.8509	0.5336	0.3966	-0.6463	28.0%
	w/	24.9975	0.7416	0.2613	0.1661	4.9824	48.2149	0.5276	0.3715	-0.7474	68.0%
	Improvement	**+1.3475**	**+0.0507**	**-0.0842**	**-0.0326**	+0.0000	**+3.3640**	-0.0060	-0.0251	**-0.1011**	**+40.0%**
FoundIR (Li et al., 2025)	**official**	22.8380	0.6975	0.3618	0.2091	4.5560	47.3037	0.5446	0.4130	-0.6632	0.0%
	w/o	22.8919	0.6881	0.3702	0.2127	5.0751	48.4214	0.5400	0.4112	-0.6722	8.0%
	w/	25.0554	0.7901	0.3346	0.2129	6.1618	50.4086	0.4704	0.3593	-0.8202	56.0%
	Improvement	**+2.1635**	**+0.1020**	**-0.0356**	+0.0002	+1.0867	**+1.9872**	-0.0696	-0.0519	**-0.1480**	**+48.0%**
FLUX-Controlnet (Labs, 2024)	w/o	20.5898	0.6079	0.4095	0.2227	4.6388	50.7200	0.5584	0.4556	-0.6629	24.0%
	w/	21.4757	0.6666	0.2797	0.1656	4.3965	62.1546	0.6034	0.5069	-0.8939	38.0%
	Improvement	**+0.8859**	**+0.0587**	**-0.1298**	**-0.0571**	**-0.2423**	**+11.4346**	**+0.0450**	**+0.0513**	**-0.2310**	**+14.0%**
Qwen-Image-Edit (Wu et al., 2025)	**official**	17.9621	0.6334	0.3941	0.2002	4.5264	68.2379	0.6160	0.5523	-1.0614	26.0%
	w/o	25.5801	0.7753	0.2352	0.1427	5.4487	50.8085	0.5479	0.3747	-0.8326	94.0%
	w/	24.2822	0.7512	0.2196	0.1253	5.0208	61.3688	0.5929	0.4657	-0.9823	88.0%
	Improvement	-1.2979	-0.0241	**-0.0156**	**-0.0174**	**-0.4279**	**+10.5603**	**+0.0450**	**+0.0910**	**-0.1497**	-6.0%

Table H.9: Comparison of representative restoration models "w/o" and "w/" GGT-100K on the **GGT100K-Low-Light** test set. For some models whose official releases can also handle real-world degradations, we additionally report their **official results**. "Improvement" indicates the performance gain brought by GGT-100K. The positive improvements are highlighted in **red**.

Model	GGT-100K	Full-reference fidelity metrics				No-reference perceptual metrics					VLM-R ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓	
MPRNet (Zamir et al., 2021)	w/o	18.5099	0.7166	0.4941	0.3778	8.6095	34.0214	0.4579	0.2767	-0.8531	18.0%
	w/	21.4928	0.7940	0.3998	0.2898	7.0979	41.2545	0.4774	0.3140	-0.8213	50.0%
	Improvement	**+2.9829**	**+0.0774**	**-0.0943**	**-0.0880**	**-1.5116**	**+7.2331**	**+0.0195**	**+0.0373**	+0.0318	**+32.0%**
NAFNet (Chen et al., 2022)	w/o	18.0173	0.7095	0.4846	0.3759	7.6623	35.7853	0.4648	0.3002	-0.8277	26.0%
	w/	22.7054	0.8123	0.3619	0.2547	6.2793	44.1940	0.4764	0.3236	-0.8196	78.0%
	Improvement	**+4.6881**	**+0.1028**	**-0.1227**	**-0.1212**	**-1.3830**	**+8.4087**	**+0.0116**	**+0.0234**	+0.0081	**+52.0%**
SwinIR (Liang et al., 2021)	w/o	16.9328	0.6797	0.5361	0.3780	6.9303	30.8260	0.4192	0.2444	-0.7589	6.0%
	w/	19.8966	0.7693	0.4307	0.2994	6.0399	38.6161	0.4279	0.2874	-0.7587	40.0%
	Improvement	**+2.9638**	**+0.0896**	**-0.1054**	**-0.0786**	**-0.8904**	**+7.7901**	**+0.0087**	**+0.0430**	+0.0002	**+34.0%**
XRestormer (Chen et al., 2024b)	w/o	18.1090	0.7399	0.4638	0.3478	7.5393	38.4316	0.4529	0.3013	-0.8473	42.0%
	w/	22.6739	0.8192	0.3616	0.2543	7.0478	45.4584	0.4766	0.3153	-0.8754	86.0%
	Improvement	**+4.5649**	**+0.0793**	**-0.1022**	**-0.0935**	**-0.4915**	**+7.0268**	**+0.0237**	**+0.0140**	**-0.0281**	**+44.0%**
PromptIR (Potlapalli et al., 2023)	w/o	18.7550	0.7231	0.4789	0.3698	7.8158	37.2343	0.4724	0.2991	-0.8759	42.0%
	w/	22.4263	0.8165	0.3541	0.2526	6.8982	45.4209	0.4805	0.3257	-0.8846	76.0%
	Improvement	**+3.6713**	**+0.0934**	**-0.1248**	**-0.1172**	**-0.9176**	**+8.1866**	**+0.0081**	**+0.0266**	**-0.0087**	**+34.0%**
MoCE-IR (Zamfir et al., 2024)	w/o	18.4825	0.7225	0.4680	0.3596	7.7570	36.5115	0.4721	0.2994	-0.8655	34.0%
	w/	22.4736	0.8177	0.3691	0.2592	7.2668	45.6688	0.4733	0.3228	-0.8901	80.0%
	Improvement	**+3.9911**	**+0.0952**	**-0.0989**	**-0.1004**	**-0.4902**	**+9.1573**	**+0.0012**	**+0.0234**	**-0.0246**	**+46.0%**
DA-CLIP (Luo et al., 2024)	w/o	16.8809	0.6615	0.5159	0.4141	8.7764	32.9001	0.4939	0.2897	-0.8783	30.0%
	w/	18.7191	0.7159	0.4115	0.3173	7.2236	37.3314	0.5036	0.3105	-0.8745	56.0%
	Improvement	**+1.8382**	**+0.0544**	**-0.1044**	**-0.0968**	**-1.5528**	**+4.4313**	**+0.0097**	**+0.0208**	+0.0038	**+26.0%**
FoundIR (Li et al., 2025)	**official**	18.9683	0.7556	0.4202	0.3376	7.4485	37.1942	0.4885	0.3093	-0.8596	58.0%
	w/o	21.3275	0.7811	0.4194	0.3258	8.2551	42.7873	0.4618	0.3236	-0.9248	64.0%
	w/	19.8822	0.7638	0.4104	0.3174	9.0116	41.5394	0.4805	0.3177	-0.9438	84.0%
	Improvement	-1.4453	-0.0173	**-0.0090**	**-0.0084**	+0.7565	-1.2479	**+0.0187**	-0.0059	**-0.0190**	**+20.0%**
FLUX-Controlnet (Labs, 2024)	w/o	19.9199	0.7312	0.4508	0.2572	6.5845	50.1438	0.5236	0.4068	-0.8537	46.0%
	w/	19.7996	0.7279	0.3383	0.1962	6.0633	57.7316	0.5753	0.4686	-0.9991	54.0%
	Improvement	-0.1203	-0.0033	**-0.1125**	**-0.0610**	**-0.5212**	**+7.5878**	**+0.0517**	**+0.0618**	**-0.1454**	**+8.0%**
Qwen-Image-Edit (Wu et al., 2025)	**official**	17.1073	0.6953	0.4166	0.2748	7.1067	50.8637	0.5216	0.3968	-0.9757	66.0%
	w/o	21.1225	0.7926	0.3344	0.2034	7.0565	48.3712	0.5091	0.3455	-0.9300	88.0%
	w/	23.0697	0.7907	0.2502	0.1458	6.6361	60.6649	0.5730	0.4580	-1.0349	94.0%
	Improvement	**+1.9472**	-0.0019	**-0.0842**	**-0.0576**	**-0.4204**	**+12.2937**	**+0.0639**	**+0.1125**	**-0.1049**	**+6.0%**

Table H.10: Comparison of representative restoration models "w/o" and "w/" GGT-100K on the **GGT100K-Old Photo** test set. For some models whose official releases can also handle real-world degradations, we additionally report their **official results**. "Improvement" indicates the performance gain brought by GGT-100K. The positive improvements are highlighted in **red**.

Model	GGT-100K	Full-reference fidelity metrics				No-reference perceptual metrics					VLM-R ↑
		PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	AFINE-NR ↓	
MPRNet (Zamir et al., 2021)	w/o	26.2769	0.8285	0.3231	0.1835	6.1129	44.5421	0.4939	0.3122	-0.6687	30.0%
	w/	29.2252	0.8686	0.2830	0.1836	6.0897	48.5279	0.4784	0.3308	-0.7995	50.0%
	Improvement	**+2.9483**	**+0.0401**	**-0.0401**	+0.0001	**-0.0232**	**+3.9858**	-0.0155	**+0.0186**	**-0.1308**	**+20.0%**
NAFNet (Chen et al., 2022)	w/o	26.8340	0.8431	0.3050	0.1723	5.9239	43.2106	0.4933	0.3029	-0.6528	48.0%
	w/	29.5942	0.8758	0.2637	0.1744	6.0141	49.2069	0.4633	0.3277	-0.8143	56.0%
	Improvement	**+2.7602**	**+0.0327**	**-0.0413**	+0.0021	+0.0902	**+5.9963**	-0.0300	**+0.0248**	**-0.1615**	**+8.0%**
SwinIR (Liang et al., 2021)	w/o	26.2180	0.8363	0.3265	0.1774	6.1039	45.1656	0.4864	0.3116	-0.6682	36.0%
	w/	28.8712	0.8666	0.2873	0.1746	6.2721	45.7231	0.4632	0.3002	-0.7664	50.0%
	Improvement	**+2.6532**	**+0.0303**	**-0.0392**	**-0.0028**	+0.1682	**+0.5575**	-0.0232	-0.0114	**-0.0982**	**+14.0%**
XRestormer (Chen et al., 2024b)	w/o	27.1331	0.8506	0.3032	0.1700	6.0296	44.4367	0.5019	0.3143	-0.6845	48.0%
	w/	29.8823	0.8769	0.2665	0.1727	6.0065	49.0715	0.4639	0.3302	-0.8300	56.0%
	Improvement	**+2.7492**	**+0.0263**	**-0.0367**	+0.0027	**-0.0231**	**+4.6348**	-0.0380	**+0.0159**	**-0.1455**	**+8.0%**
PromptIR (Potlapalli et al., 2023)	w/o	26.3384	0.8189	0.2998	0.1782	5.8692	47.2604	0.5098	0.3339	-0.6678	36.0%
	w/	29.8668	0.8780	0.2644	0.1759	6.0530	49.1353	0.4671	0.3298	-0.8159	56.0%
	Improvement	**+3.5284**	**+0.0591**	**-0.0354**	**-0.0023**	+0.1838	**+1.8749**	-0.0427	-0.0041	**-0.1481**	**+20.0%**
MoCE-IR (Zamfir et al., 2024)	w/o	26.9140	0.8378	0.3017	0.1724	5.9494	44.7693	0.5014	0.3189	-0.6697	46.0%
	w/	29.9449	0.8804	0.2610	0.1792	5.9774	50.6919	0.4667	0.3472	-0.8520	60.0%
	Improvement	**+3.0309**	**+0.0426**	**-0.0407**	+0.0068	+0.0280	**+5.9226**	-0.0347	**+0.0283**	**-0.1823**	**+14.0%**
DA-CLIP (Luo et al., 2024)	w/o	27.9405	0.8404	0.2929	0.1623	5.7807	39.3277	0.5031	0.2848	-0.6372	24.0%
	w/	28.5906	0.8473	0.2342	0.1419	5.6770	43.1152	0.4972	0.3045	-0.7158	56.0%
	Improvement	**+0.6501**	**+0.0069**	**-0.0587**	**-0.0204**	**-0.1037**	**+3.7875**	-0.0059	**+0.0197**	**-0.0786**	**+32.0%**
FoundIR (Li et al., 2025)	**official**	27.7643	0.8439	0.3062	0.1696	6.0045	39.8216	0.5091	0.2937	-0.6412	26.0%
	w/o	28.2156	0.8505	0.2944	0.1661	6.0254	42.8725	0.5014	0.3050	-0.6901	54.0%
	w/	29.5300	0.8769	0.2752	0.1727	6.3344	44.5586	0.4687	0.2985	-0.7785	68.0%
	Improvement	**+1.3144**	**+0.0264**	**-0.0192**	+0.0066	+0.3090	**+1.6861**	-0.0327	-0.0065	**-0.0884**	**+14.0%**
FLUX-Controlnet (Labs, 2024)	w/o	25.4824	0.7785	0.3341	0.1789	5.5024	43.5297	0.5163	0.3259	-0.5865	28.0%
	w/	24.1773	0.7715	0.2292	0.1332	3.9432	62.3481	0.6063	0.5151	-0.9245	68.0%
	Improvement	-1.3051	-0.0070	**-0.1049**	**-0.0457**	**-1.5592**	**+18.8184**	**+0.0900**	**+0.1892**	**-0.3380**	**+40.0%**
Qwen-Image-Edit (Wu et al., 2025)	**official**	21.6575	0.7757	0.3228	0.1788	4.3479	65.5946	0.5940	0.5241	-0.9868	74.0%
	w/o	26.1690	0.8183	0.2410	0.1281	4.9971	56.0239	0.5749	0.4267	-0.8615	74.0%
	w/	27.1337	0.8298	0.2094	0.1143	4.4643	62.7812	0.5963	0.4877	-0.9779	76.0%
	Improvement	**+0.9647**	**+0.0115**	**-0.0316**	**-0.0138**	**-0.5328**	**+6.7573**	**+0.0214**	**+0.0610**	**-0.1164**	**+2.0%**