LEARNED DEGRADATION GUIDANCE: TURNING FROZEN IMAGE EDITORS INTO CONTROLLABLE RESTORERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Instruction-based image editors have a strong image prior, and classifier-free guidance allows them to trade fidelity for perceptual quality with a single scale. Guidance pushes the output away from a second prediction of the model. In restoration, the model usually makes this second prediction under a generic low-quality prompt. Such a prompt names a degradation but cannot show the one in the image, so the model produces a degradation of its own. We introduce Learned Degradation Guidance (LDG), in which a trained adapter, not a prompt, makes the second prediction. On a frozen editor, a degradation adapter makes an image's own degradation stronger, and a restoration adapter removes it. The two adapters are trained separately, only on images that the editor generates itself, and are combined by guidance at inference. Guidance then pushes the output away from the image's own degradation, and one scale takes the editor from faithful to sharp restoration. With this second prediction, the editor reaches a higher perceptual quality than with one prompted to degrade. In real-world super-resolution tests, LDG with LoRA adapters is closer to the ground truth than the state of the art at its most faithful setting, reaches the highest perceptual quality at larger scales, and at equal or higher quality stays closer to the ground truth in almost every comparison. On real photographs with nine kinds of degradation, such as haze, rain and reflections, it achieves the best average score among the compared methods on a recent benchmark, with one adapter pair per degradation, each trained on about a hundred generated images. Source-attention calibration, an adapter that only reweights the editor's attention with a few thousand parameters, is almost as faithful as LoRA at the faithful end of the scale.

1 INTRODUCTION

Generative models have made restored photographs far more realistic. Realism and fidelity, however, pull against each other. A convincing texture need not be the one in the scene, and the reconstruction with the lowest average error tends to look blurry (Blau & Michaeli, 2018). The right balance depends on the application, so the user should set it at inference, ideally with one control.

Instruction-based image editors, such as FLUX.1 Kontext (Black Forest Labs et al., 2025), are promising restorers. They are trained to change what an instruction asks for and to keep the rest of the image, and restoration is such an edit. Current editor-based restorers, however, need paired training data. RealRestorer (Yang et al., 2026) fine-tunes Step1X-Edit (Liu et al., 2025) on about 1.7 million pairs of degraded and clean images, CoTIR (Guo et al., 2026) trains adapters on FLUX editors with about 5.2 million pairs, and Edit2Restore (Yılmaz et al., 2026) adapts FLUX.1 Kontext with 32–128 pairs per task. Each of them is released at one balance between fidelity and quality, with no trained control to change it.

Classifier-free guidance (CFG) (Ho & Salimans, 2022) is the standard single control for such a balance. It pushes the output away from a second prediction of the model: the guidance scale sets how far, and the second prediction sets the direction. In restoration, the second prediction usually comes from the same network under a low-quality prompt (Lin et al., 2024; Yu et al., 2024). S3Diff (Zhang et al., 2024a) trains the network under such a prompt to return the degraded input, on

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

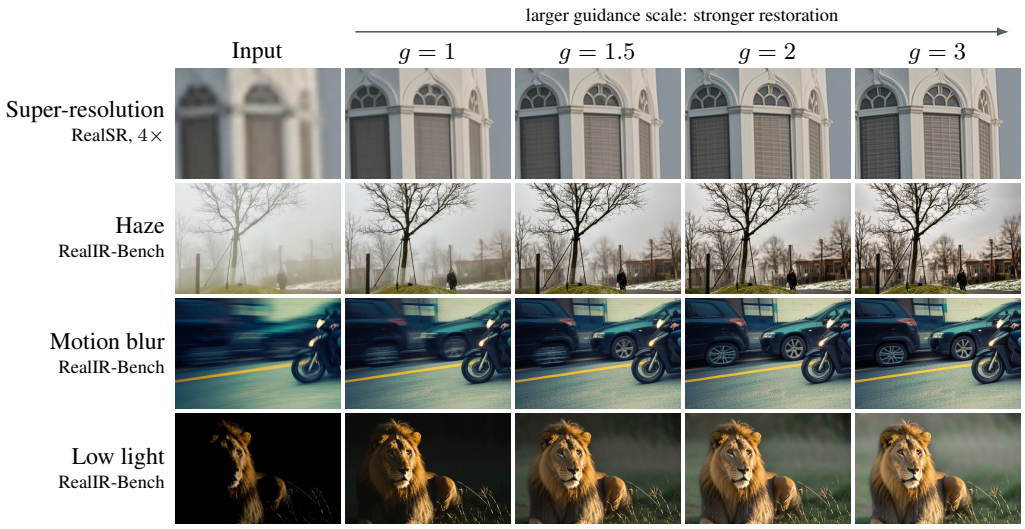


Figure 1: **One guidance scale sets how strongly a frozen editor restores.** Each row shows a degraded photograph (left) and the output of LDG-LoRA on FLUX.2 [klein] 9B at four values of the guidance scale g. At $g = 1$, the restoration adapter runs alone; a larger g pushes the output further away from the degradation adapter's prediction. Top: $4\times$ super-resolution of a RealSR image (Cai et al., 2019), where the slats of the shutters appear as g grows. Other rows: RealIR-Bench photographs (Yang et al., 2026), where the buildings behind the fog and the wheels of the smeared cars appear as g grows. A large g also fills the black background behind the lion with grass.

part of its training data. This design has two limitations. First, a prompt can name a degradation, but it cannot show the degradation of this image, e.g., which blur, where, and how strong. Even when asked to strengthen the degradation the input already has, the model produces one of its own. Guidance then also pushes away from a degradation that is not in the image. Second, when one restoration network makes both predictions, the two differ little, so guidance needs a large scale to change the output, and at a large scale it starts to repaint content. We hypothesize that guidance works better with a second prediction that shows this image's own degradation, in the same places, only stronger.

We propose *Learned Degradation Guidance* (LDG), which obtains this second prediction from a trained adapter instead of a prompt. LDG trains two small adapters on a frozen editor. A *degradation adapter* learns to map a degraded image to a more degraded version of the same image, in which the same degradation, such as blur, haze or rain, is made stronger. A *restoration adapter* learns to map the same image to its clean version. Because both adapters see the same input and differ only in their targets, the difference between their predictions points along this input's own degradation. One guidance scale g then takes the output from faithful to sharp (Figure 1); at $g = 1$, the restoration adapter runs alone. The adapters are trained separately, each on the states that its own sampler visits, and only on images that the frozen editor generates itself, each degraded twice with the same random draw. No real image is used.

We also ask how far restoration can go by only reweighting the editor's attention. For this, we introduce *source-attention calibration* (SAC), a new type of adapter with just a few thousand parameters. SAC changes no weights and adds no tokens: it only learns how strongly the generated image attends to the input image, to the instruction and to itself, separately for each attention head and varying with the noise level. We call the two variants LDG-LoRA and LDG-SAC.

We evaluate LDG on FLUX.2 [klein] 9B (Black Forest Labs, 2026) for real-world super-resolution and for the nine real degradation types of RealIR-Bench (Yang et al., 2026), and we repeat the super-resolution study on FLUX.1 Kontext [dev]. In super-resolution, LDG gives the highest perceptual quality of all compared methods on all six benchmarks. LDG-LoRA is also closer to the ground truth than ten recent methods at $g = 1$, and at equal or higher perceptual quality it stays closer in 59 of 60 comparisons (Section 4.2). On RealIR-Bench, with one adapter pair per degradation type, it achieves

the highest average score among the compared methods, ahead of editor-based restorers fine-tuned on millions of image pairs (Section 4.3). The same recipe carries over to Kontext, where LDG-LoRA is again closer to the ground truth than all ten methods at $g = 1$. SAC, with a few thousand parameters, is almost as faithful as LoRA (Hu et al., 2022) at $g = 1$. Finally, the learned second prediction follows the input's degradation: on held-out synthetic images, its guidance direction is about three times better aligned with the true restoration direction than a prompted one, and guidance with it reaches a higher perceptual quality at a smaller scale (Section 4.4).

2 Related work

Guidance with a second prediction. In CFG, one network gives both predictions (Ho & Salimans, 2022). Several methods take the second prediction from a weaker model instead; Autoguidance uses a smaller, less-trained model (Karras et al., 2024), SIMS uses a copy fine-tuned on the model's own samples (Alemohammad et al., 2024), and Diffusion-NPO uses a weight offset trained on reversed preferences (Wang et al., 2025). PAG perturbs the model's self-attention (Ahn et al., 2024), and SAG blurs the parts of the sample that the model attends to (Hong et al., 2023). All of these weaken the model or the sample. In restoration, DiffBIR (Lin et al., 2024) and the released code of SeeSR (Wu et al., 2024c) use a low-quality negative prompt with the same network. S3Diff (Zhang et al., 2024a) trains the same network under a negative prompt to return the degraded input itself, on part of its training data. LDG instead trains a separate adapter to strengthen the input's own degradation.

Generative restoration and its controls. StableSR (Wang et al., 2024a), SUPIR (Yu et al., 2024) and SeeSR (Wu et al., 2024c) condition a pretrained text-to-image diffusion model on the degraded image, and DiffBIR (Lin et al., 2024) on a first restoration of it. Later methods sample in one or a few steps (Wang et al., 2024b; Wu et al., 2024b; Sun et al., 2025; Dong et al., 2025; Yue et al., 2025; Lin et al., 2025) or use diffusion transformers (Duan et al., 2025; Wu et al., 2026). They train on real images with synthetic degradations (Zhang et al., 2021; Wang et al., 2021) or add generated images to them (Ai et al., 2024). DNF-SR (Han et al., 2026) fine-tunes FLUX.1 Kontext with positive and negative reward directions, but uses them only in its training loss, not at inference. Methods that let the user set the fidelity–quality trade-off expose one control of a trained model or its sampler: a fusion weight, the starting step, injected noise, the two LoRA scales of PiSA-SR (Sun et al., 2025) or a guidance scale (Appendix D). VOSR guides between two modes of one shared model (Wu et al., 2026). LDG instead guides between two separately trained adapters, one trained to degrade.

Compact adaptation. Small adapters train only bias terms (Ben Zaken et al., 2022), bias terms and scale factors (Xie et al., 2023), or rescale activations (Liu et al., 2022). Closer to SAC, Omini-Control, TACA and HeadRouter adjust attention between token groups or heads with hand-set biases and temperatures or training-free head weights (Tan et al., 2025; Lv et al., 2025; Xu et al., 2026), and LLaMA-Adapter learns gates for attention to the prompt tokens it adds (Zhang et al., 2024b). To our knowledge, SAC is the first adapter that consists only of learned attention biases, one per layer, head and token source, affine in the noise level, without changing any weights or adding tokens.

3 Method

LDG adapts a frozen image editor for restoration with two small adapters, which are combined at inference (Figure 2). The main novelty is the second prediction of guidance: an adapter trained to strengthen the input's own degradation. A second novelty is source-attention calibration (SAC, Section 3.5), a new type of adapter that changes no weights and only learns attention biases. The adapters learn from images that the editor generates itself, so no real image is needed.

3.1 Image editors

We consider an instruction-based editor built on a flow-matching transformer (Lipman et al., 2023). It takes an input image y and a text instruction p. Its autoencoder turns y into *reference tokens*. The transformer attends jointly over three sources of tokens: the text tokens of p, the reference tokens, and the *target tokens* of the image being generated. We write $v(z, \sigma; y, p)$ for the velocity that the model predicts for the target latent z at noise level σ. We use the linear path of rectified flow (Liu

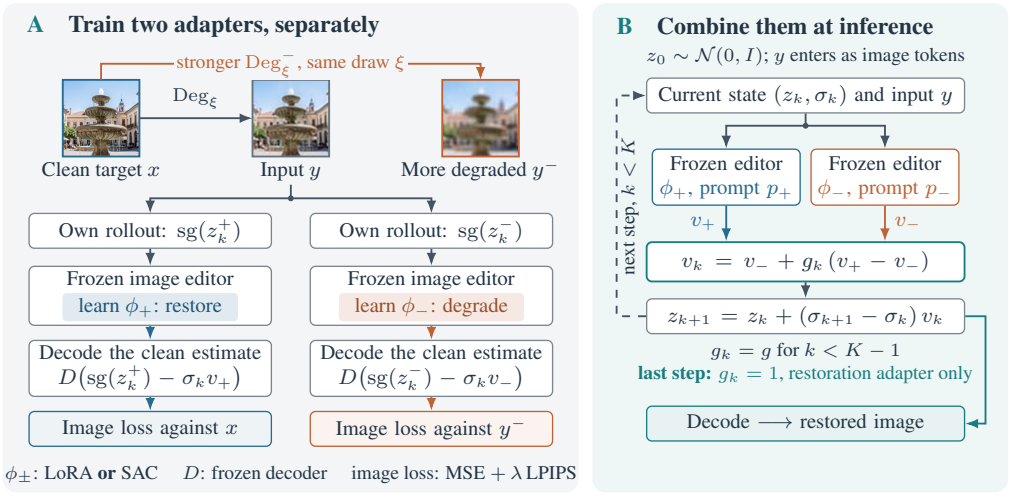


Figure 2: **Learned Degradation Guidance. A:** A random draw Deg_ξ degrades a clean render x into the input y. The same draw made stronger, Deg_ξ^-, degrades x into the target y^-. The nine RealIR degradations are built the same way. Each adapter is trained on the states of its own sampler (sg: stop-gradient). **B:** At inference, guidance moves past the restoration adapter's velocity, away from the degradation adapter's, by an amount set by g. Gray: frozen; colored: trained.

et al., 2023): a latent at noise level σ is $z = (1-\sigma)a + \sigma\epsilon$, where a is a clean latent and ϵ is Gaussian noise. Then $z - \sigma v(z, \sigma; y, p)$ is the model's estimate of a. Sampling starts from Gaussian noise z_0 and takes K Euler steps $z_{k+1} = z_k + (\sigma_{k+1} - \sigma_k)\, v(z_k, \sigma_k; y, p)$, from $\sigma_0 = 1$ to $\sigma_K = 0$. All pretrained weights stay frozen; an adapter with parameters ϕ changes the predicted velocity to v_ϕ.

3.2 Learned Degradation Guidance

Training uses triples (x, y, y^-): a clean image x, a degraded version y, and a more strongly degraded version y^- of the same image (built as described in Section 3.3). LDG trains two adapters on these triples once and uses them unchanged for every input image. The *restoration adapter* ϕ_+ is used with a restoration instruction p_+ and learns to map y to x. The *degradation adapter* ϕ_- is used with a *degradation prompt* p_- that asks to strengthen the degradation already present, and it learns to map the same y to y^-. At inference, both adapters are evaluated at the same state and combined as in classifier-free guidance (Ho & Salimans, 2022):

$$v_k = v_-(z_k) + g_k\big(v_+(z_k) - v_-(z_k)\big), \qquad g_k = \begin{cases} g, & k < K-1, \\ 1, & k = K-1, \end{cases} \tag{1}$$

where $v_\pm(z_k) = v_{\phi_\pm}(z_k, \sigma_k; y, p_\pm)$. At $g = 1$, the sampler follows the restoration adapter alone. For $g > 1$, each step moves past its prediction, away from the degradation adapter's. The last step always uses the restoration adapter alone, since a guided last step returns an extrapolated clean estimate and, on RealSR, costs 1.15 dB PSNR at MUSIQ 71 (Appendix C.1, Table 5).

What the guidance direction contains. Let $\hat{a}_\pm = z_k - \sigma_k v_\pm(z_k)$ be the clean estimates of the two adapters and $\hat{a}_g = z_k - \sigma_k v_k$ the guided one. Equation (1) gives exactly

$$\hat{a}_g = \hat{a}_+ + (g-1)\,(\hat{a}_+ - \hat{a}_-). \tag{2}$$

Each guided step thus moves the clean estimate along the difference between the two predictions, so the second prediction sets the direction of this move. If $\hat{a}_-$ shows the input's own degradation, only stronger, the difference points toward removing that degradation. If $\hat{a}_-$ shows a degradation that comes from the prompt rather than the input, the difference also points away from something that is not in the input, and moving along it changes the content. If $\hat{a}_-$ is close to $\hat{a}_+$, the difference is small and g must be large. The degradation adapter is trained for the first case (Section 3.3), and Section 4.4 measures all three. A scale $g > 1$ can remove degradation that the restoration estimate keeps. Where that estimate is already exact, $g = 1$ is the most faithful setting (Appendix C).

3.3 TRAINING DATA

The frozen editor generates 100 clean images x from generic photographic prompts. A random degradation turns each x into the input y. The same draw, made stronger, gives the target y^-, so the target is predictable from the input. In super-resolution, the draw mixes blur, ringing, resolution loss, noise and compression, and a fixed step strengthens each of them except the noise (Appendix B). The nine RealIR degradations, such as haze, rain and reflections, are strengthened the same way, in the same places (Appendix F). Another 50 triples, made from other subjects, are held out and never used for training.

3.4 TRAINING EACH ADAPTER ON ITS OWN SAMPLING STATES

Each adapter is trained alone, on the states that its own sampler visits. At each update, we draw a training triple, Gaussian noise z_0 and a step k uniformly from $\{0, \dots, K-1\}$. We run the adapter's sampler from z_0 to z_k without guidance and without gradients, so the update is a semi-gradient (Sutton & Barto, 2018), as in ReFL (Xu et al., 2023) (Appendix C.3). One evaluation at z_k, with gradients, gives a clean estimate. The frozen decoder D turns it into an image $\hat{x} = D\big(z_k - \sigma_k v_\phi(z_k, \sigma_k; y, p)\big)$, where (ϕ, p) is (ϕ_+, p_+) for the restoration adapter and (ϕ_-, p_-) for the degradation adapter. This image is compared with the adapter's target t, which is x for the restoration adapter and y^- for the degradation adapter:

$$\mathcal{L} = \mathrm{MSE}(\hat{x}, t) + \lambda\ \mathrm{LPIPS}(\hat{x}, t). \tag{3}$$

LPIPS uses VGG features (Zhang et al., 2018). The weight λ is the ratio of the MSE and LPIPS gradient norms, measured once at initialization (Appendix A).

3.5 TWO ADAPTER TYPES

LoRA. LoRA (Hu et al., 2022) adds trainable low-rank updates to linear layers, as in earlier editor-based restorers (Guo et al., 2026; Yılmaz et al., 2026). We use rank 16 on the attention and MLP projections listed in Appendix A. This gives 27.8M parameters per adapter on FLUX.2 [klein] 9B.

Source-attention calibration (SAC). In every attention layer, the target tokens attend to three sources: the reference tokens of the input image, the text tokens of the instruction, and themselves. If they attend too little to the input image, the editor may invent content. If they attend too much, the degradation may stay. The right balance may also change during sampling. SAC learns only this balance, so it shows how far restoration can go without changing the editor's weights. In attention layer ℓ and head h, SAC adds one learned bias per source to the attention logits of the target queries:

$$a_{ij} = \mathrm{softmax}_j\Big(\frac{q_i^\top k_j}{\sqrt{d}} + \theta_{\ell h, s(j)}(\sigma)\Big), \qquad \theta_{\ell h, s}(\sigma) = \theta^0_{\ell h, s} + \tilde{\sigma}\,\theta^1_{\ell h, s}. \tag{4}$$

Here i is a target token, j ranges over all keys, q_i and k_j are the query and key vectors, d is the head dimension, $s(j) \in \{\text{reference}, \text{text}, \text{target}\}$ is the source of key j, and $\tilde{\sigma}$ is the noise level, standardized over the training schedule. A logit bias multiplies all unnormalized attention weights of its source by the same factor. The total attention a target token gives to each source therefore changes, while in that layer the weights within a source keep their frozen ratios. The two coefficients let the balance change with the noise level, and zero biases give back the frozen editor. On FLUX.2 [klein] 9B, with 32 attention layers of 32 heads, SAC has $32 \times 32 \times 3 \times 2 = 6{,}144$ parameters, about 4,500 times fewer than LoRA (Appendices A and C.2).

4 EXPERIMENTS

4.1 SETUP

Editors and training. We run the full study on FLUX.2 [klein] 9B (Black Forest Labs, 2026), which we call Klein. It is a step-distilled editor that samples in $K = 4$ steps. We repeat the super-resolution study, with both adapter types, on FLUX.1 Kontext [dev] (Black Forest Labs et al., 2025), which we call Kontext. It is guidance-distilled (Meng et al., 2023), has a different architecture and autoencoder, and samples in $K = 28$ steps. Both editors use the same training recipe, each on training data it generates itself (Appendix E.4).

Table 1: **Blind super-resolution on three real-world benchmarks with ground truth** (sets without it: Table 4). PSNR, SSIM and LPIPS compare the output with the ground truth (fidelity); MUSIQ, MANIQA, CLIP-IQA and LIQE score the output alone (perceptual quality). Ours at two guidance scales, both with AdaIN color correction: the *fidelity point* $g = 1$ (the restoration adapter alone) and the *quality point*, a larger g. **Bold red**: best, underlined blue: second best.

		Related work										Ours on Klein 9B				Ours on Kontext			
												LDG-LoRA		LDG-SAC		LDG-LoRA		LDG-SAC	
Set	Metric	PiSA-SR	TSD-SR	OSEDiff	SeeSR	InvSR	VOSR	HYPIR	SinSR	DiT4SR	CoTIR	g=1	g=2.5	g=1	g=1.5	g=1	g=1.5	g=1	g=1.5
RealSR	PSNR ↑	25.50	24.81	25.15	25.46	24.45	25.33	22.85	25.97	23.41	25.48	27.12	25.13	26.95	24.86	27.48	25.66	**27.50**	23.42
	SSIM ↑	.742	.717	.734	.734	.714	.714	.673	.729	.663	.773	.770	.695	.760	.666	.781	.728	**.782**	.656
	LPIPS ↓	.267	.274	.292	.296	.285	.298	.315	.318	.324	.352	**.237**	.296	**.237**	.330	.258	.275	.257	.309
	MUSIQ ↑	70.16	71.18	69.08	68.87	68.35	70.53	66.24	61.66	68.09	40.38	66.92	71.86	61.82	72.05	66.50	71.59	63.02	**72.12**
	MANIQA ↑	.655	.635	.633	.638	.659	.646	.653	.548	.661	.381	.621	.704	.599	**.713**	.589	.662	.552	.659
	CLIP-IQA ↑	.670	.716	.669	.665	.678	.561	.647	.644	.639	.308	.585	**.732**	.516	.717	.551	.701	.450	.671
	LIQE ↑	4.10	4.27	4.07	4.03	4.04	3.82	3.94	3.30	4.00	1.59	3.69	**4.55**	3.22	4.51	3.52	4.38	3.02	4.06
DRealSR	PSNR ↑	28.32	27.77	27.92	28.45	26.40	27.87	25.91	27.93	25.63	27.04	29.75	27.15	29.52	26.88	30.10	27.95	**30.32**	25.47
	SSIM ↑	.780	.756	.784	.780	.715	.748	.713	.741	.677	**.829**	.810	.709	.797	.663	.825	.750	**.829**	.654
	LPIPS ↓	.296	.297	.297	.313	.352	.341	.348	.367	.369	.397	**.264**	.359	.273	.405	.282	.313	.267	.355
	MUSIQ ↑	66.12	66.60	64.67	64.53	65.81	63.65	60.88	57.91	64.39	29.75	62.47	**69.09**	56.86	68.89	61.54	68.41	55.17	67.43
	MANIQA ↑	.616	.587	.590	.591	.627	.580	.604	.507	.630	.336	.561	.672	.539	**.676**	.521	.634	.494	.619
	CLIP-IQA ↑	.697	**.734**	.696	.683	.716	.544	.641	.664	.666	.314	.582	.727	.515	.721	.551	.686	.441	.615
	LIQE ↑	4.04	4.13	3.94	4.04	4.04	3.47	3.71	3.30	3.96	1.28	3.30	4.25	2.68	**4.27**	3.24	4.24	2.52	3.58
ScreenSR	PSNR ↑	22.21	21.60	22.11	22.55	21.29	22.49	19.41	23.13	20.71	24.71	24.96	23.35	24.08	22.90	**25.43**	24.15	24.91	21.87
	SSIM ↑	.641	.635	.630	.632	.610	.638	.549	.657	.589	**.766**	.764	.709	.733	.686	.759	.730	.741	.649
	LPIPS ↓	.195	.195	.230	.223	.251	.216	.249	.219	.259	.205	.177	.173	.167	.171	.185	**.162**	.202	.196
	MUSIQ ↑	72.81	74.13	71.43	72.27	71.81	72.59	72.38	67.65	71.21	67.56	71.52	74.05	71.69	73.67	69.01	72.99	68.76	**74.41**
	MANIQA ↑	.727	.703	.681	.689	.709	.711	.735	.598	.708	.651	.703	.746	.719	**.751**	.653	.693	.642	.712
	CLIP-IQA ↑	.709	.752	.674	.704	.694	.580	.718	.618	.693	.568	.661	**.777**	.683	.763	.587	.741	.568	.770
	LIQE ↑	4.69	4.70	4.61	4.65	4.59	4.56	4.50	3.89	4.40	3.88	4.52	**4.84**	4.59	4.82	4.13	4.69	4.12	4.77

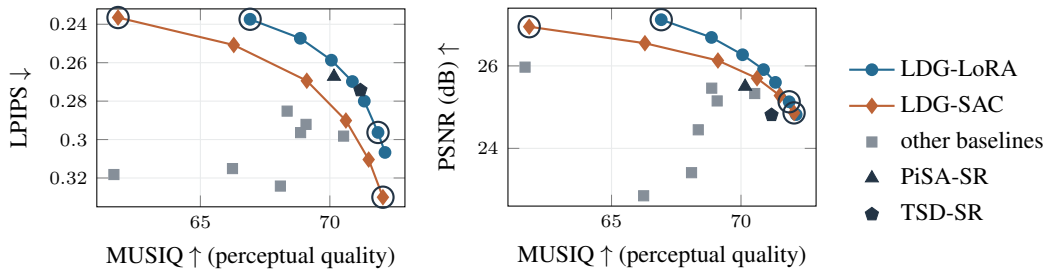


Figure 3: **One scale from fidelity to perceptual quality** on RealSR. Each curve is one adapter pair, from $g = 1$ to 3 for LDG-LoRA and to 1.5 for LDG-SAC. Circles mark the operating points of Table 1, and other markers the baselines. Up is closer to the ground truth: LDG-LoRA lies above every baseline, except TSD-SR by .0025 LPIPS. The ground truth itself scores MUSIQ 57.5.

Benchmarks, metrics and baselines. We use six real-world super-resolution sets. RealSR (Cai et al., 2019), DRealSR (Wei et al., 2020) and ScreenSR (Wu et al., 2026) have ground truth; RealPhoto60 (Yu et al., 2024), RealLQ250 (Ai et al., 2024) and RealLR200 (Wu et al., 2024c) do not (Appendix E). ScreenSR and RealPhoto60 are restored at their input size, and the other four sets at 4×. We measure fidelity to the ground truth with PSNR and SSIM (Wang et al., 2004) on luminance and with LPIPS (Zhang et al., 2018) (AlexNet), and perceptual quality with MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), CLIP-IQA (Wang et al., 2023) and LIQE (Zhang et al., 2023), all from IQA-PyTorch (Chen & Mo, 2022). We compare with ten recent methods: PiSA-SR (Sun et al., 2025), TSD-SR (Dong et al., 2025), OSEDiff (Wu et al., 2024b), SeeSR (Wu et al., 2024c), InvSR (Yue et al., 2025), VOSR (Wu et al., 2026), HYPIR (Lin et al., 2025), SinSR (Wang et al., 2024b), DiT4SR (Duan et al., 2025) and CoTIR-9B (Guo et al., 2026). Each runs with its authors' code, weights and settings, and one evaluator scores all methods (Appendix E). Our super-resolution pipeline ends with the AdaIN color correction of StableSR (Huang & Belongie, 2017; Wang et al., 2024a), as in the released code of six baselines. We report each pair at a *fidelity* point, $g = 1$, and a *quality* point (Table 1). The two guidance scales are the same on every set.

4.2 Blind Super-Resolution

LDG gives the highest perceptual quality of all compared methods on all six sets: the best MUSIQ, MANIQA and LIQE of every set, and the best CLIP-IQA of five, come from our pairs (Tables 1

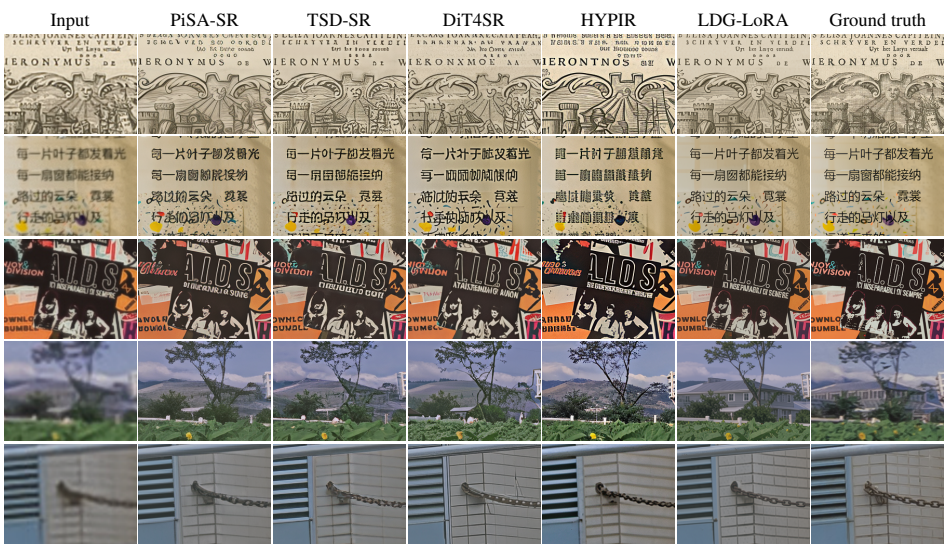


Figure 4: **Blind super-resolution: where the baselines invent content.** ScreenSR (rows 1–3) and RealSR ($4\times$) crops; baselines: PiSA-SR (Sun et al., 2025), TSD-SR (Dong et al., 2025), DiT4SR (Duan et al., 2025), HYPIR (Lin et al., 2025). The baselines rewrite the Latin title, the Chinese poem and the sticker text, paint over the house as a hillside and merge the chain into a rope. LDG-LoRA ($g = 2.5$) keeps them. Random examples and failures: Appendix E.1.

and 4). One LDG-LoRA pair covers the whole fidelity–quality trade-off (Figure 3; the whole scale is in Table 8). On all three sets with ground truth, it reaches a higher MUSIQ than every baseline (on ScreenSR, at $g = 3$). At the MUSIQ of each baseline on our scale, or at $g = 1$ where the baseline's is lower, LDG-LoRA is closer to the ground truth: its PSNR is higher and its LPIPS lower in 59 of the 60 comparisons (Appendix E). The one exception is the LPIPS of TSD-SR on RealSR, lower by .0025. Matched on MANIQA, CLIP-IQA or LIQE instead, LDG-LoRA wins 60, 53 and 54 of the 60.

At the fidelity point, LDG-LoRA has a higher PSNR and a lower LPIPS than every baseline on all three sets with ground truth, on Klein and on Kontext (Appendix E.4). Only CoTIR-9B has a higher SSIM, and it has by far the lowest MUSIQ on RealSR and DRealSR (40.38 and 29.75). At the quality point, LDG-LoRA leads every baseline in 21 of the 24 no-reference scores of the six sets, including MANIQA and LIQE on all six (Tables 1 and 4). On ScreenSR, it also has a lower LPIPS than every baseline (.173 against .195). The references of RealSR and DRealSR are less sharp than our outputs (Section 4.4), so added detail counts as error there: PiSA-SR is closer to them than our quality point, but not than LDG-LoRA at PiSA-SR's MUSIQ. In the crops of Figure 4, the baselines rewrite text that LDG-LoRA keeps.

At $g = 1$, SAC, which only rebalances attention with 12,288 trained numbers per pair, comes close to LoRA's fidelity and has a lower LPIPS than every baseline on all three sets with ground truth. Its outputs are softer: at equal MUSIQ, LDG-LoRA is .02–.03 LPIPS closer (Table 5). Guided further, LDG-SAC reaches a higher MUSIQ than every baseline on RealSR (73.63 at $g = 3$). On ScreenSR, its quality point also has a lower LPIPS than every baseline (.171).

4.3 Restoring real-world degradations

The public release of RealIR-Bench (Yang et al., 2026) has 456 real photographs, each with one of nine known degradations and no clean reference. We train one adapter pair per degradation on 100 synthetic triples from Klein's clean images (Appendix F). Each degradation uses one scale $g > 1$ (Table 12), without color correction. Every compared method, except one CoTIR-9B variant, is told the degradation (Appendix F). The benchmark's evaluator asks Qwen3-VL-8B (Bai et al., 2025) how strong the degradation is. RS is the improvement of this rating over the input, clipped at zero. LPS is the LPIPS distance to the input, and $\mathrm{FS} = 0.2\,\mathrm{RS}\,(1 - \mathrm{LPS})$ combines them.

Table 2: **RealIR-Bench** (456 images). Left: FS per degradation; right: means over the degradations of FS, of its two factors LPS (LPIPS to the input) and RS (the rater's removal score), and of three no-reference metrics. LPS is not ranked: an unchanged input has the lowest, and removal raises it. Edit2Restore covers five degradations. **Bold red**: best, underlined blue: second best.

Method	FS ↑ per degradation									Macro average			No-reference, macro		
	Blur	Compr.	Haze	Flare	Low light	Moiré	Noise	Rain	Reflect.	FS ↑	LPS	RS ↑	CLIP-IQA+ ↑	Q-Align ↑	LIQE ↑
RealRestorer (Yang et al., 2026)	.1689	.0583	.0517	**.2952**	**.1611**	.1160	.0572	.1630	.2451	.1463	.4251	1.293	.5337	3.849	2.512
CoTIR-9B, precise prompt (Guo et al., 2026)	.0385	.0344	.0126	.2052	.1364	.0795	.0579	.1743	.2400	.1087	.2486	.801	.4841	3.496	2.239
CoTIR-9B, generic prompt	.0117	.0357	.0000	.0547	.1171	.0470	.0389	.1629	.1250	.0659	.1789	.466	.4627	3.392	2.087
Edit2Restore (Yılmaz et al., 2026)	–	.0173	.0156	–	.1397	–	.0511	.1001	–	–	–	–	–	–	–
LDG-LoRA (ours)	**.2203**	.0648	**.1001**	.2788	.1565	.1316	**.0805**	.1788	**.2979**	**.1677**	.4353	**1.489**	**.6195**	**3.958**	**3.507**
LDG-SAC (ours)	.1767	**.0698**	.0899	.2647	.1368	**.1403**	.0693	**.1949**	.2442	.1541	.4076	1.364	.5924	3.911	3.275

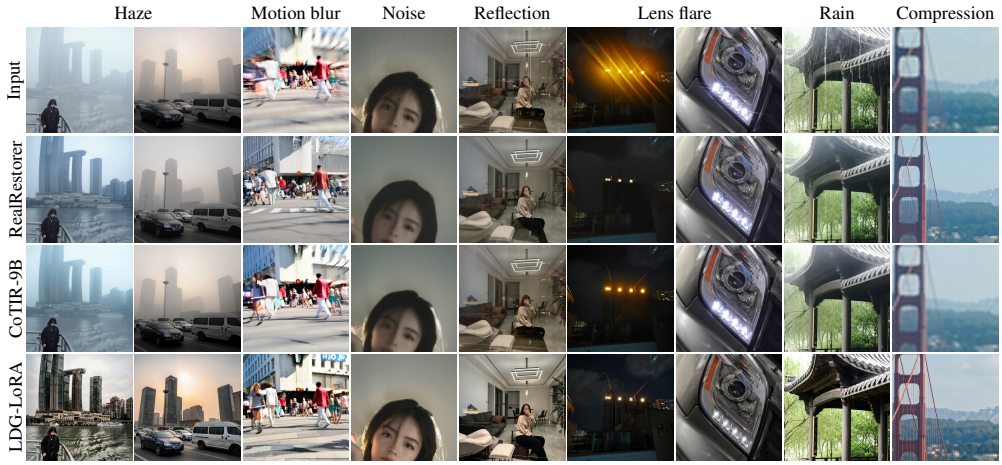


Figure 5: **RealIR-Bench: where the baselines fail** (square crops). RealRestorer (Yang et al., 2026) and CoTIR-9B (Guo et al., 2026) leave haze, the reflection veil, motion blur, rain and compression, or smooth the noisy face. RealRestorer erases most of the cranes with the night flare.

The LoRA pairs have the highest mean FS, followed by the SAC pairs, RealRestorer and CoTIR-9B (Table 2). A paired bootstrap (Efron, 1979) gives a 95% interval of $[+.0134, +.0294]$ for their lead over RealRestorer. The LoRA pairs change the input about as much as RealRestorer, but remove more of the degradation (RS 1.489 against 1.293). The SAC pairs also remove more (RS 1.364), but their FS lead is not significant. Guidance creates the lead: the restoration adapter alone scores .1436. One shared scale of 1.5, 2 or 3 also leads significantly (Appendix F.1).

The LoRA pairs have a higher FS than every baseline on seven of the nine degradations (Table 2, Figure 5). On lens flare and low light, they are second to RealRestorer, by .016 and .005. Both of our pairs also lead in CLIP-IQA+, Q-Align and LIQE (Wang et al., 2023; Wu et al., 2024a; Zhang et al., 2023), three of the no-reference metrics used by CoTIR. They learn from 100 synthetic triples per degradation, while RealRestorer and CoTIR use millions of pairs.

4.4 What the degradation adapter changes

A trained second prediction gives guidance a better direction. We fix the restoration adapter of the LoRA pair and compare three second predictions: the degradation adapter (LDG-LoRA), the restoration adapter under the degradation prompt p_- (*single-adapter CFG*) and the frozen editor under p_-.

On held-out triples, the learned direction points to the clean image. Here x, y and y^- are known (Table 7). Prompted to strengthen the degradation, the frozen editor ends further from y^- than y is (LPIPS .326 against .176): it adds a degradation of its own. The degradation adapters get closer (.115 for LoRA, .119 for SAC). After the first step, the difference of the two clean estimates has a cosine with the true direction $x - y$ about three times higher for LDG-LoRA than for single-

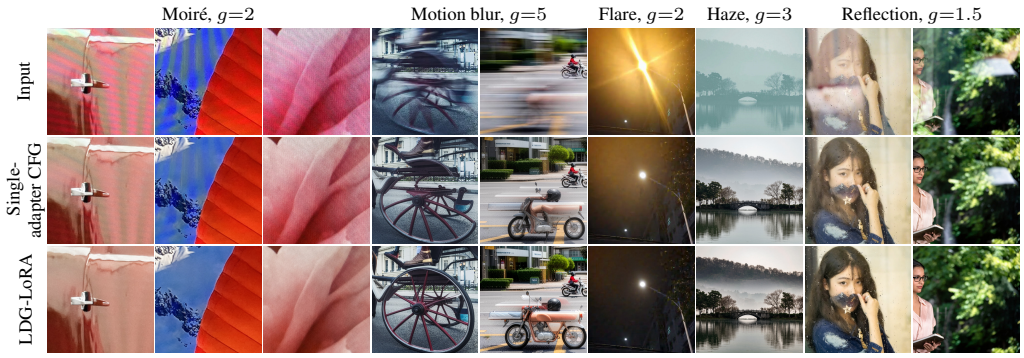


Figure 6: **What the trained degradation adapter removes** (RealIR-Bench). The output rows use the same restoration adapter, prompts and g. Their second prediction is the restoration adapter under the degradation prompt (single-adapter CFG) or the degradation adapter (LDG-LoRA). LDG-LoRA clears moiré, motion smear, glare, haze and reflections that single-adapter CFG leaves.

adapter CFG: .39 against .12 in super-resolution, and .73 against .27 on average over RealIR. LDG-LoRA is the most aligned on all ten held-out sets. Its difference is also 2.6 and 6.6 times longer, respectively, so the same g moves the output further. At $g = 1.5$, its super-resolution samples are both closer to x (LPIPS .151 against .162) and higher in MUSIQ (72.9 against 71.4).

In super-resolution, it is closer to sharp references. On the three sets with ground truth, LDG-LoRA reaches a higher MUSIQ than single-adapter CFG, at $g = 3$ against 6 or 8. With the frozen editor as the second prediction, MUSIQ on RealSR stays below 69 (Table 5). At matched MUSIQ, LDG-LoRA is closer in LPIPS at six of seven points, significantly at three: RealSR at 68 and ScreenSR at 72.5 and 73 (Table 6). Only the ScreenSR references outscore the restoration adapter's output on all four no-reference metrics, so added detail moves the output toward them. The RealSR and DRealSR references score below it (MUSIQ 57.5 and 50.6, against 66.9 and 62.5, Appendix E.2). Added detail moves the output away from them, so single-adapter CFG, which adds less, is up to 0.25 dB higher in PSNR.

On RealIR-Bench, it removes more. At the same g, LDG-LoRA has the higher RS on eight of nine degradations (1.489 against 1.364). It also has the higher CLIP-IQA+ and LIQE. The evaluator rates 196 of its outputs free of the degradation, against 175 (examples in Figure 6). Only FS prefers single-adapter CFG (.1726 against .1677). FS charges every change to the input: removal, and on blur and haze some inpainting-like artifacts (Appendix F.1). On the 311 images with the same positive rating for both, LDG-LoRA changes the input more on 309, so its FS is lower. Where its rating is higher, its FS is higher on 60 of 62. The frozen editor, an untrained second prediction, is the weakest on all four of these metrics.

5 Conclusion

A prompt can only name a degradation. LDG instead learns, from the editor's own images, a second prediction that shows the input's degradation, only stronger. With it, one scale takes a frozen editor from faithful to sharp restoration, and moves the output further than a prompt does. At its most faithful setting, LDG is closer to the ground truth than ten recent super-resolution methods, and it leads a benchmark of nine real degradations, while each adapter learns from about a hundred generated images. A few thousand attention biases nearly match LoRA at $g = 1$.

Limitations. On RealIR-Bench, each degradation type has its own adapter pair, so the type must be known. All compared methods but one CoTIR-9B variant are also given the type. A new type needs a new degradation model, but no real image. Like the baselines in Figure 4, LDG can invent detail, such as sharp but wrong letters, even at $g = 1$ (Appendix E.1).

STATEMENT ON THE USE OF GENERATIVE AI

Generative image modeling is part of the method: the frozen editor generates the synthetic training images. The code was implemented by the authors together with an AI coding assistant. A language-model assistant helped with the writing: it drafted and edited text, suggested improvements and helped correct the figures. It also helped search the literature for related work. The authors checked all AI-assisted code, text and figures and are responsible for everything the paper reports. The FS metric of RealIR-Bench uses a vision-language model, Qwen3-VL-8B (Bai et al., 2025), as its rater, as defined by the benchmark (Yang et al., 2026); we use the benchmark's evaluation protocol unchanged (Appendix F).

ETHICS STATEMENT

Restoration may generate plausible details that were not present in the original scene. Our qualitative analysis includes incorrect text, and the method should not be interpreted as recovering trustworthy evidence from unreadable content. Synthetic adaptation avoids collecting an additional real-image training corpus, but does not remove issues inherited from pretrained models and benchmark data.

REPRODUCIBILITY STATEMENT

Section 3 specifies the training objective of each adapter and the sampling rule, and Algorithm 1 the training loop. Appendix A gives the hyperparameters, instructions and implementation details, Appendix B the construction of the training data, Appendix E the evaluation protocol, the check of our scorer against published numbers and how each baseline was run, and Appendix F the training and evaluation on RealIR-Bench. Every reported number is computed from saved per-image scores, together with the settings that produced each output. On one 96 GB GPU, an adapter trains in about 16 minutes (LoRA) or 26 minutes (SAC) on Klein, and restoring a 512×512 image takes about 0.7 s at $g = 1$ and 1.1 s at $g > 1$. Per image, LDG costs up to 7 evaluations of Klein (9B) or 55 of Kontext (12B), more than the one-step baselines. We will release the code, the scripts that generate the training data, the evaluation code and the trained adapters.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *Computer Vision – ECCV 2024*, volume 15131 of *Lecture Notes in Computer Science*, pp. 1–17. Springer, 2024. doi: 10.1007/978-3-031-73464-9_1.

Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. DreamClear: High-capacity real-world image restoration with privacy-safe dataset curation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 55443–55469, 2024. doi: 10.52202/079017-1761. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6452474601429509f3035dc81c233226-Abstract-Conference.html.

Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. Self-improving diffusion models with synthetic data. *arXiv preprint arXiv:2408.16333*, 2024. URL https://arxiv.org/abs/2408.16333.

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

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.1. URL https://aclanthology.org/2022.acl-short.1/.

Black Forest Labs. FLUX.1 Kontext [dev] model card, 2025. URL https://huggingface.co/black-forest-labs/FLUX.1-Kontext-dev. Official model card, accessed 26 September 2026.

Black Forest Labs. FLUX.2 [klein] 9B model card, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-9B. Official model card, accessed 18 September 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. URL https://arxiv.org/abs/2506.15742.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6228–6237, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html.

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 (ICCV)*, pp. 3086–3095, October 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Cai_Toward_Real-World_Single_Image_Super-Resolution_A_New_Benchmark_and_a_ICCV_2019_paper.html.

Chaofeng Chen and Jiadi Mo. IQA-PyTorch: PyTorch toolbox for image quality assessment, 2022. URL https://github.com/chaofengc/IQA-PyTorch.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vmSEVL19f.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23174–23184, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dong_TSD-SR_One-Step_Diffusion_with_Target_Score_Distillation_for_Real-World_Image_CVPR_2025_paper.html.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S. Ren, Chunle Guo, and Chongyi Li. DiT4SR: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18948–18958, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Duan_DiT4SR_Taming_Diffusion_Transformer_for_Real-World_Image_Super-Resolution_ICCV_2025_paper.html.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept Sliders: LoRA adaptors for precise control in diffusion models. In *Computer Vision – ECCV 2024*, volume 15098 of *Lecture Notes in Computer Science*, pp. 172–188. Springer, 2024. doi: 10.1007/978-3-031-73661-2_10.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S. Ren, and Chao Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision – ECCV 2020*, volume 12356 of *Lecture Notes in Computer Science*, pp. 633–651. Springer, 2020. doi: 10.1007/978-3-030-58621-8_37.

Yu Guo, Zhengru Fang, Shengfeng He, Senkang Hu, Yihang Tao, Phone Lin, and Yuguang Fang. Universal image restoration via internalized chain-of-thought reasoning. *arXiv preprint arXiv:2606.17557*, 2026. URL https://arxiv.org/abs/2606.17557.

Shuhao Han, Wenjie Liao, Hayden Vance, Hang Dong, Rui Zhang, Chun-Le Guo, and Chongyi Li. DNF-SR: Dual-input and negative-aware feature fine-tuning for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38176–38186, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Han_DNF-SR_Dual-Input_and_Negative-Aware_Feature_Fine-Tuning_for_Real-World_Image_Super-Resolution_CVPR_2026_paper.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7462–7471, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hong_Improving_Sample_Quality_of_Diffusion_Models_Using_Self-Attention_Guidance_ICCV_2023_paper.html.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1510–1519, October 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, Harry Yang, and Steven Hoi. D-OPSD: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026. URL https://arxiv.org/abs/2605.05204.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52996–53021, 2024. doi: 10.52202/079017-1679. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5ee7ed60a7e8169012224dec5fe0d27f-Abstract-Conference.html.

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 (ICCV)*, pp. 5148–5157, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ke_MUSIQ_Multi-Scale_Image_Quality_Transformer_ICCV_2021_paper.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 122458–122483, 2024. doi: 10.52202/079017-3892. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/dd540e1c8d26687d56d296e64d35949f-Abstract-Conference.html.

Zhicheng Liao, Dongxu Wu, Zhenshan Shi, Sijie Mai, Hanwei Zhu, Lingyu Zhu, Yuncheng Jiang, and Baoliang Chen. Beyond cosine similarity: Magnitude-aware CLIP for no-reference image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6934–6942, 2026. doi: 10.1609/aaai.v40i9.37627. URL https://ojs.aaai.org/index.php/AAAI/article/view/37627.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Computer Vision – ECCV 2024*, volume 15117 of *Lecture Notes in Computer Science*, pp. 430–448. Springer, 2024. doi: 10.1007/978-3-031-73202-7_25.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S. Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Transactions on Graphics*, 44(6):208:1–208:21, 2025. doi: 10.1145/3763346.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1950–1965, 2022. doi: 10.52202/068431-0142. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0cde695b83bd186c1fd456302888454c-Abstract-Conference.html.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1X-Edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. URL https://arxiv.org/abs/2504.17761.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K. Wong. Rethinking cross-modal interaction in multimodal diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5934–5943, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lv_Rethinking_Cross-Modal_Interaction_in_Multimodal_Diffusion_Transformers_ICCV_2025_paper.html.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Meng_On_Distillation_of_Guided_Diffusion_Models_CVPR_2023_paper.html.

Federico Putamorsi, Leonardo Zini, Marcella Cornia, and Lorenzo Baraldi. SPARK: Input-conditioned sparse activation modulation for frozen DiT-based super-resolution. *arXiv preprint arXiv:2609.03813*, 2026. URL https://arxiv.org/abs/2609.03813.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.

Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-LoRA approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2333–2343, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_Pixel-level_and_Semantic-level_Adjustable_Super-resolution_A_Dual-LoRA_Approach_CVPR_2025_paper.html.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14940–14950, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Tan_OminiControl_Minimal_and_Universal_Control_for_Diffusion_Transformer_ICCV_2025_paper.html.

Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-NPO: Negative preference optimization for better preference aligned generation of diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iJi7nz5Cxc.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353. URL https://ojs.aaai.org/index.php/AAAI/article/view/25353.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024a. doi: 10.1007/s11263-024-02168-7.

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 (ICCV) Workshops*, pp. 1905–1914, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021W/AIM/html/Wang_Real-ESRGAN_Training_Real-World_Blind_Super-Resolution_With_Pure_Synthetic_Data_ICCVW_2021_paper.html.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25796–25805, June 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_SinSR_Diffusion-Based_Image_Super-Resolution_in_a_Single_Step_CVPR_2024_paper.html.

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. doi: 10.1109/TIP.2003.819861.

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*, volume 12353 of *Lecture Notes in Computer Science*, pp. 101–117. Springer, 2020. doi: 10.1007/978-3-030-58598-3_7.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 2024a. URL https://proceedings.mlr.press/v235/wu24ah.html.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92529–92553, 2024b. doi: 10.52202/079017-2938. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a8223b0ad64007423ffb308b0dd92298-Abstract-Conference.html.

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 (CVPR)*, pp. 25456–25467, June 2024c. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_SeeSR_Towards_Semantics-Aware_Real-World_Image_Super-Resolution_CVPR_2024_paper.html.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. VOSR: A vision-only generative model for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16311–16321, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_VOSR_A_Vision-Only_Generative_Model_for_Image_Super-Resolution_CVPR_2026_paper.html.

Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. DiffFit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4230–4239, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Xie_DiffFit_Unlocking_Transferability_of_Large_Diffusion_Models_via_Simple_Parameter-efficient_ICCV_2023_paper.html.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935, 2023. doi: 10.52202/075280-0700. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Yu Xu, Fan Tang, Juan Cao, Xiaoyu Kong, Yuxin Zhang, Jintao Li, Oliver Deussen, and Tong-Yee Lee. HeadRouter: A training-free image editing framework for MM-DiTs by adaptively routing attention heads. *ACM Transactions on Graphics*, 45(3):27:1–27:14, 2026. doi: 10.1145/3797956.

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 (CVPR) Workshops*, pp. 1191–1200, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Yang_MANIQA_Multi-Dimension_Attention_Network_for_No-Reference_Image_Quality_Assessment_CVPRW_2022_paper.html.

Yufeng Yang, Xianfang Zeng, Zhangqi Jiang, Fukun Yin, Jianzhuang Liu, Wei Cheng, Jinghong Lan, Shiyu Liu, Yuqi Peng, Gang Yu, and Shifeng Chen. RealRestorer: Towards generalizable real-world image restoration with large-scale image editing models. *arXiv preprint arXiv:2603.25502*, 2026. URL https://arxiv.org/abs/2603.25502.

Mustafa Akın Yılmaz, Ahmet Bilican, Burak Can Biner, and Ahmet Murat Tekalp. Edit2Restore: Few-shot image restoration via parameter-efficient adaptation of pre-trained editing models. *arXiv*

preprint arXiv:2601.03391, 2026. URL https://arxiv.org/abs/2601.03391. Accepted to the LoViF Workshop at ECCV 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 (CVPR)*, pp. 25669–25680, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yu_Scaling_Up_to_Excellence_Practicing_Model_Scaling_for_Photo-Realistic_Image_CVPR_2024_paper.html.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23153–23163, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yue_Arbitrary-steps_Image_Super-resolution_via_Diffusion_Inversion_CVPR_2025_paper.html.

Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024a. URL https://arxiv.org/abs/2409.17058.

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 (ICCV)*, pp. 4791–4800, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhang_Designing_a_Practical_Degradation_Model_for_Deep_Blind_Image_Super-Resolution_ICCV_2021_paper.html.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. LLaMA-Adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=d4UiXAHN2W.

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 (CVPR)*, pp. 586–595, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Blind_Image_Quality_Assessment_via_Vision-Language_Correspondence_A_Multitask_Learning_CVPR_2023_paper.html.

A IMPLEMENTATION DETAILS

Editors and token layout. FLUX.2 [klein] 9B (Black Forest Labs, 2026) has eight double-stream transformer blocks followed by 24 single-stream blocks, with 32 attention heads of width 128. A 512×512 image gives $32 \times 32 = 1{,}024$ image tokens after the autoencoder and latent packing, so the joint attention sequence is [512 text; 1,024 target; 1,024 reference]. Target and reference tokens share their horizontal and vertical position identifiers and differ in their image identifier. The reference latent is the mode of the frozen encoder's posterior, and decoded RGB values are clipped to $[0, 1]$. FLUX.1 Kontext [dev] (Black Forest Labs et al., 2025) has 19 double-stream and 38 single-stream blocks with 24 heads, uses T5 (Raffel et al., 2020) tokens as its text source, and has an embedded guidance input, which we hold at 2.5, the value of its model card (Black Forest Labs, 2025), in both adapters. This embedded input is part of the frozen model's conditioning and is distinct from the scale g. Table 11 compares the two editors.

Sampler. Klein's scheduler takes the four-point grid $(1, 3/4, 1/2, 1/4)$ and applies its resolution-dependent exponential shift (Esser et al., 2024); at 512×512 the noise levels are approximately $(1, 0.9581, 0.8840, 0.7175, 0)$. With K sampling steps, LDG needs K transformer evaluations at $g = 1$ and $2K - 1$ at $g > 1$, since the last step uses the restoration adapter alone. Kontext uses its pretrained 28-step schedule. All reported restorations use noise seed 1234.

LoRA. We use rank $r = 16$, scaling $\alpha = r$, Gaussian initialization of A and zero initialization of B, so that the initial update BA is zero (Hu et al., 2022). On Klein, the adapters target the query, key, value and output projections of the image stream in the double-stream blocks, and the fused query–key–value–MLP projection and the output projection of the single-stream blocks; the text-stream projections of the double-stream blocks are not adapted. An adapter has 27,787,264 parameters. On Kontext, the adapters also cover the text-stream attention projections and both feed-forward networks of the double-stream blocks and the attention and MLP input projections of the single-stream blocks: 54,165,504 parameters. The frozen transformer runs in bfloat16.

Source-attention calibration. Let $r_i = 1$ for target queries and 0 otherwise. In each layer and head at step k, we append $r_i\sqrt{d}$ to the query and $\theta_{\ell h, s(j)}(\sigma_k)$ to the key, and pad the value with zero. The scaled dot product then adds exactly $r_i\, \theta_{\ell h, s(j)}(\sigma_k)$ to the original logit, and a single fused attention call computes Equation (4) without materializing an attention-bias tensor. The standardized noise level is $\tilde{\sigma} = (\sigma - \mu_\sigma)/\delta_\sigma$, where μ_σ and δ_σ are the mean and population standard deviation of the training schedule's noise levels: on Klein, the four levels at 512×512 give $\mu_\sigma = 0.8899$ and $\delta_\sigma = 0.1079$, and a noise level outside their range $[0.7175, 1]$ is first clamped to it. This clamp only affects the last step at some image sizes (90 of the 200 RealLR200 images and 151 of the 456 RealIR images); that step is run by the restoration adapter alone, whose SAC biases it changes slightly. On Kontext the statistics come from its 28 training noise levels (mean 0.6213, standard deviation 0.2745, range $[0.0650, 1]$). Adding the same constant to all logits of a query leaves the softmax unchanged, so at every noise level the three biases of a head carry one redundant degree of freedom; because the biases are affine in $\tilde{\sigma}$, θ^0 and θ^1 each carry one. Klein's 6,144 parameters per adapter therefore have 4,096 effective degrees of freedom (Kontext: $57 \cdot 24 \cdot 3 \cdot 2 = 8{,}208$ and 5,472). We impose no constraint for this redundancy. Both coefficients are trained from zero, and both are used: over the 20 Klein SAC adapters of the super-resolution and RealIR pairs, the largest $|\theta^1|$ is 8.4–14.1 and the largest $|\theta^0|$ is 10.1–18.9.

Loss and optimization. The perceptual weight in Equation (3) is $\lambda = 0.308$, the ratio of the gradient norms of the two terms, $\|\nabla \mathrm{MSE}\|/\|\nabla \mathrm{LPIPS}\|$ over the trained parameters, measured at the first training step of a SAC adapter at initialization (one training triple). It is used unchanged for every adapter, both editors and all degradation types. Training LPIPS uses VGG features; the reported LPIPS uses AlexNet features. Adam (Kingma & Ba, 2015) uses a learning rate of 4×10^{-4} for LoRA and 0.1 for SAC, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$ and no weight decay. LoRA adapters train for 1,000 updates and SAC adapters for 2,000, except the RealIR LoRA adapters (Appendix F). Each update draws one training triple, uniformly with replacement.

Parameter averaging. The saved adapter is the arithmetic mean of the parameters at the last ten multiples of 50 updates: updates $550, 600, \ldots, 1000$ for LoRA and $1550, \ldots, 2000$ for SAC. For



Clean image x | Input y | More degraded target y^-

Figure 7: **One training triple generated by the editor.** The frozen FLUX.2 [klein] 9B renders the clean image x. One random draw of ringing and JPEG compression on the 128×128 degradation grid gives the input y. The same draw, with the cutoff of the ringing filter lowered by a fixed step (from 2.27 to 1.07) and the JPEG quality by a fixed step (from 84 to 54), gives the degradation adapter's target y^-.

LoRA, A and B are averaged separately, so the saved update $\bar{B}\bar{A}$ is one adapter, not an ensemble of predictions. The RealIR LoRA adapters are the exception: each chooses its number of updates on its task's held-out triples (Appendix F).

Instructions. The restoration adapter's instruction for super-resolution is *"Make this image an ultra high quality, 8K resolution photograph with sharp focus, crisp edges, crystal clear fine details and rich, natural, realistic textures. Strictly preserve the exact same content, composition, perspective, object placement, spatial layout, original colors, tone, and lighting. Change nothing else."* The degradation adapter's instruction is *"Make this image lower quality: apply the degradations it already has more strongly, in exactly the same places, so that it looks even worse."* followed by the same two preservation sentences. The phrase "8K resolution" is part of the instruction and does not set the output size. The RealIR instructions are given in Appendix F.

Image size and color correction. For $4\times$ benchmarks, the output size is four times the input size, with a minimum input short side of 128 and dimensions rounded down to multiples of 16; the input is resized once, directly to this size. For inputs restored at their own size, the input is reflect-padded to a multiple of 16 and the output cropped back. For super-resolution, AdaIN (Huang & Belongie, 2017) matches the mean and standard deviation of each RGB channel of the output to those of the input, before cropping. RealIR outputs are not color-corrected, because several of its degradations, such as low light and haze, require changing brightness and color. Ground truth is never resized to fit an output.

B Training data

Clean images. For each of the 100 training images, the frozen editor renders a 1024×1024 image, which is reduced to 512×512 with Lanczos filtering. The 50 held-out images use a separate list of subjects. Each generation prompt combines a subject, a lighting description and a camera description, followed by *"Tack sharp across the entire frame, crisp fine detail and clean edges everywhere, deep depth of field, no motion blur, no defocus, no haze, no fog, no glow."* Figure 7 shows one triple. Kontext renders its own set in the same way with its own sampler, and the RealIR sets are built on Klein's clean images (Appendix F). The images are synthetic, but they carry the visual knowledge of the pretrained editor.

Degradations. We first reduce x to 128×128 with a filter chosen uniformly among box, bilinear and bicubic. Degradation groups are activated independently with the probabilities of Table 3, and a draw with no active group is rejected. Active operations are applied in a fixed order (blur, ringing, resolution, noise, compression), and the result is upsampled to 512×512 with bicubic interpolation. The input y and the target y^- share the initial downsampling, the active groups, the operation types,

Table 3: Degradation draws for the input y. Subtype probabilities are conditional on their group being active; continuous ranges are sampled uniformly. The target y^- uses the same draw with the fixed, capped increments given in the text.

Group	Probability	Types and parameter ranges
Blur	0.50	Generalized Gaussian (Wang et al., 2021) (0.50): $\sigma_x, \sigma_y \in [0.2, 0.6]$, shape $\beta \in [0.5, 3]$ (the applied shape is drawn uniformly between β and 1 with probability 0.5 and otherwise lies within 0.02 of 1), rotation $[-90, 90]$ degrees; isotropic with probability 0.5. Motion (0.15): length $[3, 4]$, angle $[0, 360]$, direction $[-1, 1]$. Defocus (0.15): radius $[1, 1.4]$, alias blur $[0.1, 0.3]$. Zoom (0.20): factor $[1.01, 1.03]$.
Ringing	0.15	Sinc (Wang et al., 2021) cutoff $[2.2, 2.9]$ radians; odd kernel size from $[5, 9]$.
Resolution	0.15	Inner resize factor $[0.85, 0.97]$; down and up filters drawn independently from box, bilinear, bicubic.
Noise	0.50	Gaussian (0.35): standard deviation $[0.004, 0.020]$. ISO (0.25): color shift $[0.005, 0.015]$, intensity $[0.05, 0.25]$. Shot (0.25): scale $[0.001, 0.008]$. Multiplicative (0.15): deviation $[0.02, 0.06]$.
Compression	0.50	JPEG (0.70) or WebP (0.30), quality $[82, 96]$.

the parameter draws and the random seed of every operation; both start from the downsampled clean image and are processed separately.

The more degraded target. The target y^- moves each blur, ringing, resolution and compression parameter of the accepted draw by a fixed step in its own unit, with a cap (pixels are on the 128×128 degradation grid): Gaussian blur σ_x, σ_y $+0.5$ px (at most 1.2), motion length $+2$ taps (at most 7), defocus radius $+1$ px (at most 3), zoom factor $+0.03$ (at most 1.08), ringing cutoff -1.2 rad (at least 1.0), inner resize factor $\times 0.75$ (at least 0.5) and compression quality -30 (at least 20). Noise receives no increment: its type, amplitude and random seed are identical in both paths, because noise that cannot be predicted from the input would be averaged away by a regression loss. Motion lengths are rounded to odd integers, defocus radii to positive integers, and the quality is floored at 10. The inner resize factor of both paths is then clipped to $[0.15, 0.95]$, so the factor of y is also at most 0.95. A fixed step does not give a fixed loss of quality: on the training set, y^- is 0.50–5.54 dB (mean 1.58 dB) below y in RGB PSNR to the clean image. Because y^- is computed from the clean image with the same draw, not from y, it is predictable from y without being a deterministic function of it.

Acceptance. A draw is accepted only if y is at least 0.8 dB worse than a plain down- and up-sampling of the clean image and y^- is at least another 0.5 dB worse than y, in RGB PSNR to the clean image. Up to 30 attempts are allowed; from attempt 11 on, blur and compression are forced on and the base magnitudes are escalated. No accepted draw of the super-resolution sets used the escalation. The RealIR sets use their own acceptance rule (Appendix F). For two Kontext renders no draw was accepted in 30 attempts, which leaves Kontext with 98 training triples.

C Local analysis of guidance, attention, and training

This appendix states the algebraic properties behind the method. They describe the implemented operations; they do not assume that the trained adapters recover an exact degradation process.

C.1 What the guidance direction changes

Fix the condition, prompts, noise level $\sigma > 0$, and current target latent z. Write

$$\hat{a}_\pm = z - \sigma v_\pm(z), \qquad v_g = g v_+(z) + (1 - g) v_-(z). \tag{5}$$

All predictions here are evaluated at the *same* state.

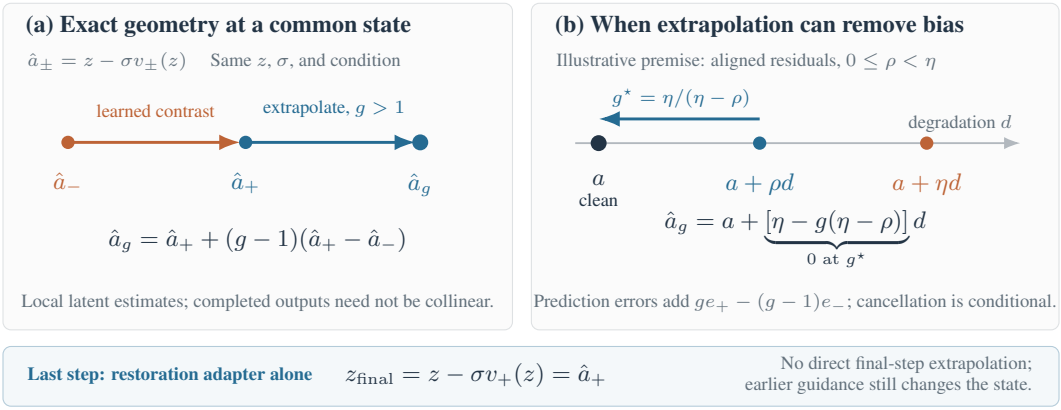


Figure 8: **What paired guidance does, and when it can help.** (a) At the same latent state, affine velocity guidance is exactly affine guidance of the clean estimates. The drawing uses $g = 2$ for visibility; the equation holds for any g. (b) Under the stated one-direction bias model and zero prediction errors, an appropriate scale can cancel the residual degradation of the restoration estimate. This premise is illustrative; prediction errors can be amplified and excessive guidance can overshoot. The final-step identity is exact at its current state. All positions are schematic.

Property 1: affine guidance of the local clean estimate. The corresponding clean estimate satisfies

$$\hat{a}_g = z - \sigma v_g = g\hat{a}_+ + (1-g)\hat{a}_- = \hat{a}_+ + (g-1)(\hat{a}_+ - \hat{a}_-). \tag{6}$$

This follows by substitution: the coefficients of z sum to one. Figure 8 draws this identity and the example below. Thus, for $g > 1$, LDG extrapolates beyond the restoration estimate, away from the degradation estimate. The sign remains the same under the decreasing-noise Euler convention. In fact, with $h = \sigma - \sigma_{\text{next}} > 0$, the difference from the Euler update of the restoration adapter alone at this same state is

$$z_{\text{next},g} - z_{\text{next},+} = \frac{h}{\sigma}(g-1)(\hat{a}_+ - \hat{a}_-). \tag{7}$$

Neither identity expresses the completed guided output as a linear combination of separately sampled outputs: after an update, the trajectories generally visit different states. Decoding and color correction are also nonlinear.

Correction and error amplification. Let a be an ideal clean latent and suppose, for this local analysis, that the two estimates can be written as $\hat{a}_+ = a + e_+$ and $\hat{a}_- = a + d + e_-$, where d is a degradation displacement and $e_\pm$ are prediction errors. Equation (6) gives exactly

$$\hat{a}_g = a - (g-1)d + ge_+ - (g-1)e_-. \tag{8}$$

Guidance introduces the opposite displacement to d, while also extrapolating the prediction errors. For $g \ge 1$, any norm obeys

$$\left\|\hat{a}_g - [a - (g-1)d]\right\| \le g\|e_+\| + (g-1)\|e_-\|. \tag{9}$$

In particular, if the restoration estimate is already exact and both errors vanish, $g > 1$ moves *past* the clean latent. A larger guidance value therefore cannot guarantee better fidelity.

A simple conditional example explains when this extrapolation can remove a residual bias. Suppose instead that

$$\hat{a}_+ = a + \rho d + e_+, \qquad \hat{a}_- = a + \eta d + e_-, \qquad 0 \le \rho < \eta. \tag{10}$$

The restoration estimate retains less of the same degradation direction than the degradation estimate. Substitution yields

$$\hat{a}_g = a + [\eta - g(\eta - \rho)]d + ge_+ - (g-1)e_-. \tag{11}$$

The directional term vanishes at $g^\star = \eta/(\eta - \rho)$, which exceeds one when $\rho > 0$. This is an explanatory model, not an assumption enforced by our loss: the errors of the two adapters need not align, and real degradations need not occupy one latent direction. It motivates measuring the whole guidance curve and checking that the degradation adapter matches its target, rather than choosing the largest scale.

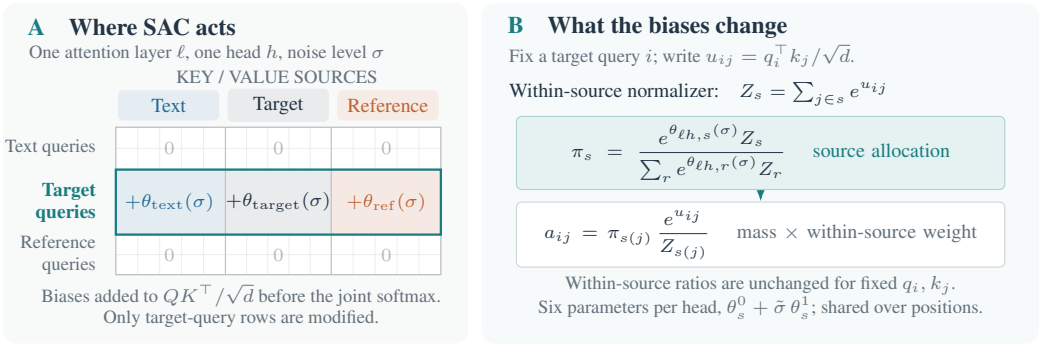


Figure 9: **Source-attention calibration (SAC). A:** In each attention layer and head, SAC adds one learned bias per key source (text, target, reference) to the attention logits of the target queries; text and reference queries are unchanged. Each bias is affine in the standardized noise level $\tilde{\sigma}$. **B:** For a given query, the attention splits exactly into the total weight π_s of each source and the weights within that source. The biases change only π_s; the weights within a source are those of the frozen model. On FLUX.2 [klein] 9B, 32 layers of 32 heads give 6,144 parameters per adapter; zero biases recover the frozen editor.

Why the last update uses the restoration adapter alone. At the last Euler step, $\sigma_{\text{next}} = 0$ and $h = \sigma$. Consequently, the update of the restoration adapter alone is exactly $z_{\text{final}} = \hat{a}_+$ at the current guided state. Guiding this update would instead produce $\hat{a}_g$, retaining the full local extrapolation term in Eq. (6). Our final-step rule removes that *direct final-step* term. It does not undo state changes or errors introduced by earlier guidance; its measured effect is in Table 5.

C.2 Source-attention calibration as a mixture over existing attention sources

Figure 9 summarizes where SAC acts and what it changes.

Consider one target query in one layer and head at one sampling step k. Suppress these indices, write θ_s for $\theta_{\ell h,s}(\sigma_k) = \theta^0_{\ell h,s} + \tilde{\sigma}_k\,\theta^1_{\ell h,s}$, fix the queries and keys, and define the original logit $u_j = q^\top k_j/\sqrt{d}$. Let J_s be the nonempty key set of source s, and define

$$Z_s = \sum_{j\in J_s} \exp(u_j), \qquad \pi_s(\theta) = \sum_{j\in J_s} a_j(\theta). \tag{12}$$

Property 2: exact separation of source mass and within-source attention. The biased attention factorizes as

$$a_j(\theta) = \pi_{s(j)}(\theta)\,\frac{\exp(u_j)}{Z_{s(j)}}, \qquad \pi_s(\theta) = \frac{\exp(\theta_s + \log Z_s)}{\sum_t \exp(\theta_t + \log Z_t)}. \tag{13}$$

To verify this, group the softmax denominator by source and factor out $\exp(\theta_s)$ within each group. The second factor is the original attention distribution conditioned on the source. SAC changes only the source masses at this fixed query and fixed keys; it preserves within-source token ratios. At a given step the learned biases are fixed numbers, and across steps they change affinely with the noise level; in both cases the resulting masses remain input- and state-dependent through Z_s.

For two sources, their attention-mass log odds obey

$$\log\frac{\pi_s(\theta)}{\pi_t(\theta)} = \theta_s - \theta_t + \log\frac{Z_s}{Z_t}. \tag{14}$$

Thus, $\theta_s - \theta_t$ has an exact interpretation as the learned change in log odds relative to the unbiased logits at these fixed query/key features. Adding the same constant to all three biases leaves every weight unchanged; each head therefore has two source-relative degrees of freedom at each noise level and, since θ^0 and θ^1 each carry one redundant common shift, four across the schedule.

Algorithm 1 Training one adapter of LDG

Require: Frozen editor and decoder D; triples; adapter $b \in \{+, -\}$ with target $t_+ = x$ or $t_- = y^-$; schedule $\{\sigma_j\}_{j=0}^{K}$

Initialize the adapter parameters ϕ_b so that the editor is unchanged
for each training iteration **do**
  Sample (x, y, y^-), $k \sim \mathrm{Uniform}\{0, \dots, K-1\}$, $z_0 \sim \mathcal{N}(0, I)$
  for $j = 0, \dots, k-1$ **do** ▷ gradients disabled
    $z_{j+1} \leftarrow z_j + (\sigma_{j+1} - \sigma_j) v_{\phi_b}(z_j, \sigma_j; y, p_b)$
  end for
  $\hat{x}_b \leftarrow D\big(\mathrm{sg}(z_k) - \sigma_k v_{\phi_b}(\mathrm{sg}(z_k), \sigma_k; y, p_b)\big)$
  Update ϕ_b by Adam on $\mathrm{MSE}(\hat{x}_b, t_b) + \lambda\,\mathrm{LPIPS}(\hat{x}_b, t_b)$ (Eq. 3)
end for
return Mean of ϕ_b at the last ten multiples of 50 updates (LoRA: $550, \dots, 1000$; SAC: $1550, \dots, 2000$; Appendix A)

The local sensitivity is also explicit:

$$\frac{\partial \pi_s}{\partial \theta_t} = \pi_s(\mathbf{1}[s = t] - \pi_t). \tag{15}$$

For fixed value vectors, let $\mu_s = \sum_{j \in J_s} \exp(u_j) V_j / Z_s$ be the source-conditioned value average. The head output and its derivative are

$$o(\theta) = \sum_s \pi_s(\theta)\mu_s, \qquad \frac{\partial o}{\partial \theta_t} = \pi_t(\mu_t - o). \tag{16}$$

The trained parameters enter through $\theta_t = \theta_t^0 + \tilde{\sigma}_k \theta_t^1$, so $\partial/\partial\theta_t^0 = \partial/\partial\theta_t$ and $\partial/\partial\theta_t^1 = \tilde{\sigma}_k\,\partial/\partial\theta_t$. Each bias moves the current mixture toward or away from an existing source summary. SAC does not learn a new value representation or prescribe a semantic role for a head. These statements are local: changing an earlier layer changes later hidden states, and hence their queries, keys, and within-source attention. They do not assert that within-source attention is invariant across complete adapted and frozen networks.

What the trained biases do. The biases relative to the mean of the three sources are gauge-free, so they can be compared across heads and adapters. For the super-resolution restoration adapter on Klein, their average over all layers and heads is $+0.46$, $+0.40$, $+0.30$ and $+0.06$ for the input image and -0.37, -0.34, -0.29 and -0.17 for the generated image at the four noise levels $\sigma = 1$, 0.958, 0.884 and 0.717: relative to the frozen editor, the calibration favors the input image over the generated image at every noise level, most strongly at the start of sampling. All 20 trained SAC adapters on Klein (the super-resolution pair and the nine RealIR pairs) show the same direction on average: the head-averaged slope of the input-image bias with respect to $\tilde{\sigma}$ is positive in all 20 ($+0.055$ to $+0.244$), and that of the generated-image bias negative in all 20 (-0.266 to -0.028). Individual heads differ, and about half of them move each way; the trend is a property of the average, not of every head.

C.3 The derivative used in training

Let ξ collect a sampled training triple, initial Gaussian noise, and step index k. For one adapter, let $R_\phi(\xi)$ be the state obtained after its k rollout updates, and let $\ell(\phi, z; \xi)$ be the decoded clean-estimate loss evaluated at state z. A full state-dependent objective would be

$$J(\phi) = \mathbb{E}_\xi\, \ell(\phi, R_\phi(\xi); \xi). \tag{17}$$

Where derivatives exist and differentiation can pass through the expectation, its derivative contains two terms:

$$\nabla_\phi J = \mathbb{E}_\xi \left[\partial_\phi \ell + (D_\phi R_\phi)^\top \nabla_z \ell\right]. \tag{18}$$

Our training uses only the first term. More precisely, with the current parameters $\bar{\phi}$ held fixed while states are sampled, define

$$\widetilde{J}(\phi; \bar{\phi}) = \mathbb{E}_\xi\, \ell(\phi, R_{\bar{\phi}}(\xi); \xi). \tag{19}$$

The sampled update estimates $\nabla_\phi \widetilde{J}(\phi; \bar{\phi})$ at $\phi = \bar{\phi}$. The rollout parameters are refreshed for the next update. This is a *semi-gradient* in the sense of Sutton & Barto (2018, Section 9.3): it keeps only part of the gradient. There, the dependence of the target on the parameters is ignored. Here, the dependence of the state is ignored: states come from the current adapter, while gradients through their construction are stopped. ReFL (Xu et al., 2023) and DRaFT with one differentiated step (Clark et al., 2024) fine-tune diffusion models in the same way. At $k = 0$, the state is independent Gaussian noise and the omitted term vanishes; at later steps it generally does not. The rule supervises the reconstruction behavior at visited states without storing the rollout's gradient graph. It is neither full differentiation through the sampler nor optimization of the coupled guided trajectory at $g > 1$.

C.4 Requirements for another editing backbone

The LDG combination applies to an editor with a velocity parameterization, a shared latent state and schedule for its two adapters, and an image-conditioning path. The clean-estimate formulas above additionally use the linear-flow convention $\hat{a} = z - \sigma v$; another prediction convention requires the corresponding conversion. SAC as defined here requires target queries that jointly attend to identifiable reference, text, and target keys under a common softmax normalization. A constant bias within a separate, single-source softmax would cancel and have no effect. Its parameter count is $6LH$ per adapter for L such layers and H heads ($4LH$ effective; Appendix A). Changing the backbone also changes its tokenization, positional convention, sampler, and suitable adapter locations. FLUX.1 Kontext [dev], the second editor of Section 4, satisfies these requirements in all 57 of its blocks.

D Extended related work

Guidance away from a degraded version of the model. Self-attention guidance (Hong et al., 2023) obtains the second prediction without training by blurring the regions of the intermediate sample that the model attends to, and perturbed-attention guidance (Ahn et al., 2024) by replacing selected self-attention maps of the model with the identity; both guide away from the resulting degraded prediction, and perturbed-attention guidance also helps inverse problems such as deblurring. In LDG the second prediction is trained, on a target that exaggerates the input's own degradation. Applying guidance only within a limited interval of noise levels improves sample quality (Kynkäänniemi et al., 2024). That interval excludes the highest and the lowest noise levels. LDG keeps guidance at high noise and leaves only its last step unguided, for the reason given in Appendix C.1. SUPIR adds generated low-quality images to its training data so that a low-quality negative prompt becomes meaningful (Yu et al., 2024).

Controlling fidelity and quality. Super-resolution methods expose this trade-off through a feature-fusion weight (StableSR (Wang et al., 2024a)), guidance toward a first restoration or the input (DiffBIR (Lin et al., 2024), SUPIR (Yu et al., 2024)), the starting step (InvSR (Yue et al., 2025)), injected noise (HYPIR (Lin et al., 2025)), separate pixel- and semantic-level LoRA scales (PiSA-SR (Sun et al., 2025)) or a guidance scale (S3Diff (Zhang et al., 2024a), VOSR (Wu et al., 2026)); Concept Sliders scale one LoRA to control an attribute (Gandikota et al., 2024).

Training on the model's own samples. Our adapters are trained on their own sampling trajectories, with one supervised step and gradients stopped through the rollout, as in ReFL (Xu et al., 2023), which DRaFT (Clark et al., 2024) treats as a special case of its general reward fine-tuning algorithm. Self Forcing (Huang et al., 2025) and D-OPSD (Jiang et al., 2026) also train on their own rollouts; unlike them, our loss compares the output directly with a target image, so we need no reward model or teacher.

Compact adaptation. Besides LoRA (Hu et al., 2022), small adapters train only bias terms (Ben Zaken et al., 2022) or rescale activations (Liu et al., 2022); SPARK modulates selected channels of frozen super-resolution transformers (Putamorsi et al., 2026). Closer to SAC, Omini-Control adds a user-set constant to the attention logits between image and condition tokens (Tan et al., 2025), TACA rescales text–image attention logits with a timestep-dependent temperature (Lv et al., 2025), and HeadRouter reweights attention heads for editing without training (Xu et al., 2026). Learned biases on attention logits also appear inside pretrained models, such as the relative position biases of T5 (Raffel et al., 2020), which depend on the distance between tokens. SAC learns such

Table 4: **Blind super-resolution on three real-world benchmarks without ground truth**, no-reference metrics, with the operating points of Table 1. **Bold red**: best, underlined blue: second best in each row.

		Related work									Ours on Klein 9B				Ours on Kontext			
											LDG-LoRA		LDG-SAC		LDG-LoRA		LDG-SAC	
Set	Metric	PiSA-SR	TSD-SR	OSEDiff	InvSR	VOSR	HYPIR	SinSR	DiT4SR	CoTIR	g=1	g=2.5	g=1	g=1.5	g=1	g=1.5	g=1	g=1.5
Real Photo60	MUSIQ ↑	73.52	75.03	71.33	71.90	73.52	73.15	69.02	73.15	57.74	71.51	74.85	71.57	74.63	70.44	74.43	70.00	**75.76**
	MANIQA ↑	.688	.674	.651	.676	.681	.693	.595	.684	.549	.668	.716	.681	**.725**	.639	.683	.630	.704
	CLIP-IQA ↑	.781	.796	.751	.763	.683	.780	.748	.781	.571	.721	.838	.726	.832	.717	.837	.692	**.846**
	LIQE ↑	4.76	4.79	4.62	4.61	4.58	4.68	4.11	4.68	3.18	4.59	**4.83**	4.57	**4.83**	4.45	**4.83**	4.34	4.82
Real LQ250	MUSIQ ↑	71.25	73.22	69.56	65.54	71.13	68.49	65.09	71.60	36.55	69.38	**74.78**	62.84	73.70	68.89	74.03	66.96	74.77
	MANIQA ↑	.605	.592	.578	.572	.612	.604	.526	.631	.361	.572	**.668**	.534	.662	.547	.621	.510	.627
	CLIP-IQA ↑	.705	.737	.672	.658	.616	.684	.709	.713	.329	.621	**.784**	.536	.758	.623	.771	.571	.780
	LIQE ↑	4.03	4.32	3.90	3.66	3.91	3.97	3.32	4.35	1.26	3.77	**4.66**	3.15	4.50	3.62	4.51	3.27	4.63
Real LR200	MUSIQ ↑	71.13	72.33	69.55	69.28	69.60	68.38	64.95	70.63	44.63	69.71	73.94	66.58	73.00	69.03	73.25	67.57	**73.99**
	MANIQA ↑	.629	.612	.602	.625	.621	.632	.546	.641	.434	.615	**.694**	.599	.691	.582	.643	.555	.657
	CLIP-IQA ↑	.711	.739	.674	.675	.599	.699	.707	.701	.410	.632	**.791**	.589	.773	.633	.769	.599	.789
	LIQE ↑	4.19	4.30	4.05	4.02	3.87	4.04	3.56	4.31	1.90	3.92	**4.70**	3.51	4.57	3.81	4.53	3.55	**4.70**

biases as an adapter, per layer, head and token source and affine in the noise level, and has no other parameters. DiffFit (Xie et al., 2023) also fine-tunes diffusion models with very few parameters, but trains the bias terms and newly added scale factors of selected layers, not attention logits.

E Evaluation protocol and additional results

Benchmarks. RealSR (Cai et al., 2019) and DRealSR (Wei et al., 2020) are evaluated on the paired $4\times$ center crops of the StableSR protocol (Wang et al., 2024a) (100 and 93 images); the low-resolution input is upsampled to the output size before it is given to the editor. ScreenSR (Wu et al., 2026) (130 images) is evaluated at its input size, without further enlargement, so it tests restoration at scale one after training on $4\times$ triples. RealPhoto60, the 60 real-world images collected by SUPIR (Yu et al., 2024), is restored at its input size; RealLQ250 (Ai et al., 2024) and RealLR200 (Wu et al., 2024c) are restored at $4\times$. Every method is scored on every image of every set it is reported on.

Scorer. PSNR and SSIM are computed on the BT.601 luminance channel without border cropping; LPIPS uses AlexNet features; MANIQA uses the PIPAL (Gu et al., 2020) weights and CLIP-IQA the RN50 backbone of CLIP (Radford et al., 2021). All metrics come from IQA-PyTorch (Chen & Mo, 2022). We checked the scorer against published numbers: on the RealSR outputs of PiSA-SR and OSEDiff in Table 1, produced with the authors' code, PSNR, SSIM and LPIPS agree with the published values to every digit, MUSIQ differs by 0.01, and MANIQA and CLIP-IQA differ by less than .001.

Baselines. Each baseline was run with its authors' released code, weights and settings, on inputs prepared so that its output has our output size; no output was resized. On ScreenSR and RealPhoto60, which are restored at input size, SinSR and InvSR, which support only $4\times$, receive the input downsampled by four with bicubic interpolation. CoTIR-9B was run with its released configuration (five steps, no CFG) and a prompt that names the task. On the $4\times$ sets, its outputs have far lower no-reference scores than those of every other method (MUSIQ 40.38 on RealSR) and one of the highest SSIMs. LDG-LoRA at $g = 1$ still has a higher PSNR and a lower LPIPS than CoTIR-9B on all three sets with ground truth. Every comparison in Section 4.2 holds with or without it. The controls that some baselines expose (the LoRA scales of PiSA-SR, the starting step of InvSR, the noise level of HYPIR, the guidance scale of VOSR) were left at their released values.

Table 4 gives the three sets without ground truth. At its quality point, LDG-LoRA has a higher MANIQA and LIQE than every baseline on all six sets. Of the baselines, only TSD-SR beats it in any no-reference metric, in three of the 24 cells: in MUSIQ on ScreenSR and RealPhoto60 (by 0.08 and 0.18) and in CLIP-IQA on DRealSR. On RealLQ250 and RealLR200, LDG-LoRA leads every baseline in all four metrics. On every one of the six sets, the best MUSIQ, MANIQA and LIQE of all methods belong to one of our four pairs, and so does the best CLIP-IQA on five.

Matched on other quality metrics. Section 4.2 compares LDG-LoRA with each baseline at the baseline's MUSIQ, along the measured scale ($g = 1$ to 3, interpolated linearly), or at $g = 1$ where the baseline's MUSIQ is lower. With MANIQA, CLIP-IQA or LIQE as the quality axis instead, LDG-LoRA has the higher PSNR and the lower LPIPS in 60, 53 and 54 of the 60 comparisons. On RealSR and ScreenSR it wins 119 of these 120 comparisons. The one loss is TSD-SR's LPIPS on RealSR on the CLIP-IQA axis. The other twelve losses are all on DRealSR, whose references score below our output at $g = 1$ on all four no-reference metrics (MUSIQ 50.6 against 62.5; Appendix E.2). Matching a baseline's CLIP-IQA or LIQE there takes a scale above one, and the fine detail this adds is absent from these references (Table 6). On CLIP-IQA, these losses are both metrics against PiSA-SR, both against TSD-SR, whose CLIP-IQA the scale does not reach by $g = 3$, LPIPS against OSEDiff and PSNR against SeeSR. On LIQE, they are both metrics against PiSA-SR and SeeSR, and LPIPS against TSD-SR and OSEDiff.

Color correction. StableSR (Wang et al., 2024a) introduced color correction for diffusion-based super-resolution: it aligns the mean and standard deviation of each color channel of the output with those of the input, which is AdaIN (Huang & Belongie, 2017) applied to the color channels. AdaIN is part of the released code or settings of PiSA-SR, TSD-SR, OSEDiff, SeeSR, VOSR and DiT4SR; HYPIR and InvSR apply the wavelet correction of StableSR (Wang et al., 2024a) (InvSR with its authors' setting for real-world images), and SinSR and CoTIR-9B apply none. Every baseline keeps the post-processing of its authors: we add none where they use none, just as we state our own. The frozen editor shifts colors strongly: AdaIN raises its PSNR at $g = 1$ from 17.50, 17.34 and 18.30 dB to 23.34, 25.54 and 21.34 dB on RealSR, DRealSR and ScreenSR. The restoration adapters remove most of this shift. Without AdaIN, LDG-LoRA at $g = 1$ scores 25.34, 26.95 and 24.37 dB PSNR and .2404, .2758 and .1771 LPIPS on RealSR, DRealSR and ScreenSR (with it: 27.12, 29.75 and 24.96 dB; .2374, .2638 and .1768). LDG-SAC at $g = 1$ without the correction scores 25.50, 27.40 and 23.52 dB and .2380, .2860 and .1675 LPIPS. With or without the correction, both pairs have a lower LPIPS than every baseline on all three sets. The correction aligns the mean and spread of each color channel with the input. PSNR charges such a global color shift heavily, and LPIPS barely registers it: the correction changes the LPIPS of LDG-LoRA by at most .012 and its PSNR by up to 2.8 dB. Without it, LDG-LoRA still has a higher PSNR than six of the ten baselines on RealSR, three on DRealSR and nine on ScreenSR.

E.1 Random examples and failures

Figure 10 shows images drawn at random, without selection, and Figure 11 more RealIR-Bench examples. The random examples also show where LDG-LoRA at its quality point fails. Where small text is blurred in the input, every method writes letters of its own: LDG-LoRA writes the label MARTINI as MARTIAI, and each of the four baselines shown also distorts some of its letters. On the shirt in the ScreenSR photograph, LDG-LoRA is the only method shown that restores the word HEALTH, and every method misspells FIRST. Every method invents knit detail for the sweater, whose ground truth is out of focus, and LDG-LoRA also changes the pattern of its stripes. On the flat gray wall from DRealSR, LDG-LoRA alone adds speckles and a jagged edge. Such texture on flat regions is consistent with the error amplification of Appendix C.1.

Figure 12 shows two RealSR crops. The large sign is restored legibly. The small labels are barely legible in the input, and every method shown, PiSA-SR and TSD-SR included, writes sharp but wrong letters there, a failure that no-reference metrics do not detect. Figure 13 shows a failure on RealIR-Bench: on a dark close-up, the low-light LoRA pair turns the black background into a gray field with blocky artifacts, while RealRestorer brightens it and the SAC pair leaves it dark, both without artifacts.

E.2 Ablations and the second prediction

Table 5 gives the ablations on RealSR; Figure 6 compares the second predictions of its first two rows on RealIR-Bench images, at the same scale. Guiding the last step lowers fidelity at equal quality (24.67 against 25.82 dB at MUSIQ 71). Shuffling the SAC biases, or replacing them by their mean over layers and heads, without retraining, lowers the PSNR at $g = 1$ from 26.95 to 22.07 and 23.87 dB. The mean control is sharper there (MUSIQ 71.13), but the SAC pair at MUSIQ 71 is still far closer to the ground truth (25.52 dB and LPIPS .2989, against 23.87 dB and .3642): the

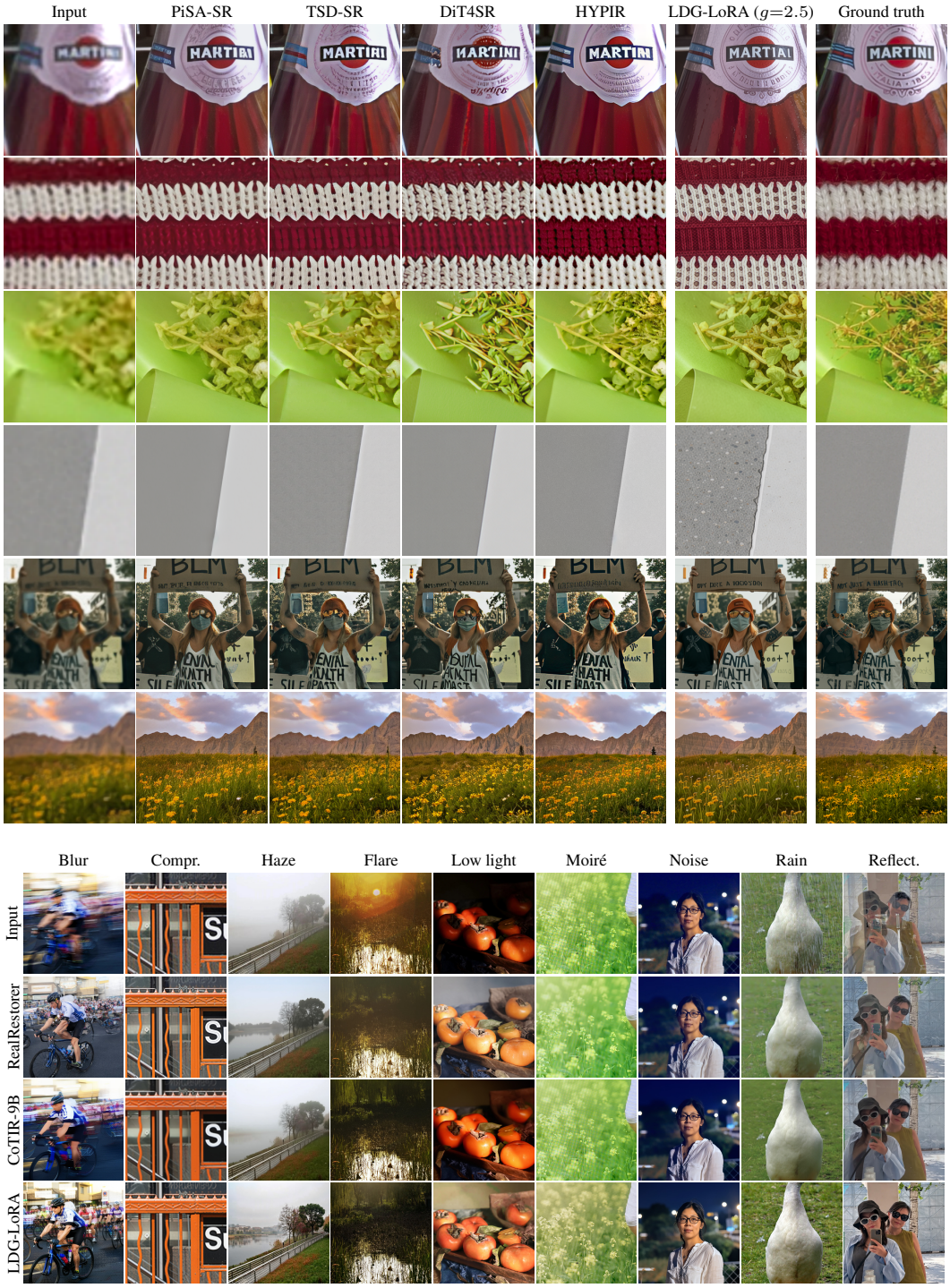


Figure 10: **Random examples.** Top: two images drawn at random from each blind super-resolution set with ground truth (RealSR, DRealSR, ScreenSR), center crops. Bottom: one image drawn at random per RealIR-Bench degradation, center square crops. No image was selected or replaced.

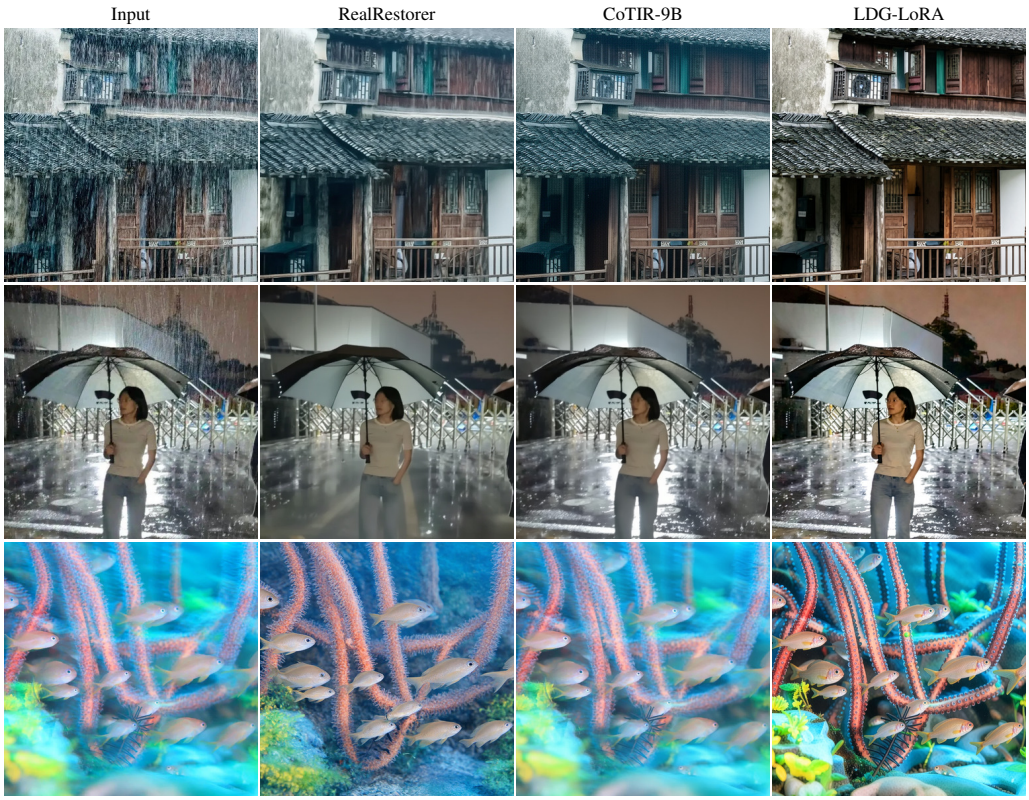


Figure 11: **More RealIR-Bench examples** (rain, rain, motion blur). LDG-LoRA removes the rain streaks and keeps the air-conditioner cage, the shutters and the carved doors, and the wet ground under the umbrella, which RealRestorer smears.



Figure 12: **Recovered structure and invented detail.** Crops of two RealSR images (Canon_004, top; Canon_002, bottom), restored by LDG-LoRA on FLUX.2 [klein] 9B at $g = 1$ and at $g = 1.75$, a scale between its fidelity and quality points, and by two baselines. The large sign is restored legibly. The small labels are barely legible in the input, and every method shown writes sharp but wrong letters there, PiSA-SR and TSD-SR included. Inputs are enlarged for display.

pattern across heads matters, not a global reweighting. Rendered alone with the method's sampler on the 50 held-out triples, the degradation adapters move toward their targets (LPIPS to y^- .115 for LoRA and .119 for SAC, against .176 for the input y), while the frozen editor under the same degradation prompt moves away (.326). Table 6 repeats the comparison of the first two rows of Table 5 on DRealSR and ScreenSR, with paired-bootstrap intervals on all three sets. LDG-LoRA reaches a higher MUSIQ than single-adapter CFG on every set, at $g = 3$ against $g = 6$ or 8. At matched MUSIQ, its LPIPS is lower at six of the seven matched points and never significantly higher. It is significantly closer to the ground truth on ScreenSR, by .0119 LPIPS at MUSIQ 72.5 and .0176 at MUSIQ 73, where its LPIPS is lower on 115 of the 130 images, and its PSNR there is not lower (+.03 and +.04 dB). It is also significantly closer on RealSR at MUSIQ 68. ScreenSR is

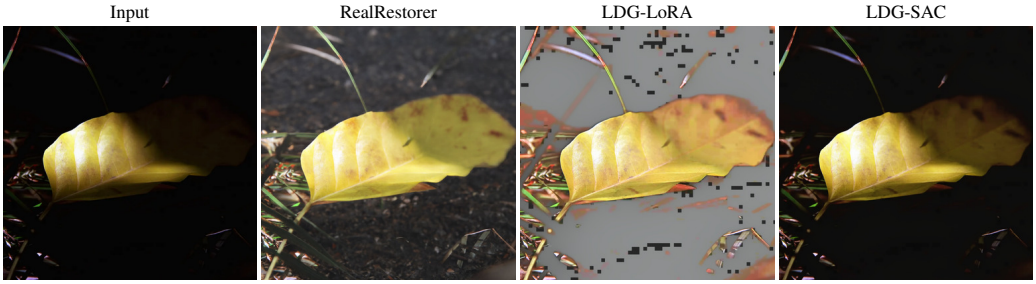


Figure 13: **A failure on RealIR-Bench** (low light, lowlight_0022, square crop). The low-light LoRA pair brightens the leaf with its veins and its shadowed half, but it turns the black background, which is almost black in the input, into a gray field with blocky artifacts. LDG-SAC leaves the background dark. Over the 46 low-light images, LDG-LoRA has the second-highest FS (.1565, after RealRestorer's .1611) and a higher removal score than single-adapter CFG (1.696 against 1.326; Tables 2 and 14).

Table 5: **Ablations on RealSR** (Klein, 100 images, color-corrected). The $g = 1$ columns are the restoration adapter alone. *Highest MUSIQ*: the highest MUSIQ over the scales measured for a row, which end at $g = 6$ for the two prompted second predictions with LoRA, $g = 2.5$ for the guided last step, $g = 1.3$ for the SAC controls and $g = 3$ otherwise. The MUSIQ of the LDG-LoRA pair still rises at $g = 3$. LDG reaches the highest MUSIQ of each block, with LoRA at half the scale of the prompted second predictions. Right: fidelity where the curve, followed in order of g and interpolated linearly, first reaches MUSIQ 68 and 71 (–: the curve does not pass it). Best of each block in bold; the two SAC controls are not marked.

		at $g = 1$				at MUSIQ 68		at MUSIQ 71	
Configuration	Trained parameters	PSNR ↑	LPIPS ↓	MUSIQ ↑	Highest MUSIQ	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
LDG-LoRA (restoration + degradation adapter)	55.6M	27.12	.2374	66.92	**72.13** (g=3)	26.88	.2429	25.82	**.2728**
second prediction: restoration adapter under p_- (single-adapter CFG)	27.8M	27.12	.2374	66.92	71.14 (g=6)	26.93	.2461	**26.07**	.2754
second prediction: frozen editor under p_-	27.8M	27.12	.2374	66.92	68.82 (g=6)	**26.98**	**.2413**	–	–
last step guided at g	55.6M	27.12	.2374	66.92	71.66 (g=1.75)	26.47	.2495	24.67	.2833
LDG-SAC (restoration + degradation adapter)	12,288	26.95	.2365	61.82	**73.63** (g=3)	26.29	.2621	25.52	.2989
second prediction: restoration adapter under p_- (single-adapter CFG)	6,144	26.95	.2365	61.82	73.24 (g=3)	26.33	.2630	**25.63**	**.2972**
second prediction: frozen editor under p_-	6,144	26.95	.2365	61.82	69.46 (g=3)	**26.50**	**.2572**	–	–
SAC biases shuffled across layers, heads and sources (not retrained)	12,288	22.07	.3709	41.98	52.96 (g=1.3)	–	–	–	–
SAC biases replaced by their mean over layers and heads (not retrained)	12,288	23.87	.3642	71.13	72.57 (g=1.3)	–	–	–	–

restored at its input size, while the adapters are trained for $4\times$ super-resolution, so this lead holds at a magnification the adapters were not trained for. MUSIQ 73 lies at the top of single-adapter CFG's range on ScreenSR (73.04 at $g = 8$), so the interval at 72.5 is the firmer of the two. On RealSR and DRealSR, single-adapter CFG keeps up to 0.25 dB more PSNR. The reason lies in the references of these two sets, which carry less fine detail than our output (below).

In Table 5, the untrained second prediction, the frozen editor under p_-, moves the output little. Its MUSIQ stops at 68.82 at $g = 6$ with LoRA and at 69.46 at $g = 3$ with SAC, the largest scales measured, below the MUSIQ 71 that both trained second predictions pass. Near the top of that range, at MUSIQ 68, it is slightly closer to the ground truth than LDG (by 0.10 dB and .0016 LPIPS with LoRA, 0.21 dB and .0049 with SAC). On RealIR-Bench it has the lowest FS, RS, CLIP-IQA+ and LIQE of the three guided variants (.1582, 1.206, .5744 and 3.038, against .1677, 1.489, .6195 and 3.507 for LDG-LoRA; Appendix F.1). With SAC, the two trained second predictions are level at equal MUSIQ on RealSR, within 0.11 dB and .0017 LPIPS. The SAC degradation adapter reaches the highest MUSIQ of any row of Table 5 (73.63), and on RealIR-Bench a higher RS than single-adapter CFG (1.364 against 1.238).

Of the three sets with ground truth, only ScreenSR has references that score above the output of our restoration adapter, and there the learned second prediction is closer to them by the largest margins. The no-reference metrics show this directly. On RealSR and DRealSR, the references score below the restoration adapter alone on all four (mean MUSIQ 57.5 and 50.6, against 66.9 and 62.5). On ScreenSR, they score above it on all four (MUSIQ 72.2 against 71.5). The high-frequency energy E of an output (above $1/8$ cycles per pixel), relative to its ground truth, gives the same picture. The restoration adapter alone has $E = 0.82$ on RealSR and 1.06 on DRealSR: it already carries about as much fine detail as these references. Every gain in MUSIQ then adds detail that the reference lacks,

Table 6: **The two second predictions at matched MUSIQ on the three sets with ground truth** (Klein, LoRA, color-corrected; the restoration adapter is the same). Each dial is followed in order of g and interpolated linearly; differences are LDG-LoRA minus single-adapter CFG, with 95% intervals from a paired bootstrap over images (2,000 resamples; resamples in which a dial does not reach the level are left out). A negative LPIPS difference means LDG-LoRA is closer to the ground truth. Bold: the interval excludes zero. E: high-frequency energy of the output relative to the ground truth ($E = 1$: as much fine detail as the reference), at $g = 1$ and at the matched MUSIQ. Highest MUSIQ: the highest MUSIQ of each dial and its g.

			E			LPIPS ↓		PSNR ↑	
Set	Highest MUSIQ (LDG-LoRA / single)	MUSIQ	$g = 1$	LDG-LoRA	single	LDG-LoRA / single	difference	LDG-LoRA / single	difference
RealSR	72.13 (g=3) / 71.14 (g=6)	68	.82	.95	.89	.2429 / .2461	**−.0032 [−.0070, −.0005]**	26.88 / 26.93	−.05 [−.10, +.00]
		71		1.58	1.19	.2728 / .2754	−.0027 [−.0096, +.0039]	25.82 / 26.07	**−.25 [−.45, −.01]**
DRealSR	69.32 (g=3) / 69.07 (g=8)	65	1.06	1.40	1.30	.2787 / .2831	−.0044 [−.0073, +.0005]	29.22 / 29.30	**−.09 [−.22, −.02]**
		67		1.85	1.59	.2998 / .3035	−.0037 [−.0092, +.0149]	28.58 / 28.71	−.13 [−.45, +.01]
		68		2.29	1.84	.3201 / .3177	+.0024 [−.0087, +.0267]	28.05 / 28.29	−.24 [−.68, +.01]
ScreenSR	74.27 (g=3) / 73.04 (g=8)	72.5	.68	.77	.73	.1684 / .1803	**−.0119 [−.0189, −.0020]**	24.60 / 24.57	+.03 [−.05, +.10]
		73		.84	.76	.1660 / .1835	**−.0176 [−.0199, −.0055]**	24.32 / 24.28	+.04 [−.05, +.14]

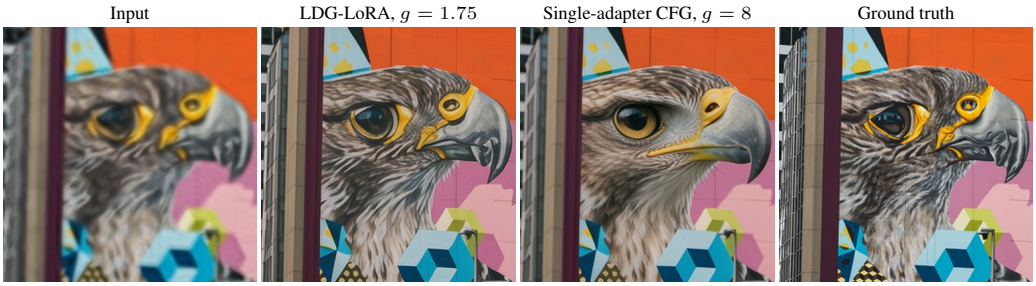


Figure 14: **The two second predictions at matched MUSIQ on ScreenSR** (image 064, color-corrected). At these scales, the set's mean MUSIQ is 73.2 and 73.0. Single-adapter CFG needs $g = 8$ to reach this quality, and at that scale it repaints the painted hawk as a photograph, changing the eye and the beak. The degradation adapter reaches it at $g = 1.75$ and keeps the painting.

and full-reference metrics charge it. Single-adapter CFG adds less of this detail at the same MUSIQ (at MUSIQ 71 on RealSR, $E = 1.19$ against 1.58), and full-reference metrics favor it for this. On ScreenSR, the restoration adapter alone has $E = 0.68$, so the references carry more fine detail than the output. There the detail that the degradation adapter adds moves the output toward the reference: its LPIPS falls from .1768 at $g = 1$ to .1654 at $g = 1.75$ while MUSIQ rises, whereas the LPIPS of single-adapter CFG only rises, to .1838 at $g = 8$. At that scale, single-adapter CFG repaints content (Figure 14).

Table 7 measures the second prediction on the held-out synthetic triples (50 for super-resolution, the first 25 of each RealIR degradation), for which x, y and y^- are known. From the same noise, the restoration adapter's own sampler runs to a state z_k; there the clean estimates of the restoration adapter ($\hat{a}_+$, under p_+) and of each second prediction ($\hat{a}_-$, under p_-) are decoded, and the difference of the two images is compared with $x - y$. The learned direction is the best aligned with $x - y$ on all ten rows. Its cosine is .387 in super-resolution, against .116 for single-adapter CFG and .144 for the frozen editor, and .731 on average over the nine RealIR degradations, against .274 and .409. The degradation adapter's estimate is also close to its target y^- (LPIPS .118 in super-resolution, .095 on average on RealIR), while the two prompted second predictions are far from it (.521 and .594 for single-adapter CFG, .419 and .528 for the frozen editor). In full samples on the super-resolution triples, LDG-LoRA is both closer to the clean image and higher in MUSIQ: going from $g = 1$ to 1.5 lowers its LPIPS to the clean image from .163 to .151 and raises MUSIQ from 71.2 to 72.9, while single-adapter CFG reaches .162 and 71.4. On the RealIR triples both end close to the clean image at $g = 1.5$ (LPIPS .03–.19; means .0917 for LDG-LoRA and .0903 for single-adapter CFG), and LDG-LoRA reaches the higher MUSIQ on eight of the nine degradations.

Table 7: **The second prediction on held-out synthetic triples** (Klein, LoRA; the clean image x, input y and stronger degradation y^- are known). Left: after the first of the four sampling steps, the cosine between the guidance direction $D(\hat{a}_+)-D(\hat{a}_-)$ and the true direction $x-y$, its length relative to $\|x-y\|$, and the LPIPS (AlexNet) of the second prediction's estimate to y^-, for the degradation adapter (LDG-LoRA), the restoration adapter under the degradation prompt (single-adapter CFG) and the frozen editor. Right: full samples at $g = 1.5$, LPIPS to x and MUSIQ. The state after the second step gives the same picture. The RealIR triples are among those on which the RealIR LoRA checkpoints were chosen (Appendix F). Best of each group in bold.

		$\cos(D(\hat{a}_+)-D(\hat{a}_-),\, x-y)\uparrow$			$\|D(\hat{a}_+)-D(\hat{a}_-)\|/\|x-y\|$		LPIPS$(D(\hat{a}_-),\, y^-)\downarrow$			LPIPS to x at $g=1.5\downarrow$		MUSIQ at $g=1.5\uparrow$	
Held-out set	n	LDG-LoRA	single	frozen	LDG-LoRA	single	LDG-LoRA	single	frozen	LDG-LoRA	single	LDG-LoRA	single
Super-resolution	50	**.387**	.116	.144	.951	.360	**.118**	.521	.419	**.1514**	.1617	**72.93**	71.37
Blur	25	**.428**	.024	.228	1.063	.259	**.090**	.451	.423	**.1750**	.1860	**72.71**	71.53
Compr.	25	**.182**	.066	.147	1.166	.461	**.292**	.503	.502	**.1616**	.1621	**73.05**	72.28
Haze	25	**.848**	.517	.580	1.277	.245	**.048**	.731	.623	.1439	**.1421**	**73.34**	73.15
Flare	25	**.969**	.408	.741	1.999	.111	**.025**	.342	.363	.0298	**.0287**	69.19	**69.26**
Low light	25	**.993**	.920	.973	1.160	.179	**.028**	.611	.357	.0457	**.0404**	**73.58**	73.53
Moiré	25	**.881**	.069	.191	2.062	.265	**.037**	.594	.583	.0404	**.0389**	**73.23**	73.05
Noise	25	**.514**	.014	.080	1.519	.238	**.160**	1.073	.915	.0953	**.0856**	**73.59**	73.17
Rain	25	**.882**	.221	.441	1.613	.133	**.099**	.672	.582	.0925	**.0895**	**73.49**	73.38
Reflect.	25	**.883**	.231	.300	1.737	.167	**.080**	.369	.404	.0412	**.0394**	**73.09**	72.92
RealIR, mean of nine		**.731**	.274	.409	1.511	.229	**.095**	.594	.528	.0917	**.0903**	**72.81**	72.47

Table 8: **Guidance with and without adapters** on the three sets with ground truth, color-corrected, along the whole scale of each system. The frozen editor uses the same sampler, prompt pair and scales as LDG, without adapters; it is shown for reference and not ranked. *Cost*: LPIPS added per MUSIQ point gained relative to the row's own $g = 1$, averaged over the three sets. The operating points of Table 1 are $g = 1$ and 2.5 for LDG-LoRA, $g = 1$ and 1.5 for LDG-SAC.

		RealSR			DRealSR			ScreenSR			
Model	g	PSNR ↑	LPIPS ↓	MUSIQ ↑	PSNR ↑	LPIPS ↓	MUSIQ ↑	PSNR ↑	LPIPS ↓	MUSIQ ↑	Cost ↓
Frozen Klein 9B, no adapter	1	23.34	.3762	71.36	25.54	.4321	68.04	21.34	.2157	73.37	–
	1.75	21.68	.4446	73.54	23.87	.5017	70.14	20.02	.2522	74.15	.0369
	2.5	20.83	.4764	74.23	22.96	.5319	71.02	19.15	.2828	74.48	.0429
LDG-LoRA (ours)	1	27.12	.2374	66.92	29.75	.2638	62.47	24.96	.1768	71.52	–
	1.25	26.69	.2473	68.86	29.13	.2811	65.41	24.73	.1703	72.23	.0006
	1.5	26.27	.2587	70.06	28.58	.2999	67.00	24.45	.1665	72.79	.0022
	1.75	25.91	.2698	70.87	28.12	.3174	67.90	24.16	.1654	73.24	.0038
	2	25.60	.2800	71.33	27.73	.3335	68.48	23.87	.1666	73.61	.0055
	2.5	25.13	.2963	71.86	27.15	.3592	69.09	23.35	.1734	74.05	.0083
	3	24.82	.3067	72.13	26.78	.3755	69.32	22.97	.1822	74.27	.0105
LDG-SAC (ours)	1	26.95	.2365	61.82	29.52	.2730	56.86	24.08	.1670	71.69	–
	1.1	26.55	.2508	66.29	28.99	.2958	62.32	23.84	.1629	72.35	.0004
	1.2	26.13	.2694	69.11	28.44	.3248	65.50	23.61	.1620	72.81	.0020
	1.3	25.70	.2901	70.62	27.91	.3533	67.25	23.37	.1634	73.16	.0038
	1.4	25.28	.3104	71.50	27.38	.3799	68.27	23.13	.1665	73.44	.0056
	1.5	24.86	.3300	72.05	26.88	.4052	68.89	22.90	.1713	73.67	.0074
	2	23.02	.4215	73.21	24.76	.5106	70.24	21.71	.2104	74.35	.0168

E.3 Guidance without adapters

Tables 8 and 9 run the frozen editor with the same sampler, prompt pair and scales as LDG, without any adapter. At $g = 1$ the frozen editor is 3.6–3.8 dB further from the ground truth on RealSR than the two adapter pairs. It gains no-reference quality by repainting the image. Figure 15 shows three examples: a flat printed illustration acquires fur and an eye, a pale leaf grows veins it does not have, and a hazy afternoon becomes a clear summer day. Each of these renders scores higher than ours on the metric of its table, and each is a different photograph. For this reason the frozen editor is not ranked with the restoration methods in Tables 1 and 2. On RealIR-Bench, its mean FS is below that of both pairs at every scale (.1336, .1229 and .1104 at $g = 1$, 3 and 5, against .1677 for LoRA and .1541 for SAC), and it changes the input more for each unit of removal (LPS/RS .349 to .363, against .292 and .299). Its FS is higher than that of both pairs on two degradations. On haze (.1166 at $g = 5$, against .1001 for LoRA and .0899 for SAC), it can repaint the scene: in Figure 15, a hazy afternoon becomes a clear summer day, and the benchmark's rater is asked only whether the haze is gone. It is also higher on noise (.0904 at $g = 1$, against .0805 and .0693) and, against the SAC pair only, on low light (.1501 at $g = 1$, against .1368, while the LoRA pair scores .1565). Figure 16 repeats the comparison on both editors, with both adapter types.

E.4 The second editor

The Kontext adapters use Klein's training recipe: the same learning rates, numbers of updates, loss weight and parameter averaging. What differs follows from the editor: the checkpoint, its 28-step schedule and the training data, which each editor renders itself (98 training triples on Kontext).

Table 9: **Guidance with and without adapters on RealIR-Bench**, all 456 images. LPS: LPIPS to the degraded input; RS: the evaluator's removal score; FS $= 0.2\,\mathrm{RS}\,(1-\mathrm{LPS})$. *Change per unit removed*: LPS/RS. For single-adapter CFG with the LoRA adapters (Table 13), this ratio is .264, but its RS is 1.364, below that of LDG-LoRA: it removes less. The frozen editor is shown for reference and not ranked.

Model	g	FS ↑	LPS ↓	RS ↑	Change per unit removed ↓	CLIP-IQA+ ↑	Q-Align ↑	LIQE ↑
Frozen Klein 9B, no adapter	1	.1336	.5383	1.484	.363	.6995	4.316	4.434
	3	.1229	.6791	1.945	.349	.7574	4.403	4.665
	5	.1104	.7238	2.028	.357	.7558	4.294	4.583
LDG-LoRA (ours)	per task	.1677	.4353	1.489	.292	.6195	3.958	3.507
LDG-SAC (ours)	per task	.1541	.4076	1.364	.299	.5924	3.911	3.275

The LoRA target modules follow each editor's diffusers training script, and Kontext's embedded guidance input is held at 2.5 (Appendix A). Our Kontext sampler reproduces the reference diffusers pipeline: its renders match at 55.0 and 53.8 dB PSNR on two held-out inputs. Table 11 lists what differs between the two editors.

The Kontext LoRA pair renders its 50 held-out triples at $g \in \{1, 1.5, 2, 2.5, 3\}$ with LPIPS to the clean images of .1567, .1528, .1699, .1839 and .1969, so $g = 1.5$ reconstructs them best. For the Kontext SAC pair we use $g = 1.5$, its quality point on Klein; its MUSIQ on the same triples rises steeply up to $g = 1.5$ and then levels off (67.34, 73.81, 74.81, 74.75 and 74.11).

In Table 1, the same recipe gives the highest PSNR of all methods at $g = 1$: the SAC pair on RealSR and DRealSR (27.50 and 30.32 dB) and the LoRA pair on ScreenSR (25.43 dB). At $g = 1$, the Kontext LoRA pair is closer to the ground truth than all ten baselines on the three sets, in PSNR (27.48, 30.10 and 25.43 dB, against at best 25.97, 28.45 and 24.71 dB) and in LPIPS (.2582, .2821 and .1846, against at best .2672, .2960 and .1951). At its quality point, the Kontext LoRA pair is closer to the ground truth than the Klein pair at the same MUSIQ on RealSR (25.66 against 25.37 dB, LPIPS .2748 against .2880) and DRealSR (27.95 against 27.77 dB, .3131 against .3316), and on ScreenSR it has the lower LPIPS (.1616 against .1660) at 0.2 dB less PSNR.

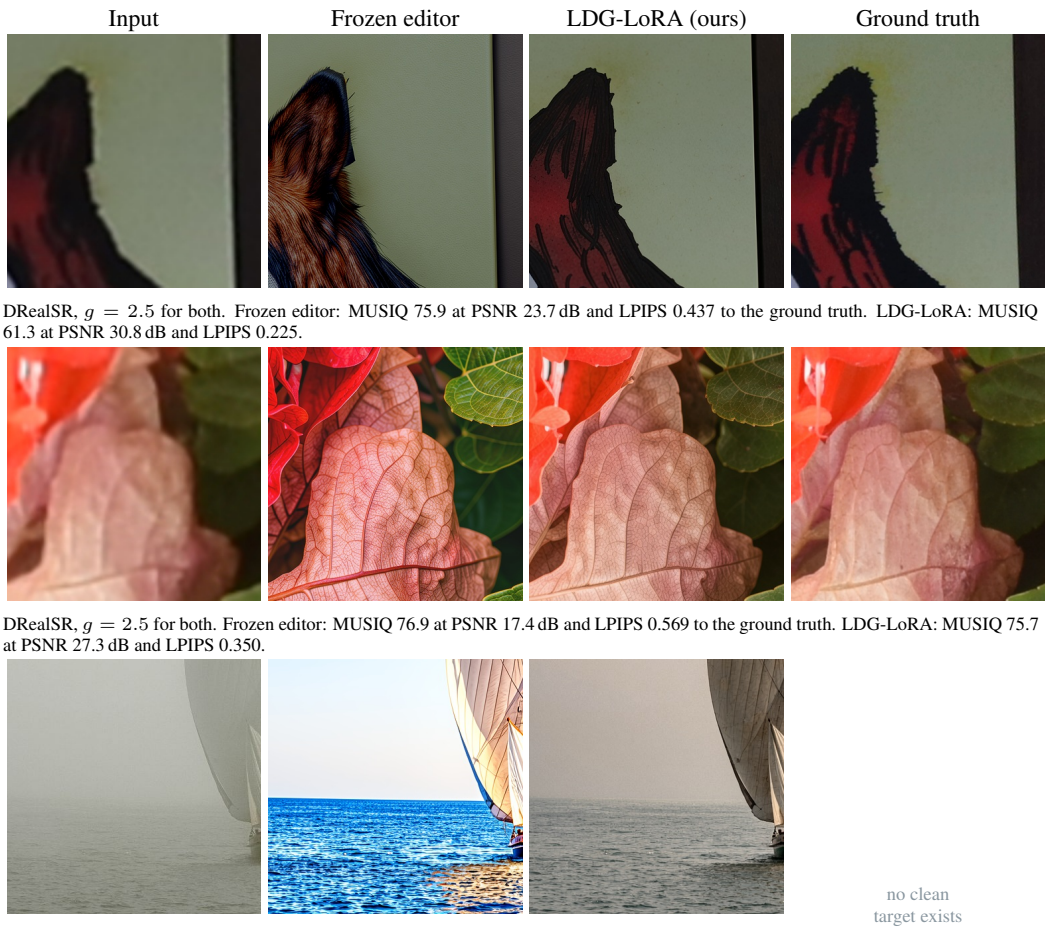
Table 10 follows both adapter types and the frozen editor on both editors along the scale.

F REALIR-BENCH: TRAINING AND EVALUATION

Scope. The comparison covers RealRestorer (Yang et al., 2026), CoTIR-9B (Guo et al., 2026) with a prompt that names the task and with a generic prompt, Edit2Restore (Yılmaz et al., 2026) on the five tasks its released adapter covers, and our LoRA and SAC pairs on FLUX.2 [klein] 9B. Our copy contains all 456 images of the public release; the RealRestorer paper states 464. The tasks have 50 (blur), 60 (compression), 45 (haze), 53 (lens flare), 46 (low light), 50 (moiré), 49 (noise), 53 (rain) and 50 (reflection) images. Every method except CoTIR with the generic prompt receives the degradation type through its own interface, so the benchmark measures restoration given the degradation type, not the identification of an unknown degradation.

Training data. Each degradation has its own 100 training and 50 held-out triples, built on the clean images of Klein's super-resolution set, with no real image and no external model. The degradation is procedural and, where the real degradation is regional, it is placed in a region: for haze, a fog layer that follows a depth ramp and cloud noise; for lens flare, a render dimmed to a night scene with a compact light core on its brightest spot. The degradation adapter's target is the same draw made stronger in the same region (for example, the same motion path or the same noise field scaled by 2.5), and outside that region the input and the target are identical. A draw is accepted when it degrades at least 8% of the frame and the degraded region falls below a task-specific PSNR ceiling, with the target further below; for renders that cannot be degraded that far, the ceiling is relaxed by 3 dB after 13 and again after 26 of at most 40 attempts (81 of the 1,350 accepted draws needed this).

Instructions and adapters. For degradation X, the restoration instruction is *"Restore this image into an ultra high quality, 8K resolution image with sharp focus, crystal clear details and clean fine textures. Remove X."* and the degradation instruction is *"Make this image lower quality: add more X, making what it already has stronger, in exactly the same places, so that it looks even*



DRealSR, $g = 2.5$ for both. Frozen editor: MUSIQ 75.9 at PSNR 23.7 dB and LPIPS 0.437 to the ground truth. LDG-LoRA: MUSIQ 61.3 at PSNR 30.8 dB and LPIPS 0.225.

DRealSR, $g = 2.5$ for both. Frozen editor: MUSIQ 76.9 at PSNR 17.4 dB and LPIPS 0.569 to the ground truth. LDG-LoRA: MUSIQ 75.7 at PSNR 27.3 dB and LPIPS 0.350.

RealIR, haze, the frozen editor at $g = 5$. Haze is one of the two degradations where its FS is higher than ours; the benchmark's rater is asked only whether the haze is gone. No clean target exists. Frozen editor: FS 0.338, distance to the input 0.436. LDG-LoRA: FS 0.147, distance 0.263.

Figure 15: **The frozen editor under guidance repaints the image.** The same window of the same image at 1:1 pixels. The frozen editor uses the same sampler and prompts as LDG, without adapters, at the scale given under each row. It scores higher on each row's metric (no-reference, or FS on RealIR-Bench) and, where a ground truth exists, it is 7 and 10 dB further from it: a printed illustration becomes an animal with fur and an eye, a pale leaf grows veins it does not have, and a hazy afternoon becomes a clear summer day. It is therefore reported apart from the restoration methods, in Tables 8 and 9.

worse.", both followed by preservation sentences. X names the degradation, for example "haze, fog, mist, low clouds and atmospheric scattering". The SAC adapters use the super-resolution recipe unchanged. Each LoRA adapter saves a snapshot every 50 updates and keeps the snapshot, or the mean of ten consecutive snapshots, that best reconstructs its own task's 50 held-out triples (LPIPS to its target); the two adapters of a pair are chosen independently, and the chosen budgets lie between 750 and 2,000 updates. Each degradation is rendered at one guidance scale $g > 1$ for all its images (Table 12); $g = 1$ is the restoration adapter alone, which Table 13 reports as an ablation.

Evaluator. For input y_i, restored image $\hat{x}_i$ and known degradation d_i, let $s_{d_i}(\cdot)$ be the degradation-specific rating of Qwen3-VL-8B-Instruct (Bai et al., 2025), an integer from 1 (extreme degradation)

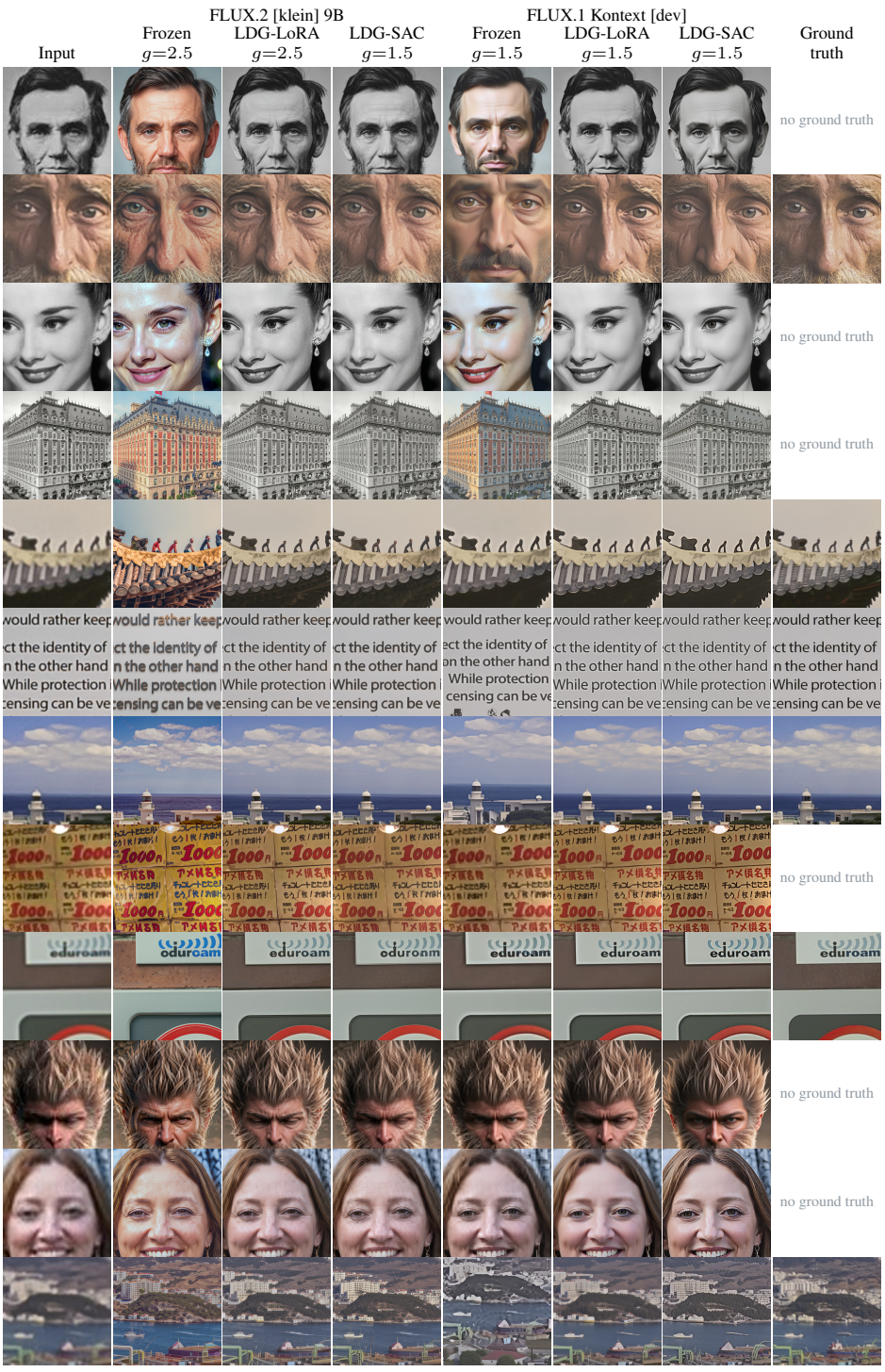


Figure 16: **The same guidance with and without adapters, on two editors.** Blind super-resolution (RealLR200, RealSR, RealLQ250, RealPhoto60), each pair at its quality point; the frozen editor uses the same sampler, prompts and scale. Without adapters, both editors repaint: they color grayscale photographs and replace the faces with other people, Klein turns the roof statuettes into people, embosses the printed text and rewrites "eduroam", and Kontext reframes the lighthouse and reflows the text. With the adapters, the same two editors restore the input.

Table 10: **The guidance scale from $g = 1$ to $g = 3$ on both editors**, RealSR (100 images), color-corrected: LDG-LoRA, LDG-SAC and the frozen editor under the same sampler and prompts. At $g = 1$ a pair runs its restoration adapter alone. †: the operating points of Table 1. For both LDG-LoRA pairs, all four no-reference metrics rise at every step of g: the scale is a quality dial. The frozen editors are shown for reference and not ranked.

		RealSR						
System	g	PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ ↑	MANIQA ↑	CLIP-IQA ↑	LIQE ↑
LDG-LoRA, Klein 9B	1†	27.12	.770	.237	66.92	.621	.585	3.69
	1.5	26.27	.742	.259	70.06	.667	.669	4.21
	2	25.60	.716	.280	71.33	.690	.709	4.44
	2.5†	25.13	.695	.296	71.86	.704	.732	4.55
	3	24.82	.682	.307	72.13	.711	.741	4.59
LDG-SAC, Klein 9B	1†	26.95	.760	.237	61.82	.599	.516	3.22
	1.5†	24.86	.666	.330	72.05	.713	.717	4.51
	2	23.02	.574	.422	73.21	.746	.752	4.71
	2.5	21.70	.508	.492	73.56	.760	.761	4.73
	3	20.56	.455	.549	73.63	.769	.768	4.72
Frozen Klein 9B	1	23.34	.626	.376	71.36	.674	.722	4.49
	1.5	22.09	.572	.429	73.15	.695	.749	4.68
	2	21.35	.542	.457	73.80	.704	.755	4.73
	2.5	20.83	.521	.476	74.23	.711	.762	4.75
	3	20.36	.502	.496	74.52	.715	.764	4.76
LDG-LoRA, Kontext	1†	27.48	.781	.258	66.50	.589	.551	3.52
	1.5†	25.66	.728	.275	71.59	.662	.701	4.38
	2	24.75	.696	.290	72.44	.674	.724	4.49
	2.5	24.24	.674	.301	72.80	.679	.733	4.51
	3	23.84	.655	.314	72.88	.682	.735	4.51
LDG-SAC, Kontext	1†	27.50	.782	.257	63.02	.552	.450	3.02
	1.5†	23.42	.656	.309	72.12	.659	.671	4.06
	2	20.32	.510	.389	72.16	.644	.690	3.79
	2.5	18.51	.402	.455	71.66	.612	.647	3.37
	3	17.22	.338	.503	69.43	.552	.537	2.55
Frozen Kontext	1	24.86	.701	.352	63.45	.520	.488	3.15
	1.5	24.39	.683	.366	65.56	.537	.536	3.41
	2	23.77	.662	.379	66.58	.547	.568	3.58
	2.5	23.36	.648	.388	67.40	.553	.592	3.76
	3	22.91	.637	.398	67.89	.558	.600	3.85

to 5 (no degradation) in the released evaluator. The per-image scores are

$$L_i = \mathrm{LPIPS}_{\mathrm{Alex}}(\hat{x}_i, y_i), \qquad R_i = \max\{0,\, s_{d_i}(\hat{x}_i) - s_{d_i}(y_i)\}, \qquad F_i = 0.2\,(1 - L_i)\,R_i, \quad (20)$$

called LPS, RS and FS. LPS measures change relative to the degraded input and is distinct from the clean-reference LPIPS of the super-resolution tables. An unchanged input has $L_i = 0$ but $R_i = F_i = 0$. We follow the released evaluator: the prediction is resized to the input size with Lanczos resampling, the VLM sees each image separately after an aspect-preserving resize to about 512^2 pixels, and greedy decoding returns one integer. Per-image scores are averaged within each task, and the reported numbers are equal-weight means over the nine tasks; the image-weighted means differ by at most .0004 in FS for our two pairs and RealRestorer.

No-reference metrics. All five no-reference metrics that CoTIR reports on its real-world images rank both of our pairs above RealRestorer and CoTIR-9B in the mean over the nine degradations: CLIP-IQA+, Q-Align, LIQE, MA-CLIP and CLIP-IQA. Table 2 reports the first three (Wang et al., 2023; Wu et al., 2024a; Zhang et al., 2023). We compute them with IQA-PyTorch and its default

Table 11: The two editors. The adapter rows count trained parameters per adapter.

	FLUX.2 [klein] 9B	FLUX.1 Kontext [dev]
Released size	9B	12B
Transformer blocks (double + single stream)	$8 + 24 = 32$	$19 + 38 = 57$
Attention heads per block	32	24
Sampling steps K	4	28
Transformer evaluations, $g = 1$ / $g > 1$	4 / 7	28 / 55
Embedded guidance input	none	2.5
Training / held-out triples	100 / 50	98 / 50
LoRA parameters	27,787,264	54,165,504
SAC parameters (effective)	6,144 (4,096)	8,208 (5,472)

Table 12: **The guidance scale of each degradation.** One value per degradation, used for all of its images. With one scale for all nine degradations ($g = 1.5$, 2 or 3), the LoRA pairs also lead RealRestorer significantly (Appendix F.1).

Variant	Blur	Compr.	Haze	Flare	Low light	Moiré	Noise	Rain	Reflect.
LDG-LoRA	5	1.5	3	2	2	2	1.5	2	1.5
LDG-SAC	2	1.5	1.5	2	1.5	1.5	1.5	2	1.5

weights, as CoTIR's released evaluator does, on RGB inputs at their native size. MA-CLIP (Liao et al., 2026) and CLIP-IQA also rank both of our pairs above RealRestorer and CoTIR (macro MA-CLIP: LoRA .5191, SAC .5229, RealRestorer .4877, CoTIR-9B .4715; CLIP-IQA: .5322, .5352, .4562, .4230).

Baselines. The RealRestorer row is a run of its released model and code (28 steps, guidance 3.0, seed 42, size setting 1024), scored by our port of its evaluator against the same inputs as every other row. On that run's own decoded copies of the inputs, the port reproduces the mean FS that the reference evaluator recorded for the run, .1461, as .1469, with LPIPS within .001 on all 456 images and an identical rating difference on 435 of them. CoTIR-9B was run with its released code and configuration (five steps, no CFG). Edit2Restore uses its released unified adapter (trained with 32 pairs per task). No RealIR row is color-corrected, because matching the input's color statistics would undo intended changes of brightness and color.

F.1 The second prediction and a shared scale on RealIR-Bench

The trained degradation adapter removes the most degradation of all second predictions (Table 13, all 456 images, each degradation at its own scale). Its RS is 1.489, against 1.364 for single-adapter CFG, 1.206 for the frozen editor under p_- and 1.036 for the restoration adapter alone. Its RS is higher than that of single-adapter CFG on eight of the nine degradations (Table 14). Per image, its RS is higher on 62 images and lower on 14. The benchmark's rater finds 196 of its outputs free of the degradation, against 175 for single-adapter CFG. LDG-LoRA also has the highest CLIP-IQA+ (.6195) and LIQE (3.507) of the four configurations, and a higher FS than single-adapter CFG on lens flare, low light and reflection. The SAC pairs give the same order: RS 1.364 for LDG-SAC, 1.238 for single-adapter CFG and .984 for the frozen editor. Figure 17 shows this removal at the same scale. Single-adapter CFG leaves a distorted wheel rim and a transparent ghost of a moving leg, and LDG-LoRA rebuilds the round wheel and much of the leg.

The second prediction needs a trained adapter. Without one, the frozen editor under p_- is the weakest guided variant on FS, RS, CLIP-IQA+ and LIQE, both with LoRA (FS .1582) and with SAC (FS .1354). Without any adapter at all, the frozen editor under guidance repaints images (Figure 15). A prompt alone does not produce the stronger degradation y^-. On the held-out triples, the frozen editor's first estimate lies at LPIPS .528 from y^- and that of single-adapter CFG at .594, while the degradation adapter's lies at .095 (means over the nine degradations, Table 7). The learned guidance direction is also far better aligned with the true direction $x - y$: its cosine is .731, against .274 for single-adapter CFG and .409 for the frozen editor.

Table 13: **The second prediction on RealIR-Bench** (Klein, all 456 images, each degradation at its own scale). At these scales, the trained degradation adapter gives the highest RS, CLIP-IQA+ and LIQE in both blocks, and the untrained frozen editor is the weakest guided variant on FS, RS, CLIP-IQA+ and LIQE. LPS: LPIPS to the degraded input. RS: the evaluator's removal score. FS $= 0.2\,\mathrm{RS}\,(1-\mathrm{LPS})$ charges every change to the input, including removal. The last column follows a row's measured scale ($g = 1, 1.5, 2, 3, 5$ for every degradation) to the point where LPS $= 0.40$ (equal change to the input) and gives its RS there, by linear interpolation. It exists for the two rows rendered at all these scales. LDG-LoRA reaches this change at a smaller scale, because its guidance direction is longer (Table 7). Best of each block in bold. LPS is not marked (an unchanged input has the lowest).

Configuration	FS ↑	LPS	RS ↑	CLIP-IQA+ ↑	LIQE ↑	RS at LPS $= 0.40$ (matched change) ↑
LDG-LoRA (restoration + degradation adapter)	.1677	.4353	**1.489**	**.6195**	**3.507**	1.361
second prediction: restoration adapter under p_- (single-adapter CFG)	**.1726**	.3609	1.364	.6044	3.377	**1.485**
second prediction: frozen editor under p_-	.1582	.3295	1.206	.5744	3.038	–
restoration adapter alone	.1436	.2888	1.036	.5590	2.860	–
LDG-SAC (restoration + degradation adapter)	.1541	.4076	**1.364**	**.5924**	**3.275**	–
second prediction: restoration adapter under p_- (single-adapter CFG)	**.1550**	.3441	1.238	.5833	3.217	–
second prediction: frozen editor under p_-	.1354	.2731	.984	.5403	2.756	–
restoration adapter alone	.1114	.2385	.782	.5223	2.575	–

Table 14: **LDG-LoRA against single-adapter CFG on RealIR-Bench, per degradation** (the same restoration adapters, each degradation at its own scale). LDG-LoRA removes more on eight of the nine degradations (higher RS). It also changes the input more on all nine (higher LPS). FS, which charges every change, prefers single-adapter CFG on six. RS: the evaluator's removal score. LPS: LPIPS to the degraded input. FS $= 0.2\,\mathrm{RS}\,(1-\mathrm{LPS})$. Better of the two in bold.

	Blur	Compr.	Haze	Flare	Low light	Moiré	Noise	Rain	Reflect.	Mean
FS ↑, LDG-LoRA	.2203	.0648	.1001	**.2788**	**.1565**	.1316	.0805	.1788	**.2979**	.1677
FS ↑, single-adapter CFG	**.2456**	**.0734**	**.1090**	.2724	.1449	**.1361**	**.0906**	**.1868**	.2943	**.1726**
RS ↑, LDG-LoRA	**2.360**	**.667**	**1.089**	**1.774**	**1.696**	**1.340**	.939	**1.377**	**2.160**	**1.489**
RS ↑, single-adapter CFG	2.220	.633	.978	1.604	1.326	1.220	**.980**	1.340	1.980	1.364
LPS ↓, LDG-LoRA	.523	.510	.517	.206	.540	.484	.505	.338	.295	.435
LPS ↓, single-adapter CFG	**.428**	**.426**	**.396**	**.144**	**.455**	**.409**	**.461**	**.287**	**.241**	**.361**

Single-adapter CFG has the higher mean FS (.1726 against .1677) because FS charges every change to the degraded input, and removing a degradation is itself such a change. FS multiplies RS by one minus LPS, and LDG-LoRA has both the higher RS and the higher LPS (.435 against .361). The FS gap therefore comes from the larger change, not from less removal. The rating is a single integer from 1 to 5, so FS credits removal only when it crosses a step. On the 311 images where both outputs receive the same positive rating, LDG-LoRA is further from the input on 309 and has the lower FS on all of them. Where it earns the higher rating, its FS is higher on 60 of 62 images. Row 1 of Figure 17 is such a tie. Both outputs are rated 4, and LDG-LoRA, which rebuilds the wheel, changes the input more (0.254 against 0.201) and has the lower FS (0.298 against 0.320).

A larger change to the input is not in itself an error. RealIR-Bench has no clean reference, and LPS is the distance to the degraded input. Where the clean image is known, on the held-out triples at $g = 1.5$, LDG-LoRA stays about as close to it as single-adapter CFG (LPIPS to x .0917 against .0903, mean of the nine degradations). It also has the higher MUSIQ on eight of the nine (72.81 against 72.47; Table 7).

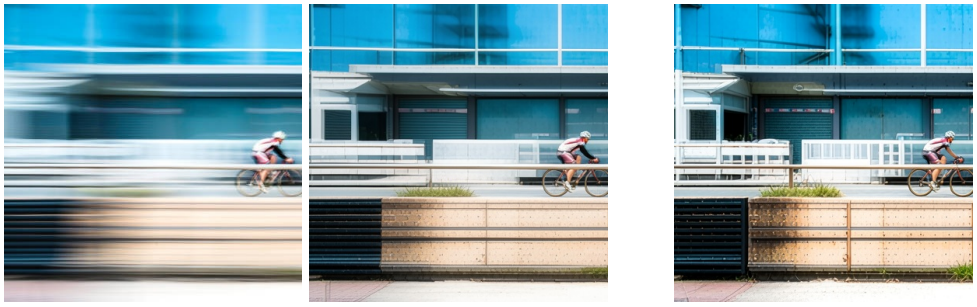
The scale sets how far the output moves. The learned direction is 6.6 times longer than that of single-adapter CFG on the RealIR triples (1.511 against .229 times $\|x - y\|$), so the same scale changes the output more. Blur and haze use the largest scales ($g = 5$ and 3), and there LDG-LoRA removes more than single-adapter CFG: RS 2.360 against 2.220 on blur, and 1.089 against .978 on haze. At these scales it rebuilds strongly smeared regions almost like inpainting. In some blur images this leaves streaks or speckles, and on haze it gives some skies an unnatural color or contrast. Single-adapter CFG changes these images less and makes fewer such errors, but it also removes less. Because the two directions differ in length, a comparison at equal change puts them at different scales, with

LDG-LoRA at the smaller one. Where each reaches LPS $= 0.40$, single-adapter CFG has the higher RS (1.485 against 1.361; last column of Table 13). At a smaller scale for all degradations, $g = 1.5$, 2 or 3, LDG-LoRA still leads RealRestorer significantly (next paragraph).

One scale for all degradations. The lead over RealRestorer does not need a scale per degradation. With one scale for all nine degradations, the LoRA pairs reach a mean FS of .1626, .1629 and .1591 at $g = 1.5$, 2 and 3. Each is significantly above RealRestorer (.1463): the lower ends of the 95% bootstrap intervals of the lead are $+.0080$, $+.0084$ and $+.0050$. The lead comes from guidance. The restoration adapter alone ($g = 1$, .1436) is not significantly different from RealRestorer (95% interval of the difference $[-.0115, +.0063]$). With $g = 5$ for all nine degradations, FS (.1473) is also not significantly different from RealRestorer (interval $[-.0072, +.0093]$). Only blur uses a scale this large (Table 12).



Blur, $g = 5$. Input rated 2. Single-adapter CFG: rated 4, distance to the input 0.201, FS 0.320. LDG-LoRA: rated 4, distance 0.254, FS 0.298.



Blur, $g = 5$. Input rated 2. Single-adapter CFG: rated 3, distance to the input 0.388, FS 0.122. LDG-LoRA: rated 5, distance 0.515, FS 0.291.



Blur, $g = 5$. Input rated 2. Single-adapter CFG: rated 2, distance to the input 0.226, FS 0.000. LDG-LoRA: rated 4, distance 0.292, FS 0.283.



Haze, $g = 3$, whole frame. Input rated 2. Single-adapter CFG: rated 2, distance to the input 0.304, FS 0.000. LDG-LoRA: rated 3, distance 0.512, FS 0.098.

Input Single-adapter CFG LDG-LoRA (ours)

Figure 17: **What the degradation adapter changes at the same scale.** RealIR-Bench, with the same restoration adapter and the same scale g in both columns. *Single-adapter CFG* uses the restoration adapter under the degradation prompt as the second prediction, and LDG-LoRA uses the trained degradation adapter. The blur rows are 1:1 crops, and the haze row is the whole frame. The benchmark's rater prefers LDG-LoRA by two grades in rows 2 and 3 and by one grade in the haze row. Single-adapter CFG leaves a distorted wheel rim in row 1 and a transparent ghost of the moving leg in row 3, and LDG-LoRA rebuilds the round wheel and much of the leg. In row 1 both outputs are rated 4, so the larger change of LDG-LoRA to the input lowers its FS: FS charges every change, including removal. Over the whole benchmark, LDG-LoRA has the higher removal score on eight of the nine degradations and 196 outputs rated free of the degradation, against 175 (Table 14, Appendix F.1).