REFERENCE REGULATION IN IN-CONTEXT IMAGE GENERATION FOR INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

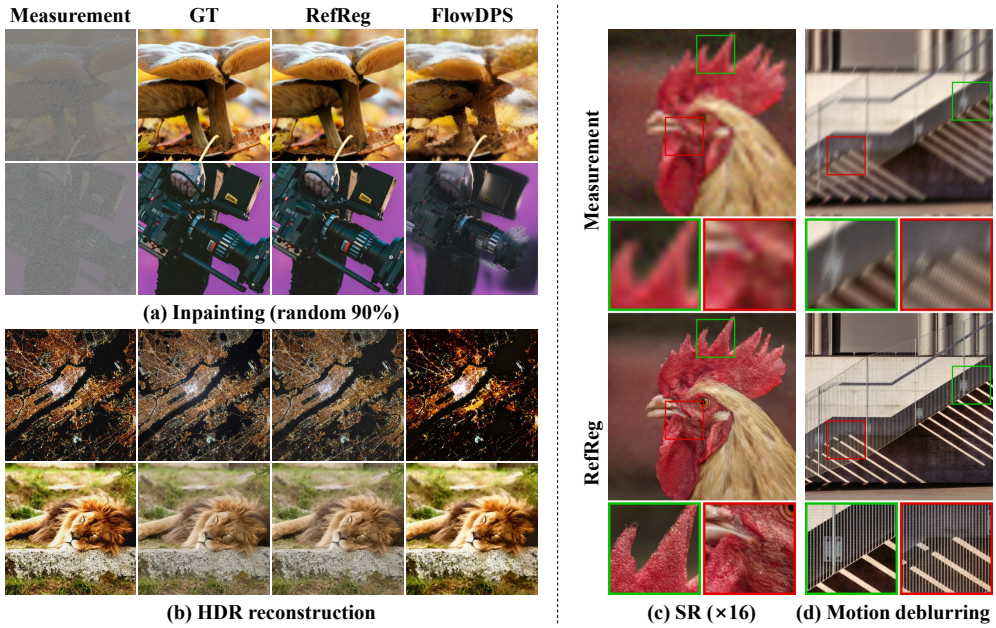

Figure 1: **Overview of RefReg**. Left: random inpainting (90% missing pixels) and high-dynamic-range (HDR) reconstruction. Right: super-resolution $\times 16$ and motion deblurring.

ABSTRACT

Pretrained in-context image generation models, which we refer to as image/text-to-image (IT2I) models, accept a reference image alongside a text prompt. Given a reference derived from a degraded measurement, such a model acts as an amortized, measurement-dependent prior that supplies the spatial structure a text prompt leaves unspecified, and data-consistency updates correct it with the exact likelihood. The measurement enters twice, and how much the prior is trusted becomes a design choice. In native IT2I generation from degraded references, we observe early structure preservation, where the observed layout appears at the first step and later steps refine details, and reduced reference influence, where generations that depart from the observed layout respond less to the reference from the start. A reference alone is thus not enough. We propose **RefReg**, a posterior sampler that regulates reference influence at three levels: constructing a measurement-derived reference, scheduling reference guidance to emphasize structure early and detail later, and suppressing generation-to-text attention during early structure formation. To compare priors on a common footing, we derive an upper bound on the average posterior approximation error of a conditional prior under data-consistency updates. By this criterion, image conditioning lowers the error bound throughout sampling in generations that preserve the observed layout on IT2I backbones. These results make a case for building inverse solvers on IT2I rather than text-only priors. Across linear, nonlinear, and partially blind inverse problems against flow-based solvers, RefReg improves perceptual quality with competitive fidelity and remains robust with as few as eight sampling steps.

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.

1 INTRODUCTION

Imaging inverse problems seek to reconstruct a clean image $x_0 \in \mathbb{R}^d$ from a measurement $y \in \mathbb{R}^m$, modeled as $y = \mathcal{A}(x_0) + n$, where $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$ is the forward operator and n is measurement noise. Since multiple images can explain a measurement, posterior sampling targets $p(x_0 \mid y) \propto p(y \mid x_0)\, p(x_0)$, where the likelihood encodes the observation model and the prior describes plausible clean images. Pretrained diffusion and flow models provide unconditional (Kawar et al., 2022; Wang et al., 2023) or text-conditioned priors (Chung et al., 2024b; Erbach et al., 2025), with measurements incorporated through likelihood guidance or data consistency updates (Chung et al., 2023; Song et al., 2024; Kim et al., 2025a). Even when derived from measurements (Chung et al., 2024b; Kim et al., 2025b), text prompts provide semantic cues but typically leave spatial structure underspecified.

Recent in-context image generators, including FLUX.2 (Black Forest Labs, 2025) and FLUX.1 Kontext (Black Forest Labs et al., 2025), accept a reference image alongside text; we refer to them as image/text-to-image (IT2I) models. These multimodal diffusion transformers use joint attention over reference image latent tokens, noisy image tokens, and text tokens (Meituan LongCat Team et al., 2025; Super Intelligence Team, 2026; Wu et al., 2025; Zhang et al., 2026). Given a measurement-derived reference $c_I = \mathcal{R}_{\mathcal{A}}(y)$ and a text prompt c_T, the model defines $p_\theta(x_0 \mid c_I, c_T)$; omitting the reference yields the text-only counterpart $p_\theta(x_0 \mid \varnothing, c_T)$. Image conditioning thus supplies spatial evidence directly to a measurement-dependent prior. Prior work fine-tunes IT2I models for restoration (Yılmaz et al., 2026); we keep the model frozen and use the original measurement and forward operator for data consistency in linear, nonlinear, and partially blind inverse problems.

This raises two questions: when image conditioning improves posterior sampling with a frozen IT2I model, and how reference influence should be regulated along the sampling trajectory. We address both by first examining how native IT2I generation uses degraded references. Whether a generation preserves the observed layout is associated with its reference influence: reference influence, measured by the relative velocity change upon removing the reference and by the reference attention share relative to text, is weaker in generations departing from the observed layout, and this difference can persist into later sampling steps. *Supplying a reference does not ensure its effective use.*

These observations motivate RefReg, a posterior sampler that uses frozen IT2I models as conditional generative priors and regulates reference influence at three levels. It constructs a reference through optimization to reduce degradation artifacts, then coordinates reference image guidance scheduling and early generation-to-text attention control to promote early structure preservation and subsequent detail refinement. Data-consistency updates to the predicted clean images further reduce disagreement with the original measurement throughout sampling.

To compare priors on a common footing, we derive an upper bound on average posterior approximation error for posterior sampling with conditional generative priors. Under matched sampling settings, the bounds differ only through the mean squared error (MSE) of clean image estimates after data consistency updates, evaluated against ground truth along each sampler's own trajectory. Evaluations using FlowDPS (Kim et al., 2025a) as a common posterior sampler under matched sampling settings show that image conditioning lowers the error bound throughout sampling across multiple IT2I backbones, and that the reduction concentrates in generations that use the reference from the initial steps. These findings make a case for IT2I priors over their text-only counterparts for posterior sampling from degraded measurements, with reference regulation as the condition under which the gain is realized. Our contributions are as follows:

- **RefReg:** Building on early structure preservation and the reduced reference influence in measurement-conditioned generation, we propose *RefReg*, which constructs a measurement-derived reference and regulates its influence in frozen IT2I models through guidance scheduling and early attention control. Data consistency updates complement it throughout sampling.

- **Posterior approximation error bound:** We derive an upper bound on average posterior approximation error in which the prior enters only through the clean-estimate error along sampler trajectories, without Lipschitz assumptions on the velocity field. Evaluations across inverse problems and backbones show image conditioning lowers this error bound throughout sampling under matched sampling settings, with the reduction concentrated in generations using the reference effectively.

- **Experimental results:** Across linear, nonlinear, and partially blind inverse problems, and against six flow-based solvers given the same measurement-derived reference, ***RefReg*** improves perceptual quality with competitive reconstruction fidelity. The gains carry over to three additional IT2I backbones and a distilled model, and RefReg remains robust with eight sampling steps.

2 BACKGROUND

2.1 GENERATIVE POSTERIOR SAMPLING FOR INVERSE PROBLEMS

Inverse problems aim to recover a clean image x_0 from a degraded measurement y modeled as

$$y = \mathcal{A}(x_0) + n, \qquad y, n \in \mathbb{R}^m, \quad x_0 \in \mathbb{R}^d, \tag{1}$$

where $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$ is a known forward operator and n is measurement noise. For white Gaussian noise $n \sim \mathcal{N}(0, \sigma_y^2 I)$ independent of x_0, the likelihood is $p(y \mid x_0) = \mathcal{N}(y; \mathcal{A}(x_0), \sigma_y^2 I)$. Posterior sampling targets $p(x_0 \mid y) \propto p(y \mid x_0)p(x_0)$, where classical priors encode assumptions such as local smoothness or sparse image gradients (Geman & Geman, 1984; Rudin et al., 1992). Diffusion and flow matching models learn expressive image priors from data (Ho et al., 2020; Lipman et al., 2023), enabling posterior inference with pretrained generative models.

Methods developed with unconditional diffusion priors incorporate measurement information through likelihood-score approximations (Chung et al., 2023; Song et al., 2023; Boys et al., 2024), such as the DPS approximation $\nabla_{x_t} \log p(y \mid x_t) \approx \nabla_{x_t} \log p(y \mid \hat{x}_{0|t})$, or data consistency updates (Kawar et al., 2022; Wang et al., 2023; Chung et al., 2024a; Zhu et al., 2023). Related approaches use MCMC or split Gibbs sampling (Zhang et al., 2025; Wu et al., 2024b), variational posterior approximation (Mardani et al., 2024), or initial-noise optimization (Wang et al., 2024). These strategies can also be adapted to text-conditional diffusion priors $p(x_0 \mid c_T)$, where c_T is a text prompt. To exploit the semantic information in these prompts more directly, subsequent methods incorporate prompt optimization, textual regularization, and score distillation (Chung et al., 2024b; Kim et al., 2025b; 2024). More recently, inverse solvers have extended the use of text-conditional priors to flow matching models (Patel et al., 2025; Kim et al., 2025a; Erbach et al., 2025; Park & Ye, 2025).

2.2 IMAGE/TEXT-TO-IMAGE MODELS

Recent large-scale image/text-to-image (IT2I) models, including FLUX.2 (Black Forest Labs, 2025) and FLUX.1 Kontext (Black Forest Labs et al., 2025), use multimodal diffusion transformers with joint attention over reference image latent tokens, noisy image tokens, and text tokens (Meituan LongCat Team et al., 2025; Super Intelligence Team, 2026; Wu et al., 2025; Zhang et al., 2026). Causal interventions in FLUX.2 show that reference information reaches generation tokens directly through image attention and indirectly through text token representations (Ge et al., 2026). Recent works adapt pretrained IT2I models for restoration through low-rank fine tuning (Yılmaz et al., 2026) or use posterior sampling guidance to preserve unedited source regions in image editing (Song et al., 2026).

We denote the conditional velocity field by $v_\theta(x_t, t, c_I, c_T)$, where c_I and c_T are the reference image and text prompt. We adopt the probability path defined by $x_t = (1-t)x_0 + tx_1$, where x_0 is a clean image and $x_1 \sim \mathcal{N}(0, I)$ is independent Gaussian noise (Lipman et al., 2023; Tong et al., 2024). Sampling integrates $\mathrm{d}x_t/\mathrm{d}t = v_\theta(x_t, t, c_I, c_T)$ from $t = 1$ to $t = 0$ with both conditions fixed. The velocity field gives the clean image estimate $\hat{x}_{0|t} = x_t - t\, v_\theta(x_t, t, c_I, c_T)$.

3 ANALYSIS: IT2I MODELS AS PRIORS FOR INVERSE PROBLEMS

Reference conditioning provides spatial information absent from text, but a degraded reference may have limited influence. We examine reference utilization in native conditional generation and its implications for posterior sampling. Throughout the analysis, we compare image-and-text conditioning with text-only conditioning within the same model:

$$v_{\mathrm{IT2I}}(x_t, t; y) = v_\theta(x_t, t, c_I, c_T), \quad v_{\mathrm{T2I}}(x_t, t; y) = v_\theta(x_t, t, \varnothing, c_T), \tag{2}$$

where $M \in \{\mathrm{IT2I}, \mathrm{T2I}\}$ indexes the two conditioning modes. We evaluate FLUX.2 and FLUX.1 Kontext (Black Forest Labs, 2025; Black Forest Labs et al., 2025) on Gaussian deblurring ($\sigma = 5$),

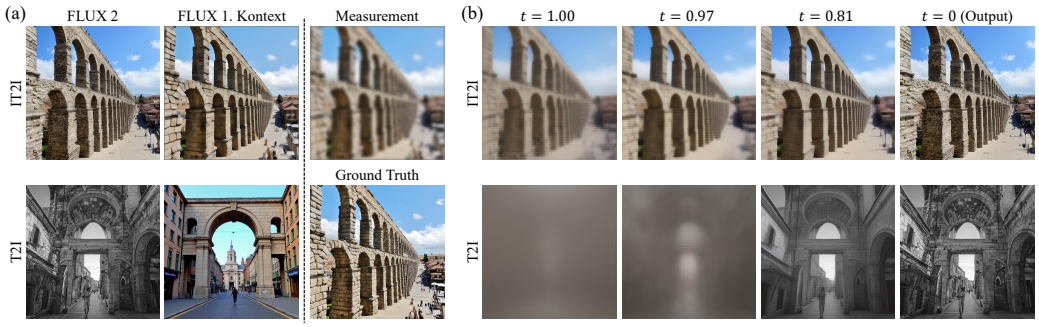

Figure 2: **Early structure preservation** in native IT2I generation. (a) FLUX.2 and FLUX.1 Kontext outputs with (IT2I) and without (T2I) image conditioning: IT2I preserves the observed layout while synthesizing details that differ from the ground truth. (b) Tweedie's estimates along FLUX.2 trajectories: with the reference, the layout is visible from the first step and later steps refine its details. Each pair shares the prompt and initial noise.

using the measurement directly as the reference ($c_I = y$). Each IT2I–T2I pair shares the same text prompt and initial noise.

3.1 Measurement-conditioned Generation in IT2I Models

Early structure preservation. Euler sampling with a degraded reference can preserve the observed layout while synthesizing fine detail (Figure 2(a)). It integrates the model's conditional velocity field without likelihood gradients or data consistency updates. With the same prompt and initial noise, the corresponding T2I outputs depart from the observed layout. Figure 2(b) shows Tweedie's estimates $\hat{x}^M_{0|t} = x^M_t - t\, v_M(x^M_t, t; y)$ along FLUX.2 trajectories. With reference conditioning, the initial estimate at $t = 1$ already shows the observed layout, and later estimates progressively refine details. Without it, structure emerges later and differs from the measurement. These observations suggest that the reference supplies scene-specific structure early, while subsequent sampling steps synthesize details within that structure. However, the synthesized details need not match the ground truth, motivating explicit data consistency to reduce disagreement with the measurement. Appendix B presents the corresponding observations for three additional IT2I backbones.

Reduced reference influence. Early structure preservation is not universal: Figure 3(a) shows some IT2I outputs that depart from the observed layout and offer little improvement in measurement consistency over the corresponding T2I outputs. To distinguish these outcomes without ground truth, we define $\Delta_{\mathrm{PSNR}} = \mathrm{PSNR}(\mathcal{A}(x_0^{\mathrm{IT2I}}), y) - \mathrm{PSNR}(\mathcal{A}(x_0^{\mathrm{T2I}}), y)$ for paired native outputs x_0^M. We label a run *reference-dominated* if $\Delta_{\mathrm{PSNR}} \geq 1\,\mathrm{dB}$ and *text-dominated* if $\Delta_{\mathrm{PSNR}} < 1\,\mathrm{dB}$. To examine how these groups differ during sampling, we use the first 100 DIV2K images in index order with five random seeds per image. We measure the relative velocity change upon removing the reference, r_t, and the reference attention share relative to text, ρ_t:

$$r_t = \frac{\|v_{\mathrm{IT2I}}(x_t, t; y) - v_{\mathrm{T2I}}(x_t, t; y)\|_2}{\|v_{\mathrm{IT2I}}(x_t, t; y)\|_2}, \qquad \rho_t = \frac{1}{L}\sum_{\ell=1}^{L} \frac{m_I^\ell(x_t)}{m_I^\ell(x_t) + m_T^\ell(x_t)}, \tag{3}$$

where L is the number of layers, and m_I^ℓ and m_T^ℓ are the attention masses from generation queries to reference-image and text keys, averaged over queries and heads with padding excluded. Both velocity fields are evaluated at the same IT2I state $x_t = x_t^{\mathrm{IT2I}}$ to isolate the effect of reference removal. For both FLUX.2 and FLUX.1 Kontext, Figure 3(b) shows that both quantities are higher on average in reference-dominated runs from the start of sampling. These results suggest that text-dominated runs already exhibit weaker reference influence during early structure formation.

3.2 Posterior Error Bound with Conditional Priors

To examine whether the benefits of native reference conditioning in Sec. 3.1 extend to posterior sampling, we derive an upper bound on average posterior approximation error and establish a sufficient condition for the IT2I prior to yield a smaller bound than its text-only counterpart. The

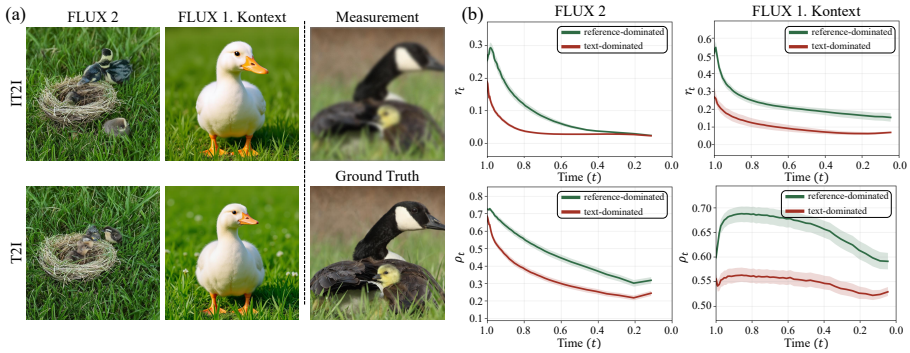

Figure 3: **Reduced reference utilization** in native IT2I generation. (a) A text-dominated example: IT2I outputs depart from the observed layout and resemble their T2I counterparts. (b) Relative velocity change upon reference removal (r_t) and reference attention share (ρ_t), grouped by measurement-consistency gain over T2I; both are lower in text-dominated runs from the first step.

score estimate corresponding to v_M in Eq. (2) is $s_M(x,t;y) = -[x + (1-t)v_M(x,t;y)]/t$. For $0 < \tau_0 < \tau_1 \leq 1$, we consider the reverse-time SDE formulation (Anderson, 1982; Song et al., 2021) with data-consistency guidance:

$$dX_t^M = \left[v_M(X_t^M, t; Y) - \frac{g_t^2}{2} s_M(X_t^M, t; Y) + G_t^M(X_t^M; Y) \right] dt + g_t \, d\overline{W}_t, \tag{4}$$

where time decreases from τ_1 to τ_0, $\overline{W}_t$ is Brownian motion in reverse time, and G_t^M denotes data consistency guidance. Both samplers share the initialization distribution, deterministic diffusion schedule g_t, and data consistency rule. The drift defines the clean image estimate after data consistency updates as $\widetilde{d}_M(x,t;y) = x - t v_M(x,t;y) - (t/\beta_t) G_t^M(x;y)$, with $\beta_t = 1 + g_t^2(1-t)/(2t)$.

Proposition 1 (Posterior approximation error bound)**.** *Let X_0 denote the clean ground-truth image associated with measurement Y, and let $\mathcal{L}$ denote probability law. Write $\pi_y = \mathcal{L}(X_0 \mid Y = y)$ for the target posterior and $q_t^{M,y} = \mathcal{L}(X_t^M \mid Y = y)$ for the distribution of the sampler in Eq. (4). Define $R_M(t) = \mathbb{E}[\|\widetilde{d}_M(X_t^M, t; Y) - X_0\|_2^2]$ as the mean squared error of clean image estimates after data-consistency updates relative to the clean ground truth X_0 along each sampler's own trajectory, averaged over data and sampler randomness. Under Assumptions A.1–A.3,*

$$\mathbb{E}_Y\left[W_2\left(q_{\tau_0}^{M,Y}, \pi_Y\right)\right] \leq C_{\text{stop}} + C_{\text{init}} + \int_{\tau_0}^{\tau_1} H_{\tau_0,t} \frac{\beta_t}{t} \sqrt{R_M(t)} \, dt, \tag{5}$$

where W_2 is the 2-Wasserstein distance and $H_{u,t} = \exp[-\int_u^t (1/r + g_r^2/(2r^2))\, dr]$. Both priors share the positive weight and the terms C_{stop} and C_{init}, which account for stopping at $\tau_0 > 0$ and the propagated initial coupling cost, respectively. (Definitions and proof in Appendix A.)

Remark 1 (Implications for comparing priors)**.** The bounds differ only through the weighted integral of $\sqrt{R_M(t)}$. In particular, $R_{\text{IT2I}}(t) \leq R_{\text{T2I}}(t)$ almost everywhere, with strict inequality on a set of positive measure, is sufficient for a smaller bound. More generally, the weighted difference of the square-root risks determines the ordering of the bounds. This comparison concerns the upper bounds and does not establish an ordering of the posterior approximation errors themselves.

Empirical validation. Remark 1 makes $R_M(t)$ the quantity to compare. Section 3.1 suggests that the priors separate in runs that use the reference. We test this with deterministic FlowDPS (Kim et al., 2025a) on the image–noise pairs of Figure 3 (Gaussian deblurring, 100 DIV2K images, five seeds, both backbones), grouped by their labels, with the measurement, prompt, initial noise, schedule, and data consistency settings shared within each IT2I–T2I pair. Along each trajectory, we measure $R_M(t)$ and report the relative difference $\Delta R_{\text{rel}}(t) = (R_{\text{IT2I}}(t) - R_{\text{T2I}}(t))/R_{\text{T2I}}(t)$ per group, which is negative exactly when the IT2I prior has the smaller error at t (Figure 4). In reference-dominated runs, $\Delta R_{\text{rel}}(t) < 0$ at every sampling time for both backbones, so the sufficient condition of Remark 1 holds on the whole sampling grid: for posterior sampling, the IT2I prior has a clear advantage over the T2I prior throughout sampling. In text-dominated runs, the reduction is smaller for both backbones. The advantage of image conditioning is thus larger in runs that use the reference from the initial steps. Appendix B repeats the comparison on three additional IT2I backbones.

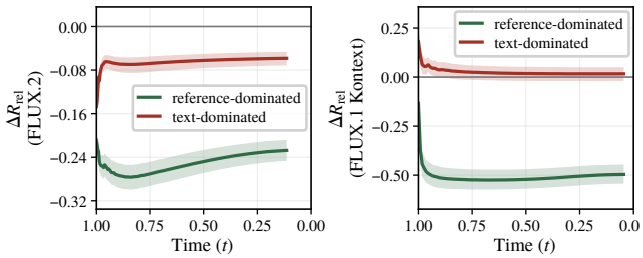

Figure 4: **Relative difference of the prediction errors of the IT2I and T2I priors.** $\Delta R_{\text{rel}}(t)$ for Gaussian deblurring with FLUX.2 and FLUX.1 Kontext. In reference-dominated runs, the IT2I prior has the smaller error at every sampling time; in text-dominated runs, the reduction is much smaller or absent.

4 Method: RefReg Framework

Building on Section 3, RefReg constructs a measurement-derived reference and regulates its influence through reference guidance scheduling and early generation-to-text attention control. Data consistency guidance complements this conditioning using the measurement and forward operator.

4.1 Measurement-derived reference construction

We construct c_I to preserve spatial information from the measurement y while reducing degradation artifacts. The construction is motivated by a regularized data-fidelity objective:

$$\min_{r \in \mathcal{C}} \frac{1}{2\sigma_y^2} \left\| \mathcal{A}(r) - y \right\|_2^2 + \lambda \mathcal{R}(r), \tag{6}$$

where r is a pixel-space image, $\mathcal{C}$ is the valid pixel range, and $\mathcal{R}$ is an image regularizer that suppresses degradation artifacts while preserving spatial structure. Starting from an image initialized from y, we alternate regularization and measurement-fidelity updates for a fixed number of iterations to construct c_I. Appendix C.1 specifies the regularizer, initialization, and operator-specific updates, with parameter values in Appendix D.2.1.

4.2 Controlling Reference Influence during Sampling

Section 3.1 shows that reference supported structure can emerge early, whereas generations departing from the observed layout exhibit weaker reference influence from the start. This motivates emphasizing references during early structure formation and allowing later detail refinement.

Reference image guidance scheduling. We first schedule reference image guidance using the velocity form of the two-scale classifier-free guidance formulation of Brooks et al. (2023):

$$\begin{aligned} \tilde{v}_\theta(x_t, t, c_I, c_T) = v_\theta(x_t, t, \varnothing, \varnothing) &+ s_I(t)\big[v_\theta(x_t, t, c_I, \varnothing) - v_\theta(x_t, t, \varnothing, \varnothing)\big] \\ &+ s_T(t)\big[v_\theta(x_t, t, c_I, c_T) - v_\theta(x_t, t, c_I, \varnothing)\big]. \end{aligned} \tag{7}$$

Here, $\varnothing$ denotes a null condition, $s_I(t)$ controls the image contribution, and $s_T(t)$ controls the additional text contribution given c_I. We use stronger image guidance early and decrease $s_I(t)$ as sampling proceeds, aiming to preserve observed structure while limiting the influence of residual reference degradations on later detail synthesis.

Early generation-to-text attention control. Guidance scheduling reweights conditional velocity predictions but leaves the attention interactions within each prediction unchanged. To complement this control during early structure formation, we additionally control generation-to-text connections within joint self-attention (Black Forest Labs et al., 2025; Ge et al., 2026). An attention-logit bias blocks generation queries from attending to selected text keys when $t \geq t_w$ (Geva et al., 2023). We remove the bias when $t < t_w$, restoring direct access to these tokens during detail refinement. Auxiliary text controls reduce indirect transmission of prompt content. Both controls are detailed in Appendix C.2, with parameter settings in Appendix D.2.2.

4.3 Overall Sampling Procedure with Data Consistency Guidance

Together, the above components encourage $\hat{x}_{0|t}$ to capture measurement-supported structure before explicit correction, reducing the structural mismatch that local data-consistency updates must

Algorithm 1 RefReg

Require: Measurement y, operator $\mathcal{A}$, text c_T, frozen v_θ, sampling times $1 = t_0 > \cdots > t_N = 0$, schedules $s_I(t), s_T(t)$, window t_w

$c_I \leftarrow$ reference constructed from y (Appendix C.1)
$x_{t_0} \sim \mathcal{N}(0, I)$
for $i = 0, \ldots, N-1$ **do**
  Use the conditioning controls in Section 4.2 and Appendix C.2
  $\tilde{v} \leftarrow \tilde{v}_\theta(x_{t_i}, t_i, c_I, c_T)$ using equation 7
  $\hat{x}_{0|t_i} \leftarrow x_{t_i} - t_i \tilde{v}$; $\hat{x}_{1|t_i} \leftarrow x_{t_i} + (1 - t_i)\tilde{v}$
  $\widetilde{x}_{0|t_i} \leftarrow \arg\min_x \ell_y(x)$ ▷ init. at $\hat{x}_{0|t_i}$
  $x_{t_{i+1}} \leftarrow (1 - t_{i+1})\widetilde{x}_{0|t_i} + t_{i+1}\hat{x}_{1|t_i}$
end for
return x_{t_N}

resolve. At each timestep, we reduce the data-consistency objective $\ell_y(x) := \mathcal{L}(\mathcal{A}(x), y)$ from the clean estimate and reinterpolate the corrected estimate with the retained noise endpoint $\hat{x}_{1|t} = x_t + (1 - t)\tilde{v}_\theta$. Algorithm 1 summarizes the procedure, where $\arg\min$ denotes a fixed number of gradient iterations initialized at the clean estimate; the discrepancy $\mathcal{L}$, update rule are detailed in Appendices C.3 and D.2.3.

5 Experiments

5.1 Experimental Setup

Problem setting. We evaluate linear inverse problems: super resolution at $\times 8$ and $\times 16$, box and 90% random inpainting, Gaussian deblurring with $\sigma = 5$, and motion deblurring with a 61×61 kernel. Nonlinear tasks include deblurring (Tran et al., 2021; Zhang et al., 2025) and high dynamic range reconstruction. We add Gaussian noise with $\sigma_y = 0.03$ to all measurements on the $[-1, 1]$ pixel scale. Text conditioning uses the fixed prompt "a high quality photo of a face, clean, high-resolution, 8k" for FFHQ and DAPE captions (Wu et al., 2024a) for each DIV2K image.

Datasets and metrics. We evaluate 800 DIV2K (Agustsson & Timofte, 2017) and 1,000 FFHQ images (Karras et al., 2019) at 768^2 resolution. We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and FID (Heusel et al., 2017) using non-overlapping 256^2 patches.

Baselines. We compare against six flow-based inverse solvers: FlowChef (Patel et al., 2025), FlowDPS (Kim et al., 2025a), Flower (Pourya et al., 2026), FLAIR (Erbach et al., 2025), MPC (Webber et al., 2026), and TriPS (Bang et al., 2026). All methods use FLUX.2-Klein-Base-4B with 50 NFEs. Implementation details are provided in Appendix D.

5.2 Experimental Results

Inverse problems. Table 1 shows that our method achieves the lowest LPIPS and FID in most reported settings, with competitive reconstruction fidelity. Figure 5 presents representative qualitative comparisons. Appendix E.1 reports all eight tasks, with further examples in Appendix F.

Reconstruction quality at low NFEs. Table 2 evaluates reconstruction with fewer NFEs on 100 DIV2K images. With as few as 8 NFEs, RefReg stays near its 50 NFE quality, whereas removing image conditioning degrades it more, consistent with the early structure preservation in Section 3.1. Computational costs and detailed settings are provided in Appendix E.5. Additional experiments using the distilled FLUX.2-klein-4B model at 4 NFEs are presented in Appendix E.6.

Partially blind inverse problems. Table 3 compares the partially blind extension of RefReg with GDP (Fei et al., 2023) and DMPlug (Wang et al., 2024), both adapted to the same IT2I model. For Gaussian deblurring and HDR reconstruction, it attains higher PSNR and lower LPIPS and operator errors, supporting joint image and operator estimation. Task formulations, estimation procedures, and settings are in Appendix D.4, with qualitative comparisons in Appendix F.

Table 1: **Quantitative comparison** averaged over 800 DIV2K and 1,000 FFHQ samples using 50 NFEs under additive Gaussian noise ($\sigma_y = 0.03$). (All eight tasks: Table A6)

Task	Method	DIV2K (768 × 768)				FFHQ (768 × 768)			
		PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Super-Res. ×8	FlowChef	22.06	0.510	0.392	68.34	27.54	0.706	0.308	92.78
	FlowDPS	21.78	0.584	0.550	82.14	25.86	0.784	0.385	83.59
	Flower	20.01	0.516	0.467	69.39	23.09	0.745	0.334	63.51
	FLAIR	22.73	0.587	0.348	25.79	28.43	0.752	0.235	22.76
	MPC	22.31	0.555	0.406	37.09	28.46	0.765	0.327	55.07
	TriPS	22.29	0.575	0.467	42.40	28.79	0.794	0.300	39.23
	Ours	**23.01**	**0.600**	**0.266**	**15.36**	**29.85**	**0.806**	**0.194**	**13.56**
Gaussian Deblur	FlowChef	21.54	0.538	0.467	97.68	27.16	0.757	0.354	86.35
	FlowDPS	19.09	0.512	0.650	130.86	22.28	0.737	0.454	88.49
	Flower	21.29	0.472	0.476	81.97	26.05	0.636	0.362	66.21
	FLAIR	19.80	0.487	0.392	25.14	24.69	0.673	0.255	16.43
	MPC	20.76	0.498	0.398	37.12	26.81	0.745	0.272	27.14
	TriPS	20.99	0.499	0.367	20.37	27.32	0.740	0.229	16.40
	Ours	**22.13**	**0.563**	**0.324**	**19.51**	**28.64**	**0.773**	**0.217**	**16.21**
Inpaint (Random)	FlowChef	23.96	0.709	0.234	43.76	30.68	0.852	0.212	57.15
	FlowDPS	22.29	0.639	0.314	61.14	28.55	0.807	0.216	83.06
	Flower	16.36	0.366	0.496	154.85	18.16	0.367	0.569	263.74
	FLAIR	23.78	0.695	0.272	88.05	29.73	0.814	0.226	111.14
	MPC	22.76	0.632	0.317	48.99	30.05	0.826	0.179	55.52
	TriPS	24.54	0.714	0.226	19.39	**31.99**	**0.872**	0.166	15.28
	Ours	**24.61**	**0.722**	**0.169**	**13.32**	31.70	0.854	**0.108**	**9.04**
Nonlinear Deblur	FlowChef	**26.79**	**0.779**	**0.148**	15.64	**30.75**	**0.843**	0.183	27.50
	FlowDPS	21.51	0.641	0.387	41.37	24.55	0.803	0.283	32.85
	Flower	23.71	0.643	0.277	37.59	27.55	0.715	0.244	32.72
	FLAIR	23.57	0.656	0.208	**13.62**	28.44	0.778	0.154	12.31
	MPC	23.98	0.695	0.225	23.23	29.00	0.824	0.181	20.87
	TriPS	23.47	0.661	0.286	21.56	28.92	0.820	0.208	17.84
	Ours	25.20	0.724	0.186	14.44	29.67	0.805	**0.144**	**10.63**
HDR	FlowChef	24.85	0.820	0.108	10.98	24.32	0.844	0.139	18.38
	FlowDPS	18.08	0.707	0.193	12.91	18.96	0.777	0.189	15.95
	Flower	21.43	0.802	0.112	8.87	20.89	0.816	0.131	13.91
	FLAIR	24.02	0.783	0.157	12.96	24.66	0.822	0.180	21.08
	MPC	24.22	0.813	0.109	8.79	23.73	0.846	0.112	11.96
	TriPS	23.87	0.750	0.191	9.86	25.12	0.827	0.149	7.72
	Ours	**26.92**	**0.847**	**0.068**	**4.93**	**28.89**	**0.861**	**0.066**	**4.32**

Table 2: **Reconstruction quality at low NFEs** on 100 DIV2K images. Parentheses show changes from each variant's 50 NFE result. IC: image conditioning.

NFEs	IC	SR ×16 PSNR↑	SR ×16 LPIPS↓	Gaussian deblurring PSNR↑	Gaussian deblurring LPIPS↓
25	✗	18.74 (−0.60)	0.5019 (−0.0002)	20.63 (−0.64)	0.4523 (+0.0013)
	✓	**20.11** (−0.18)	**0.3628** (−0.0081)	**21.92** (−0.23)	**0.3037** (−0.0174)
15	✗	18.47 (−0.86)	0.5101 (+0.0080)	20.19 (−1.07)	0.4614 (+0.0105)
	✓	**20.09** (−0.20)	**0.3592** (−0.0116)	**21.88** (−0.27)	**0.2980** (−0.0231)
8	✗	18.54 (−0.79)	0.5320 (+0.0299)	19.92 (−1.35)	0.4900 (+0.0390)
	✓	**20.02** (−0.27)	**0.3832** (+0.0123)	**21.71** (−0.44)	**0.3093** (−0.0118)

Table 3: **Partially blind reconstruction** on 100 DIV2K images. We jointly estimate the clean image and either the blur kernel for Gaussian deblurring or the exposure scale for HDR reconstruction. Operator errors are $E_{\mathbf{k}} = \|\hat{\mathbf{k}} - \mathbf{k}^\star\|_2 / \|\mathbf{k}^\star\|_2$ and $E_s = |\hat{s} - s^\star|$.

Method	Gaussian deblurring PSNR↑	Gaussian deblurring LPIPS↓	Gaussian deblurring $E_{\mathbf{k}}$↓	HDR reconstruction PSNR↑	HDR reconstruction LPIPS↓	HDR reconstruction E_s↓
GDP	19.61	0.289	1.659	21.32	0.130	0.463
DMPlug	21.52	0.330	0.701	21.17	0.136	0.607
Ours	**23.71**	**0.283**	**0.175**	**24.67**	**0.077**	**0.170**

5.3 Ablation Studies

Table 4 reports the component ablations of RefReg. All variants retain data consistency (DC) updates. Qualitative comparisons are provided in Appendix E.3.

Image conditioning (IC). Removing c_I degrades reconstruction despite DC updates: the reference supplies scene structure, while DC enforces measurement agreement. Appendix E.2 separately examines reference input and optimization.

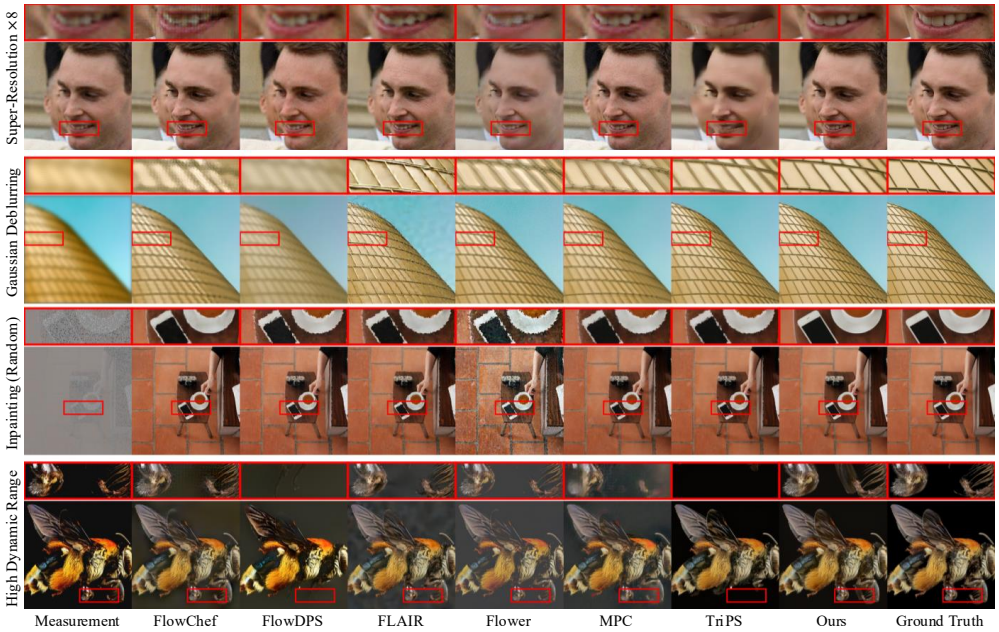

Figure 5: **Qualitative comparison** on FFHQ and DIV2K (All eight tasks: Figure A5).

Reference image guidance scheduling (IGS). Replacing the schedule $s_I(t)$ with fixed unit image guidance ($s_I = 1$) reduces reconstruction quality. The schedule emphasizes the reference early and relaxes it during refinement, which limits the transfer of degraded appearance from the reference.

Early generation-to-text attention blocking (EAB). Removing early control of attention from generation to text (Section 4.2) allows generation queries to attend to all text keys throughout sampling and degrades reconstruction even with IGS. Blocking selected text keys limits textual influence during early structure formation. Appendix E.4 examines the timing of this intervention.

Table 4: **Ablation studies** on 100 DIV2K images. For IGS, ✗ denotes $s_I = 1$.

IC	IGS	EAB	Motion deblurring		Inpainting (90%)	
			PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
✓	✓	✓	**27.629**	**0.1563**	**24.659**	**0.1648**
✗	–	✓	25.158	0.3128	19.949	0.5318
✓	✗	✓	27.055	0.1574	24.317	0.1882
✓	✓	✗	27.177	0.1907	23.890	0.2323

5.4 Discussion

We evaluate RefReg on three additional backbones: FLUX.1 Kontext, FireRed-Image-Edit, and LongCat-Image-Edit. The evaluation covers super-resolution $\times 8$, Gaussian deblurring, and motion deblurring (Table A11). Our method achieves the best LPIPS in all nine settings and the best PSNR in five, including all three tasks on FLUX.1 Kontext. These results are consistent with those on FLUX.2 and support the applicability of our method to large IT2I models (Appendix E.7).

6 Conclusion

In this paper, we presented RefReg, which uses frozen in-context image generation models as a conditional generative prior. It regulates reference influence through reference construction, guidance scheduling, and early attention control, complemented by data-consistency updates. Our analysis establishes a sufficient condition for image conditioning to yield a smaller posterior error bound. Experiments across linear, nonlinear, and partially blind inverse problems demonstrate improved perceptual quality with competitive fidelity across multiple backbones.

Limitation. Our method inherits the limitations of the pretrained models and may generate inaccurate details when measurements provide insufficient information. We follow the same principle of regulating reference influence across backbones, but implementation requires adjustments for differences in reference image encoding and token representation.

AI USE STATEMENT

We used generative AI tools to support method implementation and experimental coding, as well as data analysis and interpretation. These tools also assisted with drafting and cross-checking mathematical proofs and identifying relevant literature. During manuscript preparation, AI assisted with translation, editing for clarity, and LaTeX formatting. We also used AI feedback to review the paper's organization, logical consistency, and scope of claims. Four authors reviewed and cross-checked all AI-assisted work. We take full responsibility for the accuracy and integrity of the final manuscript and accompanying materials.

REPRODUCIBILITY STATEMENT

We provide the assumptions and proofs for our theoretical results in Appendix A. The proposed algorithm is described in Section 4 and Algorithm 1, with detailed procedures in Appendix C. Section 5.1 and Appendix D describe the experimental setup, including data preprocessing, forward operators, evaluation protocols, and implementation settings for our method and the baselines. Settings for additional experiments and ablations are provided in Appendix E. We will release the code upon acceptance of the paper.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, July 2017.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Junseo Bang, Dong Ju Mun, Hoigi Seo, Seongmin Hong, and Se Young Chun. Triadic dynamics aware diffusion posterior sampling for inverse problems: Optimizing guidance and stochasticity schedules. In *ICML*, 2026.

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

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, 2025. URL `https://arxiv.org/abs/2506.15742`.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and O. Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *TMLR*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pp. 18392–18402, 2023.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023.

Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *ICLR*, 2024a.

Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. In *ICML*, 2024b.

Julius Erbach, Dominik Narnhofer, Andreas Dombos, Bernt Schiele, Jan Eric Lenssen, and Konrad Schindler. Solving inverse problems with flair. *NeurIPS*, 38:136667–136704, 2025.

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, pp. 9935–9946, 2023.

Tomer Garber and Tom Tirer. Image restoration by denoising diffusion models with iteratively preconditioned guidance. In *CVPR*, 2024.

Chris Ge, Rohit Gandikota, Antonio Torralba, and Tamar Rott Shaham. Vision-language binding in in-context image generation. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE TPAMI*, 6(6):721–741, 1984.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *EMNLP*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pp. 4396–4405, 2019.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *NeurIPS*, 35:23593–23606, 2022.

Jeongsol Kim, Geon Yeong Park, and Jong Chul Ye. DreamSampler: Unifying diffusion sampling and score distillation for image manipulation. In *ECCV*, 2024.

Jeongsol Kim, Bryan Sangwoo Kim, and Jong Chul Ye. Flowdps: Flow-driven posterior sampling for inverse problems. In *ICCV*, pp. 12328–12337, 2025a.

Jeongsol Kim, Geon Yeong Park, Hyungjin Chung, and Jong Chul Ye. Regularization by texts for latent diffusion inverse solvers. In *ICLR*, 2025b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Muyu Liu, Xuanyu Tian, Chenhe Du, Qing Wu, Hongjiang Wei, and Yuyao Zhang. Resolving blind inverse problems under dynamic range compression via structured forward operator modeling. *arXiv preprint arXiv:2603.01890*, 2026. URL https://arxiv.org/abs/2603.01890.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *ICLR*, 2024.

Meituan LongCat Team et al. LongCat-Image technical report. *arXiv preprint arXiv:2512.07584*, 2025. URL https://arxiv.org/abs/2512.07584.

Jonghyun Park and Jong Chul Ye. FlowLPS: Langevin-proximal sampling for flow-based inverse problem solvers. *arXiv preprint arXiv:2512.07150*, 2025.

Maitreya Patel, Song Wen, Dimitris N. Metaxas, and Yezhou Yang. Flowchef: Steering of rectified flow models for controlled generations. In *ICCV*, 2025.

Mehrsa Pourya, Bassam El Rawas, and Michael Unser. Flower: A flow-matching solver for inverse problems. In *ICLR*, 2026.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *ICLR*, 2024.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*, 2023.

Rui Song, Guo-Hua Wang, Qing-Guo Chen, Weihua Luo, Tongda Xu, Zhening Liu, Yan Wang, Zehong Lin, and Jun Zhang. Training-free image editing with visual context integration and concept alignment. *arXiv preprint arXiv:2604.04487*, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Super Intelligence Team. FireRed-Image-Edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026. URL https://arxiv.org/abs/2602.13344.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *TMLR*, 2024.

Phong Tran, Anh Tuan Tran, Quynh Phung, and Minh Hoai. Explore image deblurring via encoded blur kernel space. In *CVPR*, pp. 11956–11965, 2021.

Hengkang Wang, Xu Zhang, Taihui Li, Yuxiang Wan, Tiancong Chen, and Ju Sun. DMPlug: A plug-in method for solving inverse problems with diffusion models. In *NeurIPS*, 2024.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023.

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

George Webber, Alexander Denker, Riccardo Barbano, and Andrew J Reader. Solving inverse problems with flow-based models via model predictive control. 2026.

Chenfei Wu et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pp. 25456–25467, 2024a.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine L Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. In *NeurIPS*, 2024b.

M Akın Yılmaz, Ahmet Bilican, Burak Can Biner, and A Murat Tekalp. Edit2restore: Few-shot image restoration via parameter-efficient adaptation of pre-trained editing models. *arXiv preprint arXiv:2601.03391*, 2026.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *CVPR*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Xinjie Zhang et al. Mage-Flow: An efficient native-resolution foundation model for image generation and editing. *arXiv preprint arXiv:2607.19064*, 2026. URL https://arxiv.org/abs/2607.19064.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *CVPRW*, 2023.

APPENDIX

A THEORETICAL ANALYSIS: ASSUMPTIONS, PROOFS, AND ADDITIONAL VALIDATION

This appendix makes Proposition 1 precise and proves it. Appendix A.1 fixes the notation, defines the constants C_{init} and C_{stop}, and states the assumptions. Appendix A.2 gives the proof and the special case evaluated in the experiments. Appendix A.3 proves the comparison of Remark 1 and specifies how the bound is estimated from sampler trajectories, and Appendix B validates the comparison on three additional IT2I backbones.

A.1 SETTING, NOTATION, AND ASSUMPTIONS

Notation. $X_0 \in \mathbb{R}^d$ is the clean image, Y the measurement, and $X_1 \sim \mathcal{N}(0, I)$ the Gaussian noise of the probability path of Section 2.2 (written x_1 there), independent of (X_0, Y). We write $X_t = (1-t)X_0 + tX_1$ for the interpolant and $p_t^y = \mathcal{L}(X_t \mid Y = y)$ for its conditional law, so that $p_0^y = \pi_y$. Equation (4) runs from τ_1 down to τ_0. With $S = \tau_1 - \tau_0$, $t(s) = \tau_1 - s$, and $\widehat{X}_s^M = X_{t(s)}^M$, it is the forward-time Itô equation

$$d\widehat{X}_s^M = -\, b_{t(s)}^M(\widehat{X}_s^M; Y)\, ds + g_{t(s)}\, dB_s, \qquad 0 \le s \le S, \tag{A.1}$$

with a standard Brownian motion B and the drift

$$b_t^M(x;y) = \beta_t v_M(x,t;y) + \frac{g_t^2}{2t}\, x + G_t^M(x;y) = a_t\, x - c_t\, \widetilde{d}_M(x,t;y), \quad a_t = \frac{1}{t} + \frac{g_t^2}{2t^2}, \quad c_t = \frac{\beta_t}{t}. \tag{A.2}$$

The first form substitutes $s_M(x,t;y) = -[x + (1-t)v_M(x,t;y)]/t$ into Eq. (4); the second follows from the definition $\widetilde{d}_M(x,t;y) = x - tv_M(x,t;y) - (t/\beta_t)G_t^M(x;y)$ in Section 3, which gives $\beta_t v_M = c_t(x - \widetilde{d}_M) - G_t^M$. We write $e_t^M = \widetilde{d}_M(X_t^M, t; Y) - X_0$ for the corrected-estimate error along the sampler trajectory, so that $R_M(t) = \mathbb{E}\|e_t^M\|_2^2$, and

$$H_{u,v} = \exp\Big(-\int_u^v a_r\, dr\Big) = \frac{u}{v}\exp\Big(-\frac{1}{2}\int_u^v \frac{g_r^2}{r^2}\, dr\Big), \qquad \tau_0 \le u \le v \le \tau_1, \tag{A.3}$$

which is the weight $H_{\tau_0,t}$ of Proposition 1. All random variables take values in $\mathbb{R}^d$, the space in which the backbone generates, and $R_M(t)$ is measured there. The text prompt is fixed and suppressed from the notation; with image-specific prompts, every conditional statement below conditions on (Y, c_T) in place of Y.

Initialization and the constants of the bound. Both samplers start from the same initial law μ_{τ_1}, which has a finite second moment: $X_{\tau_1}^M = U \sim \mu_{\tau_1}$. For the proof, (U, X_1) is any coupling with these marginals that is independent of (X_0, Y) (for Gaussian initialization $\mu_{\tau_1} = \mathcal{N}(0, I)$ one may take $U = X_1$), and B is independent of (X_0, Y, U, X_1); neither choice changes the law of either sampler given Y. The constants in Eq. (5) are

$$C_{\text{init}} = H_{\tau_0,\tau_1}\big(\mathbb{E}\|U - X_{\tau_1}\|_2^2\big)^{1/2}, \qquad C_{\text{stop}} = \mathbb{E}_Y\big[W_2(p_{\tau_0}^Y, \pi_Y)\big]. \tag{A.4}$$

C_{init} is the cost of coupling the initial state with the interpolant at τ_1, contracted by H_{τ_0,τ_1} over the sampling interval; C_{stop} is the distance between the interpolant law at the stopping time τ_0 and the posterior. Both depend on μ_{τ_1}, g, τ_0, τ_1, and the data, but not on M.

Assumption A.1 (Schedule)**.** $g : [\tau_0, \tau_1] \to [0, \infty)$ is deterministic with $\int_{\tau_0}^{\tau_1} g_t^2\, dt < \infty$.

Since $\tau_0 > 0$, a_t and c_t are then positive and integrable on $[\tau_0, \tau_1]$, and $0 < H_{u,v} \le 1$. The assumption holds for the deterministic sampler ($g \equiv 0$) used in our experiments and for every bounded schedule.

Assumption A.2 (Well-posed sampler)**.** $v_M(x,t;y)$ and $G_t^M(x;y)$ are Borel functions of (x,t,y) that depend neither on the trajectory history nor on auxiliary randomness. For each M, Eq. (A.1) started from U has a unique continuous adapted strong solution with finite second moments at every time, and $\mathbb{E}\|X_0\|_2^2 < \infty$.

The first part states that the samplers of Section 3 are Markov in their state (a velocity network and a data-consistency correction of the current state); the second part is the minimum needed for $q_t^{M,y}$ and the W_2 distances to be defined. No Lipschitz constant of the learned velocity field is assumed: the nonlinearity of the network and any departure of the trajectory from the interpolant enter the bound only through $R_M(t)$, which is measured on the trajectory itself.

Assumption A.3 (Integrable corrected-estimate error)**.** For $M \in \{\mathrm{IT2I}, \mathrm{T2I}\}$, $R_M(t) < \infty$ for almost every t and $\int_{\tau_0}^{\tau_1} H_{\tau_0,t}\, c_t \sqrt{R_M(t)}\, dt < \infty$.

This is finiteness of the prior-dependent term of the bound; measurability of $t \mapsto R_M(t)$ follows from the joint measurability of $(t, \omega) \mapsto e_t^M$ under Assumption A.2.

With these definitions, Proposition 1 reads as follows.

Proposition 1 (restated). *Under Assumptions A.1–A.3, for $M \in \{\mathrm{IT2I}, \mathrm{T2I}\}$,*

$$\mathbb{E}_Y\big[W_2\big(q_{\tau_0}^{M,Y}, \pi_Y\big)\big] \le C_{\mathrm{stop}} + C_{\mathrm{init}} + \int_{\tau_0}^{\tau_1} H_{\tau_0,t}\, \frac{\beta_t}{t} \sqrt{R_M(t)}\, dt, \tag{A.5}$$

with C_{init} and C_{stop} as in Eq. (A.4). Moreover, $C_{\mathrm{stop}} \le \tau_0 \sqrt{d + \mathbb{E}\|X_0\|_2^2}$, and for Gaussian initialization with $U = X_1$, $C_{\mathrm{init}} = H_{\tau_0,\tau_1}(1-\tau_1)\sqrt{d + \mathbb{E}\|X_0\|_2^2}$, which vanishes when $\tau_1 = 1$.

A.2 Proof of Proposition 1

Proof idea. The proof has three steps. First, we construct an auxiliary process Z that is driven by the same Brownian motion as the sampler, whose drift is the sampler drift (A.2) with the corrected estimate $\widetilde{d}_M$ replaced by the clean image X_0, and whose law given $Y = y$ is exactly p_t^y (Lemma A.1). Because the two processes share the diffusion term, their difference solves a random linear differential equation whose forcing term is precisely $c_t e_t^M$. Second, solving this equation and applying Minkowski's inequality bounds the conditional mean-square distance between $X_{\tau_0}^M$ and Z_{τ_0}, hence $W_2(q_{\tau_0}^{M,y}, p_{\tau_0}^y)$, by the propagated initial cost plus a weighted integral of the conditional root-mean-square error. Third, averaging over Y with Jensen's inequality replaces the conditional errors by $\sqrt{R_M(t)}$, and the triangle inequality for W_2 at the stopping time adds C_{stop}. The auxiliary process is a device of the proof only: it is never run by a solver, and its drift is anchored at X_0 rather than at the posterior drift, which is what makes the prior-dependent term observable along each sampler's own trajectory.

Lemma A.1 (Auxiliary process anchored at the clean image)**.** *Let Z solve, in reverse time and with the Brownian motion of the samplers,*

$$dZ_t = \big[a_t Z_t - c_t X_0\big]dt + g_t\, d\overline{W}_t, \qquad Z_{\tau_1} = X_{\tau_1} = (1-\tau_1)X_0 + \tau_1 X_1. \tag{A.6}$$

Then, for all $t \in [\tau_0, \tau_1]$: (i) $\mathcal{L}(Z_t \mid X_0, Y) = \mathcal{N}\big((1-t)X_0,\, t^2 I\big)$, and hence $\mathcal{L}(Z_t \mid Y = y) = p_t^y$; (ii) $b_t^M(x;Y) - \big[a_t x - c_t X_0\big] = -c_t\big(\widetilde{d}_M(x,t;Y) - X_0\big)$ for every x; in particular, the drift mismatch along the sampler trajectory is $-c_t e_t^M$.

Proof. (i) In forward time, $\widehat{Z}_s = Z_{t(s)}$ solves the linear equation $d\widehat{Z}_s = \big[-a_{t(s)}\widehat{Z}_s + c_{t(s)}X_0\big]ds + g_{t(s)}\, dB_s$, which has a unique strong solution under Assumption A.1. Conditionally on (X_0, Y), the initial value is Gaussian and independent of B, and the solution is an affine function of the initial value and of a Wiener integral, hence conditionally Gaussian. Its conditional mean m_s and covariance V_s satisfy $\dot{m}_s = -a_t m_s + c_t X_0$ and $\dot{V}_s = -2a_t V_s + g_t^2 I$ with $t = t(s)$, $m_0 = (1-\tau_1)X_0$, and $V_0 = \tau_1^2 I$. By (A.2), $c_t - a_t(1-t) = 1$ and $g_t^2 - 2a_t t^2 = -2t$, while $dt/ds = -1$ gives $\frac{d}{ds}[(1-t)X_0] = X_0$ and $\frac{d}{ds}[t^2 I] = -2tI$; hence $m_s = (1 - t(s))X_0$ and $V_s = t(s)^2 I$ solve both equations with the correct initial values, and uniqueness for linear equations identifies them as the conditional mean and covariance. Averaging $\mathcal{N}((1-t)x_0, t^2 I)$ over $x_0 \sim \pi_y$ gives p_t^y. (For $g_t \not\equiv 0$ this is an identity of marginals only; Z_t is not the interpolant with a fixed X_1.) (ii) Subtract $a_t x - c_t X_0$ from the second form of (A.2), and evaluate at $x = X_t^M$. □

Why not couple with the interpolant X_t itself? When $g \equiv 0$, $Z_t = X_t$ and nothing is lost. When $g \not\equiv 0$, $X_t = (1-t)X_0 + tX_1$ has no Brownian component, so the difference $X_t^M - X_t$ would

retain the diffusion term $g_t\, d\overline{W}_t$ of the sampler together with the drift $\frac{g_t^2}{2t} X_1$ (since $a_t X_t - c_t X_0 = (X_1 - X_0) + \frac{g_t^2}{2t} X_1$), and both add terms of order $\sqrt{d}$ to the mean-square distance that do not depend on the prior. The process Z is the interpolant re-expressed with the sampler's noise: it has the same conditional law as X_t at every time, but it shares the Brownian motion of the sampler, so the noise cancels exactly in the difference.

Proof of Proposition 1. Fix M. Run the sampler from $X^M_{\tau_1} = U$ and the process Z of Lemma A.1 from $Z_{\tau_1} = X_{\tau_1}$, both driven by the same Brownian motion.

Coupling. The initial state and the Brownian motion of the sampler have the laws prescribed by the sampling protocol and are independent of (X_0, Y), so $\mathcal{L}(X^M_t \mid Y = y) = q^{M,y}_t$; by Lemma A.1(i), $\mathcal{L}(Z_t \mid Y = y) = p^y_t$. For fixed y, $(X^M_{\tau_0}, Z_{\tau_0})$ is therefore a coupling of $(q^{M,y}_{\tau_0}, p^y_{\tau_0})$, and by the definition of W_2 as an infimum over couplings,

$$W_2\big(q^{M,y}_{\tau_0}, p^y_{\tau_0}\big) \le \Big(\mathbb{E}\big[\|X^M_{\tau_0} - Z_{\tau_0}\|_2^2 \,\big|\, Y = y\big]\Big)^{1/2}. \tag{A.7}$$

Difference equation. Let $\xi_s = \widehat{X}^M_s - \widehat{Z}_s$ with $t = t(s)$. Subtracting the forward-time equations of the two processes, the identical diffusion terms cancel, so ξ is almost surely absolutely continuous, and Lemma A.1(ii) gives

$$\frac{d\xi_s}{ds} = -a_t\, \xi_s + c_t\, e^M_t. \tag{A.8}$$

Only the Brownian term has disappeared: e^M_t is evaluated on the random trajectory of the sampler and carries the nonlinearity of the network. Variation of constants and the substitution $t = \tau_1 - s$ yield the pathwise identity

$$X^M_{\tau_0} - Z_{\tau_0} = H_{\tau_0,\tau_1}\big(U - X_{\tau_1}\big) + \int_{\tau_0}^{\tau_1} H_{\tau_0,t}\, c_t\, e^M_t\, dt, \tag{A.9}$$

whose integral is almost surely finite because $\mathbb{E}\int_{\tau_0}^{\tau_1} H_{\tau_0,t} c_t \|e^M_t\|_2\, dt \le \int_{\tau_0}^{\tau_1} H_{\tau_0,t} c_t \sqrt{R_M(t)}\, dt < \infty$ by Tonelli's theorem, Jensen's inequality, and Assumption A.3.

Conditional mean-square bound. Taking the L^2 norm of (A.9) conditionally on $Y = y$ and applying Minkowski's integral inequality,

$$\begin{aligned}\Big(\mathbb{E}\big[\|X^M_{\tau_0} - Z_{\tau_0}\|_2^2 \,\big|\, Y = y\big]\Big)^{1/2} &\le H_{\tau_0,\tau_1}\Big(\mathbb{E}\big[\|U - X_{\tau_1}\|_2^2 \,\big|\, Y = y\big]\Big)^{1/2} \\ &\quad + \int_{\tau_0}^{\tau_1} H_{\tau_0,t}\, c_t \Big(\mathbb{E}\big[\|e^M_t\|_2^2 \,\big|\, Y = y\big]\Big)^{1/2} dt.\end{aligned} \tag{A.10}$$

Averaging and stopping. Take $\mathbb{E}_Y$ on both sides of (A.10). The kernel $H_{\tau_0,t} c_t$ is deterministic and the integrand nonnegative, so Tonelli's theorem exchanges $\mathbb{E}_Y$ and $\int dt$, and Jensen's inequality for the concave square root gives $\mathbb{E}_Y[(\mathbb{E}[\|e^M_t\|_2^2 \mid Y])^{1/2}] \le \sqrt{R_M(t)}$ and $\mathbb{E}_Y[(\mathbb{E}[\|U - X_{\tau_1}\|_2^2 \mid Y])^{1/2}] \le (\mathbb{E}\|U - X_{\tau_1}\|_2^2)^{1/2}$. Together with (A.7) and $c_t = \beta_t/t$,

$$\mathbb{E}_Y\big[W_2\big(q^{M,Y}_{\tau_0}, p^Y_{\tau_0}\big)\big] \le C_{\mathrm{init}} + \int_{\tau_0}^{\tau_1} H_{\tau_0,t}\, \frac{\beta_t}{t} \sqrt{R_M(t)}\, dt. \tag{A.11}$$

Finally, the triangle inequality $W_2(q^{M,y}_{\tau_0}, \pi_y) \le W_2(q^{M,y}_{\tau_0}, p^y_{\tau_0}) + W_2(p^y_{\tau_0}, \pi_y)$, averaged over Y and combined with (A.11), gives (A.5). For the bound on C_{stop}, the pair (X_0, X_{τ_0}) is a coupling of $(\pi_y, p^y_{\tau_0})$ given $Y = y$ with $X_0 - X_{\tau_0} = \tau_0(X_0 - X_1)$, so $W_2(p^y_{\tau_0}, \pi_y) \le \tau_0(\mathbb{E}[\|X_0 - X_1\|_2^2 \mid Y = y])^{1/2}$, and Jensen's inequality over Y gives $C_{\mathrm{stop}} \le \tau_0(\mathbb{E}\|X_0 - X_1\|_2^2)^{1/2} = \tau_0\sqrt{d + \mathbb{E}\|X_0\|_2^2}$, because X_1 is independent of X_0 with zero mean and $\mathbb{E}\|X_1\|_2^2 = d$. For $U = X_1$, $U - X_{\tau_1} = (1 - \tau_1)(X_1 - X_0)$, which gives the stated value of C_{init}. All terms are finite by Assumptions A.2 and A.3. □

Corollary 1 (Deterministic sampler with Gaussian initialization)**.** *Let $g \equiv 0$, $\tau_1 = 1$, and $\mu_{\tau_1} = \mathcal{N}(0, I)$ with $U = X_1$. Then $\beta_t = 1$, $H_{\tau_0,t} = \tau_0/t$, $C_{\mathrm{init}} = 0$, and*

$$\mathbb{E}_Y\big[W_2\big(q^{M,Y}_{\tau_0}, \pi_Y\big)\big] \le \tau_0\sqrt{d + \mathbb{E}\|X_0\|_2^2} + \tau_0 \int_{\tau_0}^{1} \frac{\sqrt{R_M(t)}}{t^2}\, dt. \tag{A.12}$$

Proof. For $g \equiv 0$, $a_t = c_t = 1/t$ and $\beta_t = 1$, so $H_{\tau_0,t} = \tau_0/t$ by (A.3); $\tau_1 = 1$ and $U = X_1$ give $C_{\text{init}} = 0$ in the restated proposition, and the bound on C_{stop} gives the first term. In this case, $Z_t = (1-t)X_0 + tX_1$ is the interpolant itself. □

Equation (A.12) is the form that applies to the experiments of Section 3 and Appendix B: the deterministic sampler used there is Eq. (4) with $g \equiv 0$, started from Gaussian noise at $\tau_1 = 1$, and $R_M(t)$ is measured along its trajectory.

Remark A.1 (Sources of slack)**.** Three steps of the proof discard information: the coupling in (A.7) is synchronous rather than optimal; Minkowski's inequality in (A.10) ignores cancellations of the vector error e_t^M over time; and the triangle inequality at the stopping time adds C_{stop} regardless of the sampler. These steps are the same for both priors, so the prior enters the bound only through $R_M(t)$; nevertheless, as stated in Remark 1, the comparison of Section 3 orders upper bounds and not the posterior errors themselves.

A.3 Comparison of the Bounds and Measurement of the Corrected-Estimate Error

Proof of Remark 1. Write $\mathcal{B}_M$ for the right-hand side of (5). The constants are shared, so

$$\mathcal{B}_{\text{IT2I}} - \mathcal{B}_{\text{T2I}} = \int_{\tau_0}^{\tau_1} H_{\tau_0,t}\, \frac{\beta_t}{t} \Big[\sqrt{R_{\text{IT2I}}(t)} - \sqrt{R_{\text{T2I}}(t)} \Big] dt, \tag{A.13}$$

where both integrals are finite by Assumption A.3. The weight $H_{\tau_0,t}\beta_t/t$ is positive and the square root is strictly increasing. Hence, if $R_{\text{IT2I}}(t) \le R_{\text{T2I}}(t)$ for almost every t with strict inequality on a set of positive measure, the integrand is nonpositive almost everywhere and strictly negative on a set of positive measure, and $\mathcal{B}_{\text{IT2I}} < \mathcal{B}_{\text{T2I}}$. In general, the sign of (A.13) is that of the weighted difference of the root errors. □

Measurement along sampler trajectories. Let $t_0 = 1 > t_1 > \cdots > t_J = \tau_0 > 0$ be the time grid of the sampler, and let $x_{t_j}^{M,i,k}$ be the state of sampler M for measurement y_i ($i = 1, \ldots, n$) and initial noise k ($k = 1, \ldots, K$). The corrected-estimate error is estimated by

$$\widehat{R}_M(t_j) = \frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} \big\| \widetilde{d}_M\big(x_{t_j}^{M,i,k}, t_j; y_i\big) - x_{0,i} \big\|_2^2, \tag{A.14}$$

where the logged corrected estimate is the one that drives the sampler, as required by (A.2). Both priors use the same measurements, prompts, initial noises, schedule, and data-consistency rule, so the comparison is paired. Figures 4 and A4 report the relative difference $\Delta R_{\text{rel}}(t_j) = [\widehat{R}_{\text{IT2I}}(t_j) - \widehat{R}_{\text{T2I}}(t_j)]/\widehat{R}_{\text{T2I}}(t_j)$, whose sign at t_j is the condition of Remark 1 on the grid; the averages in (A.14) run over the runs of one group. Confidence bands are obtained by a paired image-cluster bootstrap that resamples images together with all their noises and both priors and recomputes $\Delta R_{\text{rel}}(t_j)$. Two caveats apply. The bound concerns the continuous-time sampler, so a statement about a finite-step solver would additionally require a discretization error term, and the sign of $\Delta R_{\text{rel}}(t_j)$ on a finite grid supports, but does not prove, the almost-everywhere condition of Remark 1. Moreover, $R_M(t)$ is measured in the space in which the backbone generates; pixel-space or perceptual errors computed after a nonlinear decoder are not substitutes.

B Measurement-Conditioned Generation in Additional IT2I Backbones

Section 3.1 characterizes native IT2I generation from a degraded reference with FLUX.2 and FLUX.1 Kontext. This appendix repeats its analyses on LongCat-Image-Edit (Meituan LongCat Team et al., 2025), FireRed-Image-Edit (Super Intelligence Team, 2026), and Mage-Flow-Edit-Base (Zhang et al., 2026): early structure preservation, text-dominated runs, and the relation of the two groups to the corrected-estimate error.

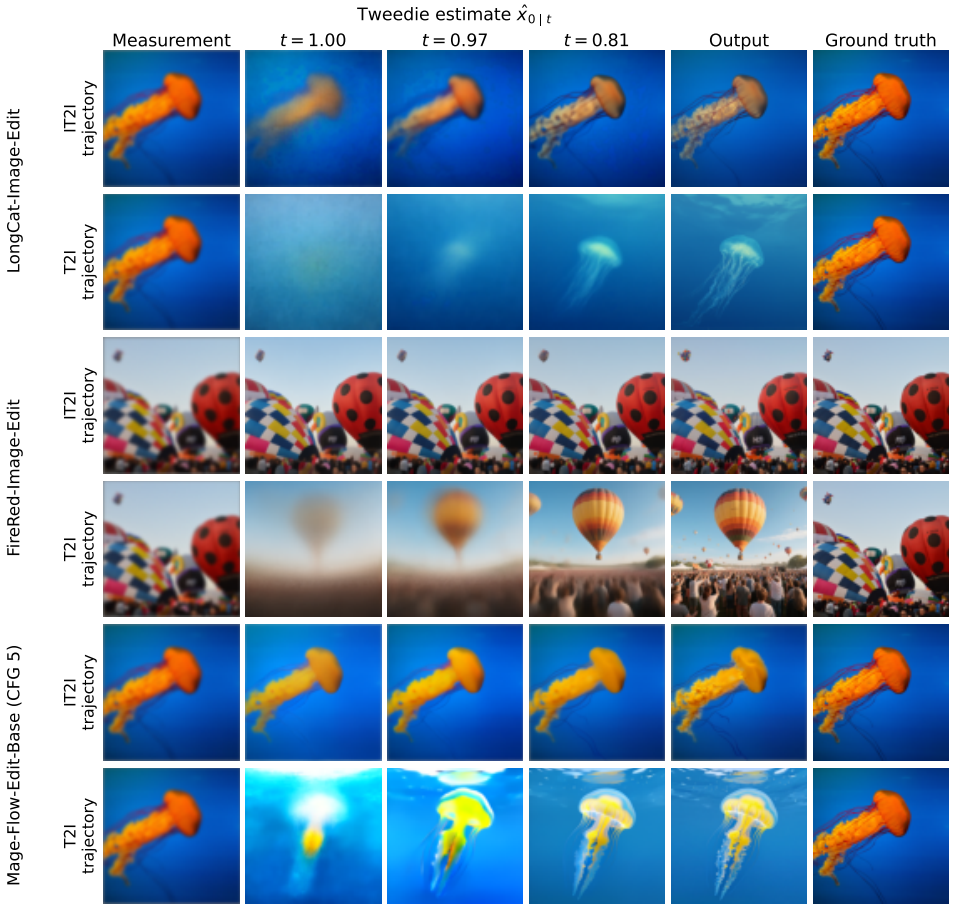

Figure A1: **Early structure preservation in additional IT2I backbones.** Native generation from a Gaussian-blurred measurement used as the reference ($c_I = y$) with LongCat-Image-Edit, FireRed-Image-Edit, and Mage-Flow-Edit-Base. For each backbone, the top row follows the IT2I trajectory and the bottom row the T2I trajectory with the same prompt and initial noise: Tweedie's estimates $\hat{x}_{0|t}$ at $t = 1.00$, 0.97, and 0.81 and the output, next to the measurement and the ground truth. As in Figure 2, the reference fixes the observed layout from the first step and later steps refine its details, whereas without the reference the first estimate does not carry the observed layout and later steps form a different scene.

Protocol. As in Section 3.1, we use Gaussian deblurring ($\sigma = 5$) with the measurement as the reference ($c_I = y$) on the first 100 DIV2K images, with five initial noises per image for LongCat-Image-Edit and FireRed-Image-Edit and three for Mage-Flow-Edit-Base. Each IT2I–T2I pair shares the prompt and the initial noise, and runs are labeled reference- or text-dominated by the $1\,\mathrm{dB}$ rule of Section 3.1. All three models are conditioned on the same per-image text prompt as in Section 3.1; LongCat-Image-Edit and FireRed-Image-Edit are sampled with a text guidance scale of 1.0 (no classifier-free guidance), and Mage-Flow-Edit-Base with its default scale of 5.0. The corrected-estimate error is measured with the deterministic sampler and data-consistency updates of Section 3.

Early structure preservation. Figure A1 shows the counterpart of Figure 2 for the three backbones. With the reference, each model preserves the observed layout and synthesizes plausible detail, whereas its T2I counterpart, with the same prompt and initial noise, produces a different scene. Along the IT2I trajectory, Tweedie's estimate shows the observed layout already at $t = 1$, and later estimates refine details within it; along the T2I trajectory with the same initial noise, the estimate at $t = 1$ does not carry the observed layout, and later estimates form a different scene. The reference thus supplies scene-specific structure early in all three models, as it does in FLUX.2.

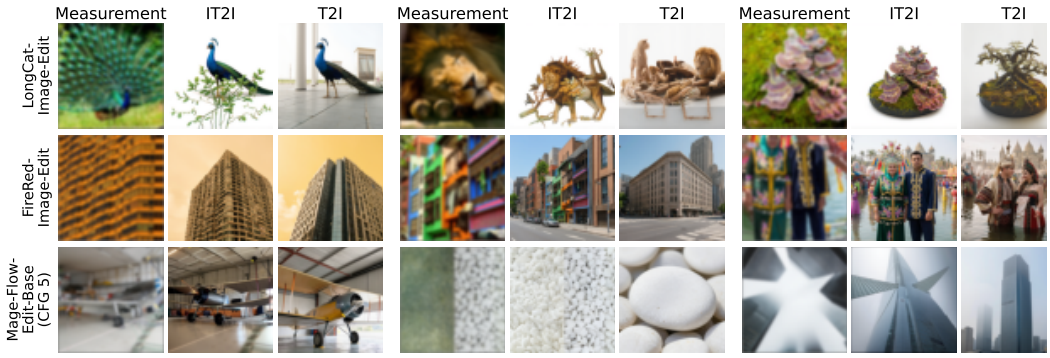

Figure A2: **Text-dominated runs in additional IT2I backbones.** Native generation with LongCat-Image-Edit, FireRed-Image-Edit, and Mage-Flow-Edit-Base (rows) for three runs per backbone labeled text-dominated, each shown as measurement, IT2I output, and T2I output: the IT2I output departs from the observed layout and resembles its T2I counterpart, which shares the prompt and the initial noise. Examples are chosen among the text-dominated runs whose IT2I output differs most from the measurement.

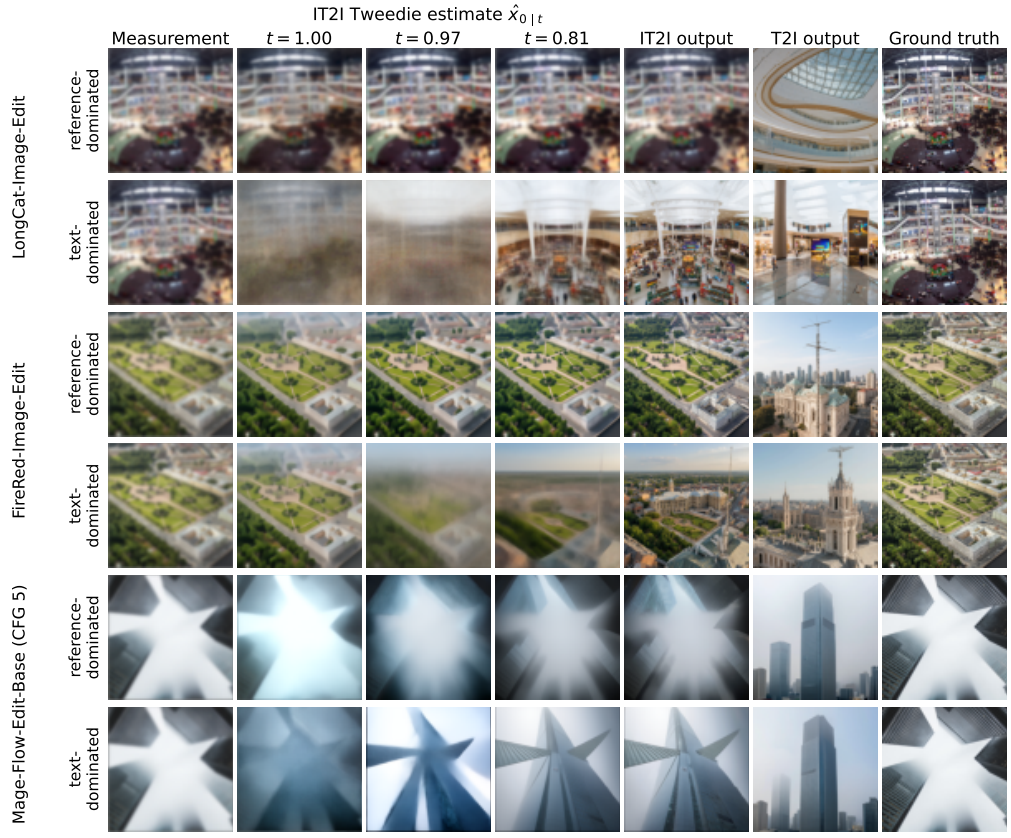

Figure A3: **The initial noise decides whether the reference is used.** For one measurement per backbone, native IT2I generation with the same prompt from two initial noises, one labeled reference-dominated (top row of each pair) and one labeled text-dominated (bottom row): the measurement, Tweedie's estimates $\hat{x}_{0|t}$ along the IT2I trajectory at $t = 1.00$, 0.97, and 0.81, the IT2I output, the T2I output with the same initial noise, and the ground truth. Along the reference-dominated trajectory the first estimate already carries the observed layout; along the text-dominated trajectory the early estimates do not, and the output converges to a scene that resembles the T2I output.

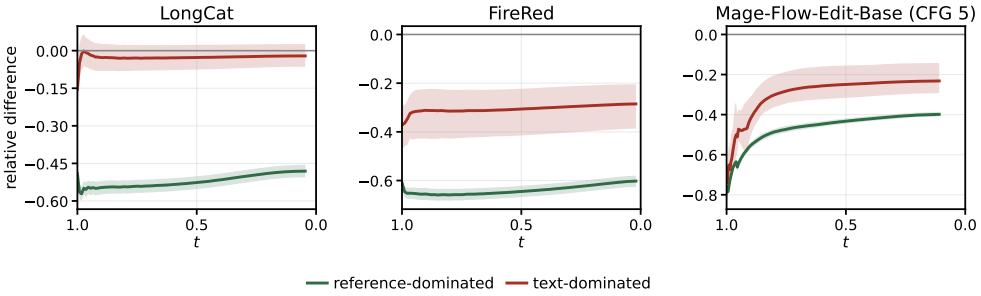

Figure A4: **Relative difference $\Delta R_{\mathrm{rel}}(t)$ of the corrected-estimate error** in reference- and text-dominated runs of LongCat-Image-Edit, FireRed-Image-Edit, and Mage-Flow-Edit-Base, with the deterministic sampler and data-consistency updates of Section 3. Negative values favor IT2I; the reduction is larger in reference-dominated runs for all three backbones.

Reduced reference influence. Text-dominated runs occur in every backbone. Figure A2 shows three per backbone: with the same prompt and initial noise, the IT2I output departs from the observed layout and resembles the T2I output, as in Figure 3(a). Figure A3 isolates the role of the initial noise. For the same measurement and prompt, one initial noise yields a reference-dominated run and another a text-dominated run; along the reference-dominated trajectory, Tweedie's estimate carries the observed layout from $t = 1$, whereas along the text-dominated trajectory the early estimates do not, and the output converges to a scene that resembles the T2I output. The departure is visible at $t = 1$ for LongCat-Image-Edit and within the first few steps for FireRed-Image-Edit and Mage-Flow-Edit-Base. Whether the reference is used is therefore settled in the initial steps in these backbones as well, which is what the readouts of Section 3.1 quantify for FLUX.2 and FLUX.1 Kontext.

Relation to posterior sampling. Figure A4 reports $\Delta R_{\mathrm{rel}}(t)$ for the same runs, grouped by label. In reference-dominated runs, image conditioning lowers the corrected-estimate error from the first step and throughout sampling in all three backbones, and the reduction is smaller in text-dominated runs. The three backbones thus show the same tendency as FLUX.2 and FLUX.1 Kontext in Figure 4.

C Further Details on the RefReg Algorithm

This appendix specifies the operations in Algorithm 1. Sampling proceeds from $t = 1$ to $t = 0$ in latent space. Pixel-space images lie in $\mathcal{C} = [-1, 1]^{C\times H\times W}$. Parameter values are provided in Appendix D.2.

C.1 Measurement-Derived Reference Construction

Initialization. We construct c_I once before sampling. The initial image r^0 is the measurement for deblurring and HDR, and its bicubic upsampling for super-resolution. For inpainting, we place the observed pixels on the reconstruction grid, set missing pixels to zero, and rescale the image to $[-1, 1]$ using its minimum and maximum. Missing regions are then filled inward from their boundaries by averaging available 3×3 neighbors. All initial images are clipped to $\mathcal{C}$.

Regularization and measurement updates. We instantiate $\mathcal{R}$ as smoothed isotropic total variation (TV) (Rudin et al., 1992):

$$\mathcal{R}(r) = \sum_{c,i,j} \sqrt{(D_h r)^2_{c,i,j} + (D_v r)^2_{c,i,j} + \delta^2}, \tag{C.15}$$

where D_h and D_v denote horizontal and vertical forward differences, and $\delta > 0$ ensures differentiability. At iteration k, we take a TV-gradient descent step, normalized by the gradient's ℓ_∞ norm and scaled by α, to obtain u^k. An operator-specific measurement update then produces r^{k+1} using

the discrepancy $f_y(r) = \frac{1}{2}\|\mathcal{A}(r) - y\|_2^2$. Both updates are followed by clipping to $\mathcal{C}$. We retain the iterate with the smallest measurement residual as the reference c_I.

Operator-specific measurement updates. For a linear operator $\mathcal{A}(r) = Ar$, the proximal measurement update has the closed form

$$\mathrm{prox}_{\eta^{-1} f_y}(u) = u + A^\top (AA^\top + \eta I)^{-1}(y - Au). \tag{C.16}$$

For HDR, where $\mathcal{A}(u) = \mathrm{clip}(su, -1, 1)$, we instead use $u + \frac{s}{s^2+\eta}(y - \mathcal{A}(u))$, based on the gain of the unsaturated branch. For nonlinear deblurring, the measurement update consists of two gradient-descent steps on f_y, starting from u^k. Each step has length $\gamma\|w\|_2$ at the current iterate w and is followed by clipping to $\mathcal{C}$.

C.2 Controlling Reference Influence During Sampling

Reference image guidance scheduling. For the primary FLUX.2 implementation, the text and image guidance scales in Eq. equation 7 decrease linearly with the sampling-step index. Their endpoints are specified in Appendix D.2.2. The null text condition is the encoded empty prompt, while the null image condition removes the reference-image tokens.

Early generation-to-text attention control. Let $\mathcal{G}$ index generation queries and $\mathcal{S}$ index selected prompt-content keys, excluding the fixed quality suffix. In the text-conditioned forward pass, we add the following bias to the joint self-attention logits:

$$(B_i)_{qk} = \begin{cases} -b, & t_i \geq t_w, \quad q \in \mathcal{G}, \quad k \in \mathcal{S}, \\ 0, & \text{otherwise}, \end{cases} \tag{C.17}$$

where q and k index queries and keys, respectively, and $b > 0$ controls the suppression strength. This bias suppresses direct access to the selected prompt content during the early window without masking reference-image or quality-suffix keys. Auxiliary text-embedding replacement and text-to-text attention control limit indirect transmission of prompt content through other text tokens. Token selection and application settings are specified in Appendix D.2.2.

C.3 Overall Sampling Procedure With Data Consistency Guidance

Measurement discrepancy. The data-consistency objective in Section 4.3 uses a task-specific measurement discrepancy. For linear operators, we use the regularized back-projection loss (Garber & Tirer, 2024)

$$\mathcal{L}_{\mathrm{BP}}(p, y) = \left\| A_\xi^+ (p - y) \right\|_2, \qquad A_\xi^+ = A^\top (AA^\top + \xi I)^{-1}, \tag{C.18}$$

where p denotes the predicted measurement. For HDR, we replace A_ξ^+ with the scalar gain $s/(s^2 + \xi)$. For nonlinear deblurring, we use the measurement residual $\mathcal{L}_{\mathrm{res}}(p, y) = \|p - y\|_2$. Both discrepancies use the unsquared norm.

Correction and continuation. Starting from the predicted clean latent $x^0 = \widehat{x}_{0|t_i}$, we take J gradient steps:

$$x^{j+1} = x^j - \beta_i \nabla \ell_y(x^j), \qquad j = 0, \ldots, J-1. \tag{C.19}$$

The corrected estimate $\widetilde{x}_{0|t_i} = x^J$ replaces the clean endpoint in Algorithm 1. Sampling then continues by re-interpolating this estimate with the retained noise endpoint, without additional noise. The step-size schedule, update counts are specified in Appendix D.2.3.

D Experimental Details

D.1 Datasets, Operators, and Evaluation Protocol

The main comparisons use 800 DIV2K training images and 1,000 FFHQ images. Images are resized along the shorter side and center-cropped to 768×768. Measurements contain Gaussian noise with

Table A1: Forward operators and degradation settings used in the reconstruction experiments.

Task	Operator specification
SR $\times 8$, $\times 16$	Downsampling by the stated factor using an antialiased bicubic filter with reflective padding.
Gaussian deblurring	Standard deviation $\sigma = 5$ on a 60×60 kernel, with zero padding.
Motion deblurring	A 61×61 kernel generated by the random-path generator of DPS (intensity 0.5, one kernel per image), applied using circular convolution.
Box inpainting	A missing box of size 384×384 at the image center.
Random inpainting	90% missing pixels using a fixed pixel-wise uniform random mask shared across images and channels.
Nonlinear deblurring	The learned blur model of Tran et al. (2021), as used in Zhang et al. (2025), with the `GOPRO_wVAE` checkpoint and a fixed kernel code (seed 0, scale 1.2).
HDR reconstruction	Saturation with gain 2: $\mathcal{A}(x) = \mathrm{clip}(2x, -1, 1)$.

$\sigma_y = 0.03$ on the $[-1, 1]$ pixel scale. Text prompts follow Appendix D.2, and Table A1 lists the forward operators. The component ablation study and the sampling-step comparison each use 100 DIV2K images.

We measure reconstruction fidelity with PSNR and SSIM, and perceptual quality with LPIPS and patch-based FID. FID uses InceptionV3 features from nine non-overlapping 256×256 patches per image, aggregated separately for each dataset and task.

D.2 Settings for RefReg

The settings below describe our primary FLUX.2 implementation on DIV2K. FFHQ uses the same measurement-condition construction and conditioning controls, with task-specific differences listed in Table A4.

Backbone and sampling settings. We use FLUX.2-klein-base-4B in bfloat16 at a resolution of 768×768, with $N = 50$ NFEs. The measurement noise is $\sigma_y = 0.03$ on the $[-1, 1]$ pixel scale. The base autoencoder's image encoder encodes c_I, and the decoder is used for all decoding during sampling and for the final output.

Text-conditioning settings. We use text features from layers 9, 18, and 27 of the text encoder. All prompts, including the empty prompt for null conditioning, use the model's chat template. For DIV2K, each prompt consists of an image-specific DAPE caption followed by the fixed quality suffix ", clean, high-resolution, 8k". FFHQ uses the shared prompt "a high quality photo of a face, clean, high-resolution, 8k".

D.2.1 Measurement-derived reference construction

Table A2 lists the parameters of the reference construction in Appendix C.1. We evaluate Eq. equation C.16 using the singular value decomposition of A for Gaussian deblurring, super-resolution, and inpainting, and the FFT for motion deblurring with circular boundaries.

Table A2: Parameters for constructing the measurement condition c_I (Appendix C.1).

Parameter	Value
Iterations K	60 (10 for Gaussian deblurring and super-resolution)
TV step bound α per pixel	0.005
TV smoothing parameter δ	10^{-3}
Proximal parameter η	0.05
Gradient step γ (nonlinear blur only)	0.01

D.2.2 Controlling Reference Influence during Sampling

Reference image guidance scheduling. Text guidance decreases from 4.0 to 1.0 for all tasks, while Table A3 lists the task-specific image guidance endpoints. Both schedules follow Appendix C.2.

Table A3: Image guidance endpoints $s_I^{\text{start}} \to s_I^{\text{end}}$ across tasks for FLUX.2 on DIV2K.

	SR×8	SR×16	Gauss.	Motion	Nonlin.	HDR	Inp. 90%	Box
s_I	$2 \to 0$	$2 \to 0$	$3 \to 0$	$4 \to 1$	$3 \to 0$	$2 \to 1$	$3.5 \to 0$	$4 \to 0$

Early generation-to-text attention control. The early window covers the first three sampling steps when $N = 50$, corresponding to $t_w = 0.99$. During this window, generation queries receive an attention bias of -30 toward prompt-content keys $\mathcal{S}$ in every attention layer and head. The selected keys exclude the quality suffix, chat-template tokens, and padding. For box inpainting, an additional bias of $\log 10^{-6}$ suppresses attention from text and generation queries to reference-image keys in the missing region.

Auxiliary text controls. Before sampling, we replace quality-suffix, closing-template, and padding embeddings with aligned embeddings obtained by encoding the quality suffix alone. These replacements remain throughout sampling. During the early window, the bias of -30 also suppresses attention from template-prefix, quality-suffix, and closing-template queries to $\mathcal{S}$. All attention biases are applied only in the text-conditioned pass, leaving empty-prompt passes unchanged.

D.2.3 Overall Sampling Procedure with Data Consistency Guidance

The step size β_i in Eq. equation C.19 combines an exponentially varying correction strength ρ_i with the factor $f + (1 - f)t_i^2$, divided by the number of gradient updates J. We interpolate ρ_i from 30 to 10 using an exponential schedule with shape parameter 6, and set $f = 1.5$ and $\xi = \sigma_y^2$. The task-specific discrepancies are defined in Appendix C.3. For DIV2K, we use $J = 12$ updates per sampling step, except for motion deblurring ($J = 10$) and nonlinear deblurring ($J = 30$).

After sampling, we clip the reconstructed image to $\mathcal{C}$ and apply the linear or HDR measurement update described in Appendix C.1 with $\eta = \xi' = 0.05$, followed by clipping to $\mathcal{C}$.

Table A4 lists the FFHQ-specific changes. All other settings follow DIV2K.

Table A4: FFHQ settings that differ from DIV2K for FLUX.2. Dashes indicate unchanged settings.

Task	Image guidance s_I	Data-consistency correction
SR×8, SR×16, Box	$3 \to 0$	—
Motion	$3 \to 1$	$\rho = 70 \to 23.3$
Inpainting 90%	$4 \to 0$	$\rho = 60 \to 20$
HDR	—	$J = 20$

D.3 Baseline Adaptations

For the main comparisons, we adapt six flow-based inverse solvers to the FLUX.2 backbone and resolution in Appendix D.2, using 50 sampling steps and the same measurements and text prompts. Unless stated otherwise, text and image guidance scales are both 2.0. Method-specific updates and settings are described below. All experiments are conducted on a single NVIDIA A100 GPU. Results on additional backbones are provided in Appendix E.7.

- **FlowChef** (Patel et al., 2025). At each sampling step, we optimize the current noisy latent with 15 Adam iterations, holding the velocity fixed and minimizing the measurement MSE of the decoded clean estimate. Learning rates are 0.02 for SR×8, 0.0025 for SR×16, 0.05 for Gaussian, motion, and nonlinear deblurring, 0.1 for box inpainting, 0.4 for random inpainting, and 0.07 for HDR. For HDR, text guidance is 4.0 and image guidance decreases linearly from 1.0 to 0.
- **FlowDPS** (Kim et al., 2025a). We refine the predicted clean latent with three gradient-descent updates per sampling step at step scale 15, followed by time-dependent blend-

ing and stochastic renoising. The objective is the ℓ_2 norm of the pseudoinverse residual for super-resolution ($\eta = 10^{-8}$), the adjoint residual for other linear tasks, and the measurement-space residual for HDR and nonlinear deblurring.

- **Flower** (Pourya et al., 2026). We use 50 conjugate-gradient iterations for pixel-space proximal refinement of the decoded clean estimate, followed by 15 Adam iterations to match the latent to the refined image. Latent-matching learning rates are 0.4 for both SR tasks, 0.2 for Gaussian deblurring, 0.8 for motion deblurring, 0.1 for box inpainting and nonlinear deblurring, 0.025 for random inpainting, and 0.05 for HDR. Nonlinear deblurring uses one Gauss–Newton step with 50 conjugate-gradient iterations. HDR uses the exact pointwise proximal map and text guidance 4.0. We return the final refined image.
- **FLAIR** (Erbach et al., 2025). The data-consistency iteration budget per sampling step is 15 for SR$\times$8, motion deblurring, and inpainting, and 20 otherwise. The data-term learning rate is $17694.72/m$, where m is the measurement dimension. Regularizer learning rates are 1.25 for SR$\times$16, 0.25 for motion deblurring and inpainting, and 1.0 otherwise. Text guidance is 4.5 for Gaussian deblurring, 5.0 for SR$\times$16, and 3.0 for HDR and nonlinear deblurring. Image guidance is 1.0 except for motion deblurring and inpainting, which use 2.0.
- **MPC** (Webber et al., 2026). We use MPC-RHC with a single-step planning horizon. At each sampling step, we hold the model velocity fixed and optimize an additive clean-latent correction against the squared measurement residual. The correction is then used to adjust the sampling velocity. It is initialized to zero and optimized with a fresh Adam optimizer for 15 iterations at learning rate 0.1. Unlike the original formulation, we omit the control-energy penalty.
- **TriPS** (Bang et al., 2026). We adapt TriPS-T to FLUX.2 with six gradient-descent updates per sampling step and the released implementation's time-dependent weighting of back-projection and least-squares losses. Data-consistency guidance decreases linearly from 30 to 20. Text CFG increases from 1 to 6 and stochasticity decreases from 1 to 0, following the released exponential templates.

D.4 Partially Blind Inverse Problems

We evaluate partially blind Gaussian deblurring and HDR reconstruction on 100 DIV2K training images (Table 3), using the preprocessing, measurement noise, and image-quality metrics in Appendix D.1. Gaussian deblurring estimates a nonnegative, unit-sum 61×61 kernel for circular convolution. The true kernel is Gaussian with standard deviation 5. HDR estimates $s \in [0.5, 8]$ in $\mathcal{A}_s(u) = P(sP(u))$, where $P(u) = \mathrm{clip}(u, -1, 1)$ and $s^\star = 2$.

All methods share the frozen FLUX.2 backbone, measurements, and initial noise, and initialize HDR exposure at $s = 1$. For Gaussian deblurring, Ours and GDP start from a centered delta kernel, while DMPlug uses normally distributed logits with a softmax parameterization. Ours uses image-specific captions and a measurement reference constructed under the initial operator and then held fixed. GDP and DMPlug use a fixed raw-measurement reference and the prompt "a high quality photo". Qualitative comparisons are provided in Figures A19 and A20.

D.4.1 Extension of RefReg

We retain the conditioning and sampling framework of Algorithm 1 and the conditioning settings in Appendix D.2. Task-specific operator estimation and data-consistency updates start from the predicted clean latent $q_i^{(0)} = \widehat{x}_{0|t_i}$. We write u for the image estimate associated with the clean latent q, and $u_i^{(0)}$ for the estimate associated with $q_i^{(0)}$. After correction, sampling continues by re-interpolation with the retained noise endpoint.

Gaussian deblurring. At sampling steps $i = 16, 21, \ldots, 46$, we fix the clipped image prediction $\bar{u}_i = P(u_i^{(0)})$ and update the kernel before BP correction. At each pyramid level, the update solves

$$h^{(j+1)} = \arg\min_h \ \frac{1}{2} \sum_c \|h * \bar{u}_{i,c} - y_c\|_2^2 + \frac{\gamma_k}{2} \|h - Ek^{(j)}\|_2^2, \tag{D.20}$$

where h is an image-grid kernel and E embeds the current kernel on that grid. The unconstrained solution is computed by FFT, cropped to the prescribed support, clipped to nonnegative values, and normalized. Each scheduled update starts from the previous kernel and uses average-pooling factors $4, 2, 1$, supports $15, 31, 61$, ten iterations per level, and $\gamma_k = 10^6$. Kernels are transferred between levels by bilinear interpolation and renormalization. The estimated kernel is then held fixed for the BP correction in Appendix C.3

HDR reconstruction. We jointly refine the clean latent and exposure scale while penalizing deviation from the fixed image prediction $a_i = u_i^{(0)}$, motivated by the HQS image–operator subproblem of CaMB-Diff (Liu et al., 2026):

$$\mathcal{L}_i^{(j)}(u, s) = b_j \sqrt{\|\mathcal{A}_s(u) - y\|_2^2 + \lambda_{\mathrm{p}} \|u - a_i\|_2^2}, \qquad b_j = \frac{s^{(j)}}{(s^{(j)})^2 + \xi}. \tag{D.21}$$

We set $\lambda_{\mathrm{p}} = 0.5$ and $\xi = 0.0009$. The BP factor b_j is recomputed but held constant during each gradient evaluation. The objective is evaluated in image space, with gradients propagated to the clean latent q. At each sampling step, we reset Adam and take fifteen updates of q, updating s jointly only at $i = 9, 11, \ldots, 49$. Among these iterates and the initial pair, we retain the pair minimizing the squared-residual sum inside the square root, without b_j.

D.4.2 Baseline Adaptations

Guidance scales and learning rates are listed in Table A5.

- **GDP** (Fei et al., 2023). We convert the FLUX velocity to a VP noise prediction for 50-step DDPM sampling. At each step, measurement-MSE gradients at the clean estimate update the image proposal and operator from the same incoming pair, with an image-gradient multiplier of 1500. Gradients do not propagate through the velocity model. Kernel updates use gradient descent with projection onto the probability simplex.
- **DMPlug** (Wang et al., 2024). We retain optimization through the full generator, replacing DDIM with native Euler flow sampling. Adam jointly optimizes the initial noise and operator to minimize measurement MSE. Gaussian kernels use softmax logits. After 200 Adam updates, we regenerate the output from the final optimized noise using the three-step generator.

All Adam optimizers use $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$.

Table A5: Guidance and learning-rate settings for partially blind reconstruction (Gaussian deblurring / HDR reconstruction).

Setting	Ours	GDP	DMPlug
Text guidance	$4 \to 1$	2	2 / 1
Image guidance	$3 \to 0$ / $2 \to 1$	1	1 / 0.5
Latent/noise Adam LR	— / 0.05	—	0.001 / 0.003
Operator LR	— / 0.01	1 / 3	0.1

E Additional Quantitative Results and Ablations

E.1 Results across All Tasks

Table A6 reports the complete results for all eight inverse problems on DIV2K and FFHQ, extending the selected tasks in Table 1. These comprise super-resolution at $\times 8$ and $\times 16$, Gaussian and motion deblurring, box and random inpainting, nonlinear deblurring, and HDR reconstruction. All results follow the evaluation protocol in Section 5.1; the entries shared with Table 1 are reproduced for completeness.

Figure A5 presents qualitative comparisons across the same eight tasks using DIV2K and FFHQ images different from those in Figure 5.

Table A6: **Complete quantitative comparison** across all eight inverse problems on 800 DIV2K and 1,000 FFHQ images using 50 NFEs and Gaussian measurement noise ($\sigma_y = 0.03$). Bold and underlined values indicate the best and second-best results, respectively.

Task	Method	DIV2K				FFHQ			
		PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Super-Resolution ×8	FlowChef	22.06	0.510	0.392	68.34	27.54	0.706	0.308	92.78
	FlowDPS	21.78	0.584	0.550	82.14	25.86	0.784	0.385	83.59
	Flower	20.01	0.516	0.467	69.39	23.09	0.745	0.334	63.51
	FLAIR	22.73	0.587	0.348	25.79	28.43	0.752	0.235	22.76
	MPC	22.31	0.555	0.406	37.09	28.46	0.765	0.327	55.07
	TriPS	22.29	0.575	0.467	42.40	28.79	0.794	0.300	39.23
	Ours	**23.01**	**0.600**	**0.266**	**15.36**	**29.85**	**0.806**	**0.194**	**13.56**
Super-Resolution ×16	FlowChef	**20.42**	0.489	0.496	84.79	25.59	0.717	0.371	82.30
	FlowDPS	19.87	**0.516**	0.664	187.00	23.55	0.741	0.479	176.50
	Flower	15.81	0.449	0.610	130.29	16.96	0.653	0.469	129.03
	FLAIR	19.52	0.456	0.466	48.37	24.29	0.633	0.337	61.61
	MPC	20.04	0.462	0.475	54.05	25.78	0.720	0.369	67.58
	TriPS	19.91	0.492	0.531	55.98	25.67	0.737	0.347	43.02
	Ours	20.27	0.492	**0.367**	**21.86**	**26.55**	**0.759**	**0.288**	**29.01**
Motion Deblurring	FlowChef	25.14	0.698	0.280	35.63	29.97	0.816	0.263	46.20
	FlowDPS	21.34	0.649	0.388	36.74	24.27	0.800	0.309	44.48
	Flower	25.79	0.704	0.271	33.74	30.35	0.777	0.229	50.48
	FLAIR	26.29	**0.746**	0.246	22.73	**31.63**	**0.853**	0.226	31.56
	MPC	25.67	0.713	0.192	15.64	30.95	0.823	0.152	19.20
	TriPS	24.89	0.673	0.228	**11.71**	29.90	0.794	0.156	**9.31**
	Ours	**27.61**	0.741	**0.154**	14.70	31.27	0.760	**0.137**	13.89
Gaussian Deblurring	FlowChef	21.54	0.538	0.467	97.68	27.16	0.757	0.354	86.35
	FlowDPS	19.09	0.512	0.650	130.86	22.28	0.737	0.454	88.49
	Flower	21.29	0.472	0.476	81.97	26.05	0.636	0.362	66.21
	FLAIR	19.80	0.487	0.392	25.14	24.69	0.673	0.255	16.43
	MPC	20.76	0.498	0.398	37.12	26.81	0.745	0.272	27.14
	TriPS	20.99	0.499	0.367	20.37	27.32	0.740	0.229	16.40
	Ours	**22.13**	**0.563**	**0.324**	**19.51**	**28.64**	**0.773**	**0.217**	**16.21**
Inpainting (Box)	FlowChef	18.53	**0.795**	0.210	18.86	19.85	**0.875**	0.189	34.60
	FlowDPS	18.23	0.697	0.287	16.37	21.81	0.835	0.185	11.57
	Flower	19.04	0.773	0.198	30.93	21.63	0.821	0.169	22.52
	FLAIR	18.64	0.766	0.226	18.74	20.91	0.855	0.187	37.52
	MPC	19.01	0.774	0.191	9.01	23.29	0.874	0.135	8.44
	TriPS	18.72	0.751	0.195	9.86	**23.67**	0.863	0.119	**5.62**
	Ours	**19.57**	0.762	**0.147**	**6.30**	23.65	0.817	**0.112**	6.07
Inpainting (Random)	FlowChef	23.96	0.709	0.234	43.76	30.68	0.852	0.212	57.15
	FlowDPS	22.29	0.639	0.314	61.14	28.55	0.807	0.216	83.06
	Flower	16.36	0.366	0.496	154.85	18.16	0.367	0.569	263.74
	FLAIR	23.78	0.695	0.272	88.05	29.73	0.814	0.226	111.14
	MPC	22.76	0.632	0.317	48.99	30.05	0.826	0.179	55.52
	TriPS	24.54	0.714	0.226	19.39	**31.99**	**0.872**	0.166	15.28
	Ours	**24.61**	**0.722**	**0.169**	**13.32**	31.70	0.854	**0.108**	**9.04**
Nonlinear Deblurring	FlowChef	**26.79**	**0.779**	**0.148**	15.64	**30.75**	**0.843**	0.183	27.50
	FlowDPS	21.51	0.641	0.387	41.37	24.55	0.803	0.283	32.85
	Flower	23.71	0.643	0.277	37.59	27.55	0.715	0.244	32.72
	FLAIR	23.57	0.656	0.208	**13.62**	28.44	0.778	0.154	12.31
	MPC	23.98	0.695	0.225	23.23	29.00	0.824	0.181	20.87
	TriPS	23.47	0.661	0.286	21.56	28.92	0.820	0.208	17.84
	Ours	25.20	0.724	0.186	14.44	29.67	0.805	**0.144**	**10.63**
High Dynamic Range	FlowChef	24.85	0.820	0.108	10.98	24.32	0.844	0.139	18.38
	FlowDPS	18.08	0.707	0.193	12.91	18.96	0.777	0.189	15.95
	Flower	21.43	0.802	0.112	8.87	20.89	0.816	0.131	13.91
	FLAIR	24.02	0.783	0.157	12.96	24.66	0.822	0.180	21.08
	MPC	24.22	0.813	0.109	8.79	23.73	0.846	0.112	11.96
	TriPS	23.87	0.750	0.191	9.86	25.12	0.827	0.149	7.72
	Ours	**26.92**	**0.847**	**0.068**	**4.93**	**28.89**	**0.861**	**0.066**	**4.32**



Figure A5: **Qualitative comparisons across all eight inverse problems** using additional DIV2K and FFHQ examples. All methods are compared on the same measurement for each example.

E.2 Effect of Reference Optimization

Table A7 evaluates reference optimization while retaining image conditioning in both variants. Applying the regularization and measurement-fidelity updates to the initial reference improves reconstruction PSNR and LPIPS on both tasks.

Table A7: **Effect of reference optimization** on 100 DIV2K images. Both variants retain image conditioning, IGS, EAB, and DC updates.

	Motion deblurring		Inpainting (90%)	
Variant	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
w/o optimization	25.373	0.316	24.620	0.168
w/ optimization	**27.623**	**0.158**	**24.670**	**0.165**

E.3 Qualitative Component Ablations

Figure A6 provides qualitative comparisons for the component ablations in Section 5.3. The variants correspond to Table 4 and retain data consistency updates.

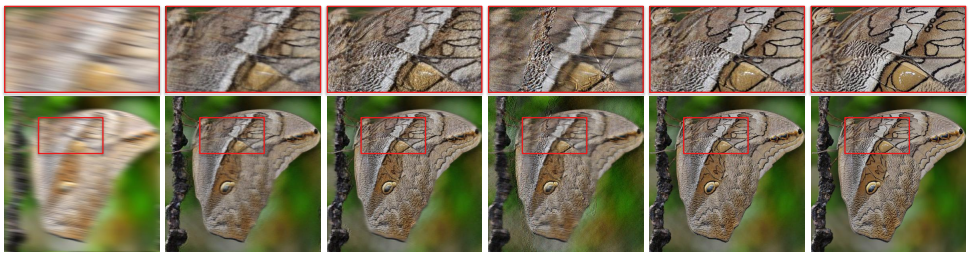

Figure A6: **Component ablations** of RefReg. Qualitative comparison of the full method with variants that remove image conditioning (IC), replace image guidance scheduling (IGS) with fixed unit guidance, or remove early-step attention blocking (EAB), corresponding to Table 4. All variants include data consistency updates.

E.4 Timing of Generation-to-Text Attention Blocking

We examine the timing of generation-to-text attention blocking (AB) on 100 DIV2K images using FLUX.2 Klein Base 4B with 50 NFEs. Our early-step AB (EAB), Middle-step AB, and Late-step AB apply the same attention bias of -30 to the same connections for five NFEs at the positions listed in Table A8. Measurements, prompts, initial noise, guidance schedules, and data consistency settings are fixed across variants. We keep the auxiliary text controls described in Appendix C.2 fixed across variants. The text-to-text bias is active at NFE indices 0–4. No AB disables only generation-to-text blocking.

Table A8 shows that our EAB achieves the best PSNR, SSIM, and LPIPS on both tasks. Middle and late blocking provide negligible gains over No AB. In Figure A7, these three variants produce similar structural deviations and artifacts, consistent with the text-dominated behavior described in Section 3.1. Our EAB reduces these artifacts and better preserves the observed structure. Only early blocking provides a consistent improvement, supporting EAB during scene formation.

Table A8: Effect of generation-to-text attention blocking (AB) timing on 100 DIV2K images. Best results are in bold.

		SR $\times 16$			Gaussian deblurring ($\sigma = 5$)		
AB window	NFE indices	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
No AB	–	20.54	0.5194	0.4528	22.44	0.5871	0.4066
EAB (ours)	0–4	**20.83**	**0.5397**	**0.4242**	**22.74**	**0.6056**	**0.3867**
Middle-step AB	23–27	20.55	0.5194	0.4529	22.45	0.5874	0.4066
Late-step AB	45–49	20.54	0.5192	0.4530	22.44	0.5870	0.4067

NFE indices are zero-based. Each active window spans five NFEs under the same nonuniform noise schedule. The main configuration applies EAB during the first three NFEs.

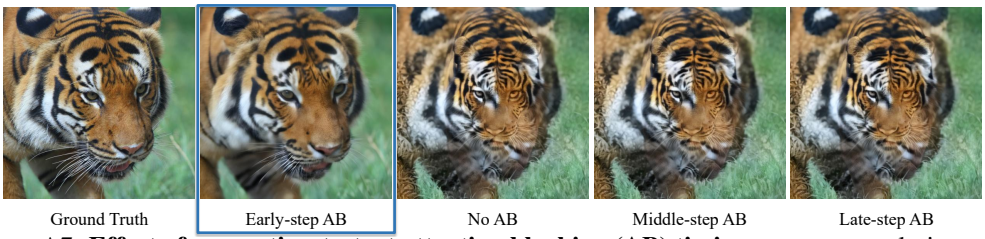

Figure A7: **Effect of generation-to-text attention blocking (AB) timing** on super-resolution $\times 16$. Early-step AB better preserves the observed structure and reduces artifacts, while middle-step and late-step AB produce results similar to No AB. The comparison uses the same measurement, text prompt, and initial noise.

Table A9: Reconstruction quality and computational cost at different NFEs on 100 DIV2K images. IC: image conditioning. Sampling time and model FLOPs are reported per image; T denotes 10^{12} operations.

		SR $\times 16$				Gaussian deblurring ($\sigma = 5$)			
NFEs	IC	PSNR↑	LPIPS↓	Time (s)↓	FLOPs (T)↓	PSNR↑	LPIPS↓	Time (s)↓	FLOPs (T)↓
50	✗	19.33	0.502	22.1	2,339	21.26	0.451	24.1	2,339
	✓	20.29	0.371	40.4	5,323	22.15	0.321	42.0	5,323
25	✗	18.74	0.502	16.1	1,352	20.63	0.452	18.0	1,352
	✓	20.11	0.363	25.4	2,835	21.92	0.304	27.3	2,835
15	✗	18.47	0.510	13.7	957	20.19	0.461	15.6	957
	✓	20.09	0.359	19.4	1,848	21.88	0.298	21.5	1,868
8	✗	18.54	0.532	12.0	681	19.92	0.490	13.9	681
	✓	20.02	0.383	15.2	1,177	21.71	0.309	17.2	1,177

E.5 Few-NFE Reconstruction and Computational Cost

Table 2 evaluates sensitivity to reducing sampling over $N \in \{50, 25, 15, 8\}$ on the same 100 DIV2K images. Both variants use FLUX.2 Klein Base 4B with shared measurements, text prompts, initial noise, and the sampling grid at each N. The variant without image conditioning also disables image guidance and early-window attention blocking; text guidance and data consistency remain active. We hold the total data-consistency budget fixed at 600 inner updates, using 12, 24, 40, and 75 updates per sampling step, respectively. The per-step correction parameters are scaled by $50/N$ to preserve the correction update scale. For the full method, the early window spans 3, 2, 2, and 1 steps, respectively. TV preprocessing of the measurement condition is disabled in this sweep.

Table A9 includes the 50-NFE reference results, while Table 2 reports PSNR and LPIPS with paired changes from each variant's own 50-NFE run in parentheses; small differences from subtracting the displayed means are due to rounding. Both variants are compared at the same number of NFEs under a fixed total data-consistency budget. Image conditioning reduces PSNR loss at every tested step count, and LPIPS improves at 25 and 15 NFEs. At 8 NFEs, LPIPS remains below its 50-NFE value for Gaussian deblurring but increases for SR $\times 16$. These results assess robustness to shorter trajectories when image conditioning and its associated controls are present; they do not directly isolate early structure preservation as the causal mechanism.

E.6 Evaluation on a Distilled Backbone

We evaluate RefReg on the distilled FLUX.2-klein-4B checkpoint with four NFEs. Table A10 compares native IT2I sampling, data-consistency correction alone, and RefReg on 100 DIV2K images. RefReg improves both PSNR and LPIPS over data-consistency correction alone, demonstrating its effectiveness on a distilled backbone. With only four NFEs, it also remains competitive with the 50-NFE base-backbone results, achieving similar or better LPIPS with a 0.56–0.69 dB decrease in PSNR.

Table A10: Reconstruction quality on 100 DIV2K images. Four-step variants use distilled FLUX.2-klein-4B; the 50-NFE reference uses FLUX.2-klein-base-4B.

		SR $\times 16$		Gaussian deblurring ($\sigma = 5$)	
Variant	Steps	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Native IT2I	4	10.28	0.574	10.25	0.564
Native + correction	4	18.57	0.441	20.19	0.390
RefReg	4	19.73	0.372	21.46	0.299
RefReg (base)	50	20.29	0.371	22.15	0.321

E.7 Evaluation on Additional IT2I Backbones

To evaluate the applicability of RefReg beyond FLUX.2, we extend the framework to FLUX.1 Kontext, FireRed-Image-Edit, and LongCat-Image-Edit. We compare against FlowChef, FlowDPS, TriPS-T, and FLAIR on 100 DIV2K images for super-resolution $\times 8$, Gaussian deblurring ($\sigma = 5$), and motion deblurring. Within each backbone, all methods share the measurements, prompts, and measurement-derived image conditions, with settings selected for each backbone and task.

As shown in Table A11, RefReg achieves the lowest LPIPS across all nine backbone–task combinations while maintaining competitive reconstruction fidelity. Relative to the best baseline in each setting, LPIPS decreases by 9.6–42.7%. These results demonstrate that the perceptual gains of our framework extend beyond FLUX.2 and remain consistent across the three additional IT2I backbones. Figures A8, A9, and A10 present the corresponding qualitative comparisons.

Table A11: Quantitative comparison on additional IT2I backbones using 100 DIV2K images at 768×768, 50 NFEs, and Gaussian measurement noise ($\sigma_y = 0.03$). **Bold** and underlined values indicate the best and second-best results within each backbone and task.

	Super-resolution $\times 8$			Gaussian deblurring ($\sigma = 5$)			Motion deblurring		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
FLUX.1 Kontext									
FlowChef	22.53	0.541	0.533	20.82	0.462	0.555	25.89	0.722	0.282
FlowDPS	22.84	0.589	0.555	21.21	**0.544**	0.530	26.42	0.751	0.218
TriPS-T	22.87	0.595	0.533	20.58	0.498	0.592	25.52	0.719	0.336
FLAIR	22.30	0.530	0.503	20.66	0.490	0.569	25.54	0.717	0.357
Ours	**23.14**	**0.604**	**0.340**	**21.29**	0.523	**0.346**	**27.38**	**0.782**	**0.125**
FireRed-Image-Edit									
FlowChef	22.99	0.591	0.480	**21.80**	0.556	0.515	24.90	0.688	0.337
FlowDPS	**23.04**	**0.612**	0.319	21.41	**0.557**	0.361	25.45	0.735	0.202
TriPS-T	20.73	0.560	0.304	21.61	0.556	0.292	26.05	0.745	0.186
FLAIR	21.82	0.554	0.233	19.59	0.466	0.305	26.25	0.767	0.164
Ours	22.05	0.562	**0.201**	21.48	0.541	**0.264**	**27.24**	**0.772**	**0.122**
LongCat-Image-Edit									
FlowChef	23.10	0.593	0.534	21.55	0.515	0.523	26.46	0.733	0.250
FlowDPS	22.86	0.580	0.564	20.35	0.516	0.477	22.98	0.669	0.294
TriPS-T	22.83	0.577	0.538	**22.12**	**0.572**	0.437	**27.65**	**0.780**	0.235
FLAIR	21.80	0.474	0.525	20.42	0.465	0.500	25.94	0.745	0.325
Ours	**23.28**	**0.608**	**0.369**	21.81	0.541	**0.319**	26.85	0.744	**0.149**

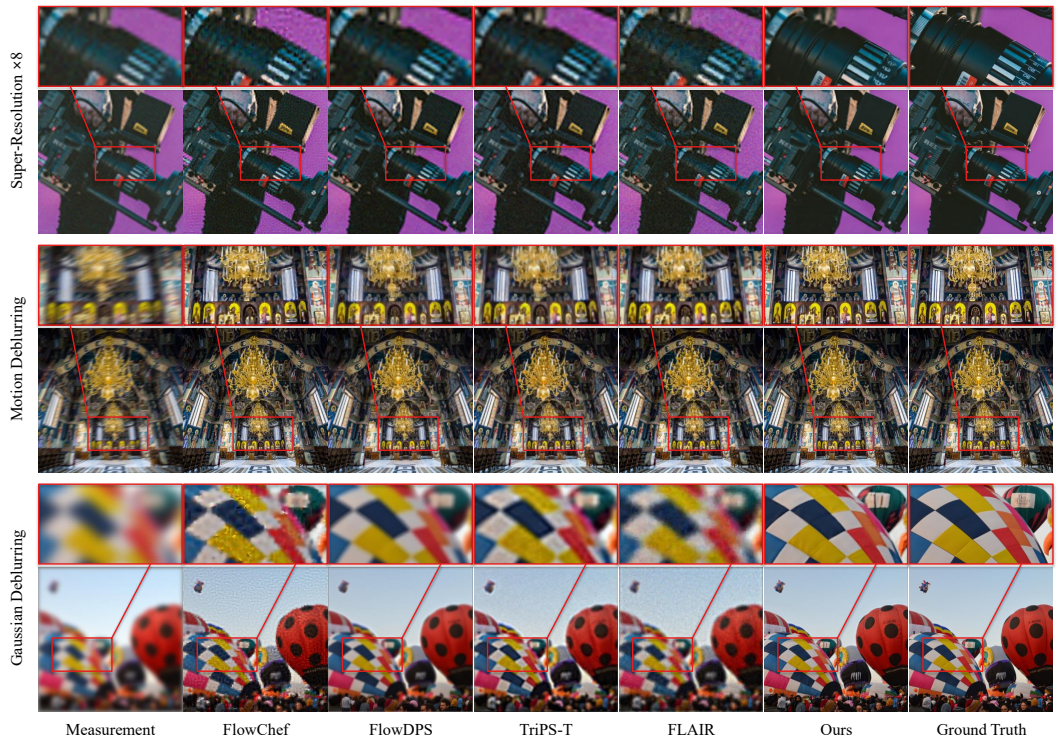

Figure A8: Qualitative comparison on DIV2K using FLUX.1 Kontext.

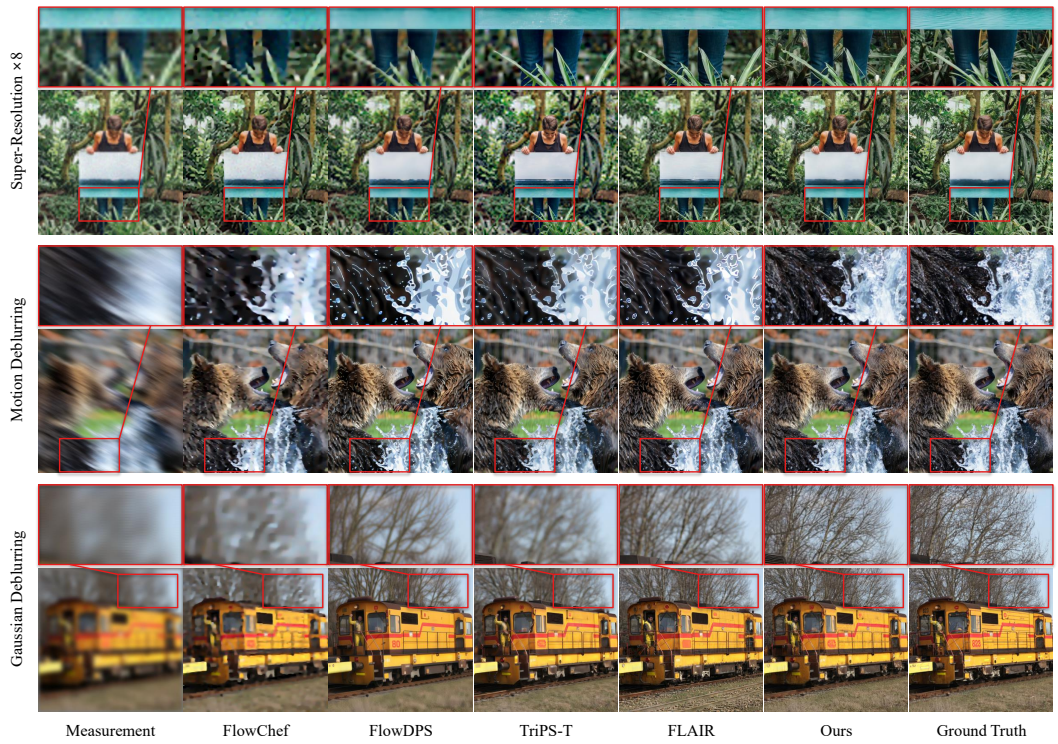

Figure A9: Qualitative comparison on DIV2K using FireRed-Image-Edit.

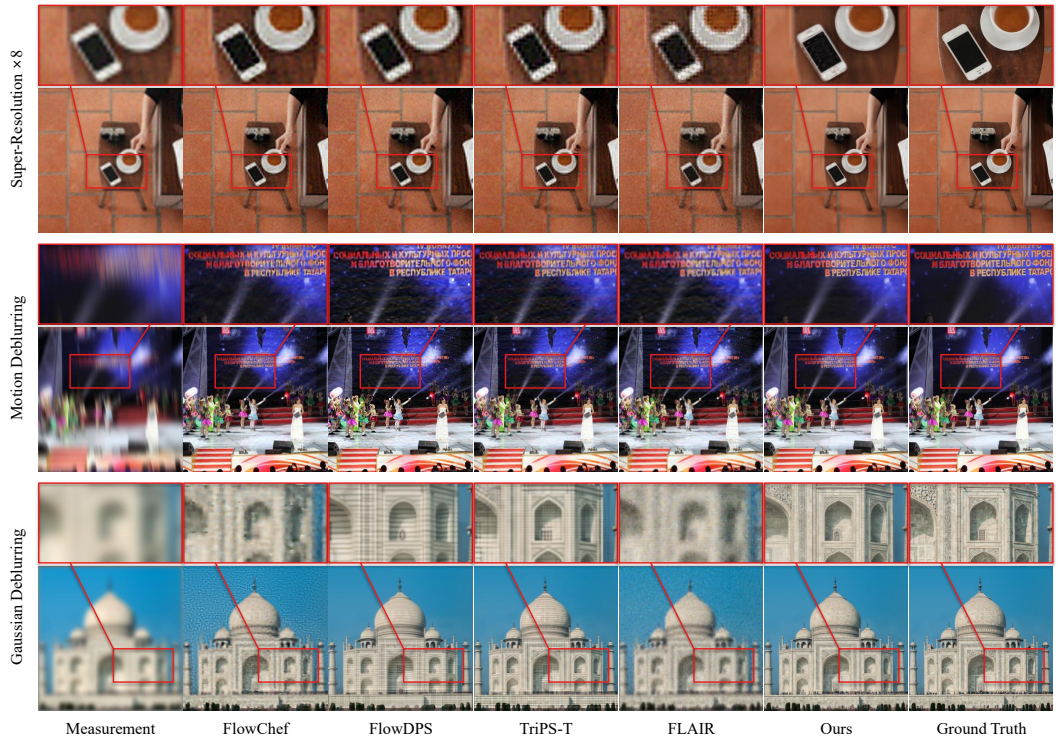

Figure A10: Qualitative comparison on DIV2K using LongCat-Image-Edit.

F Additional Qualitative Results

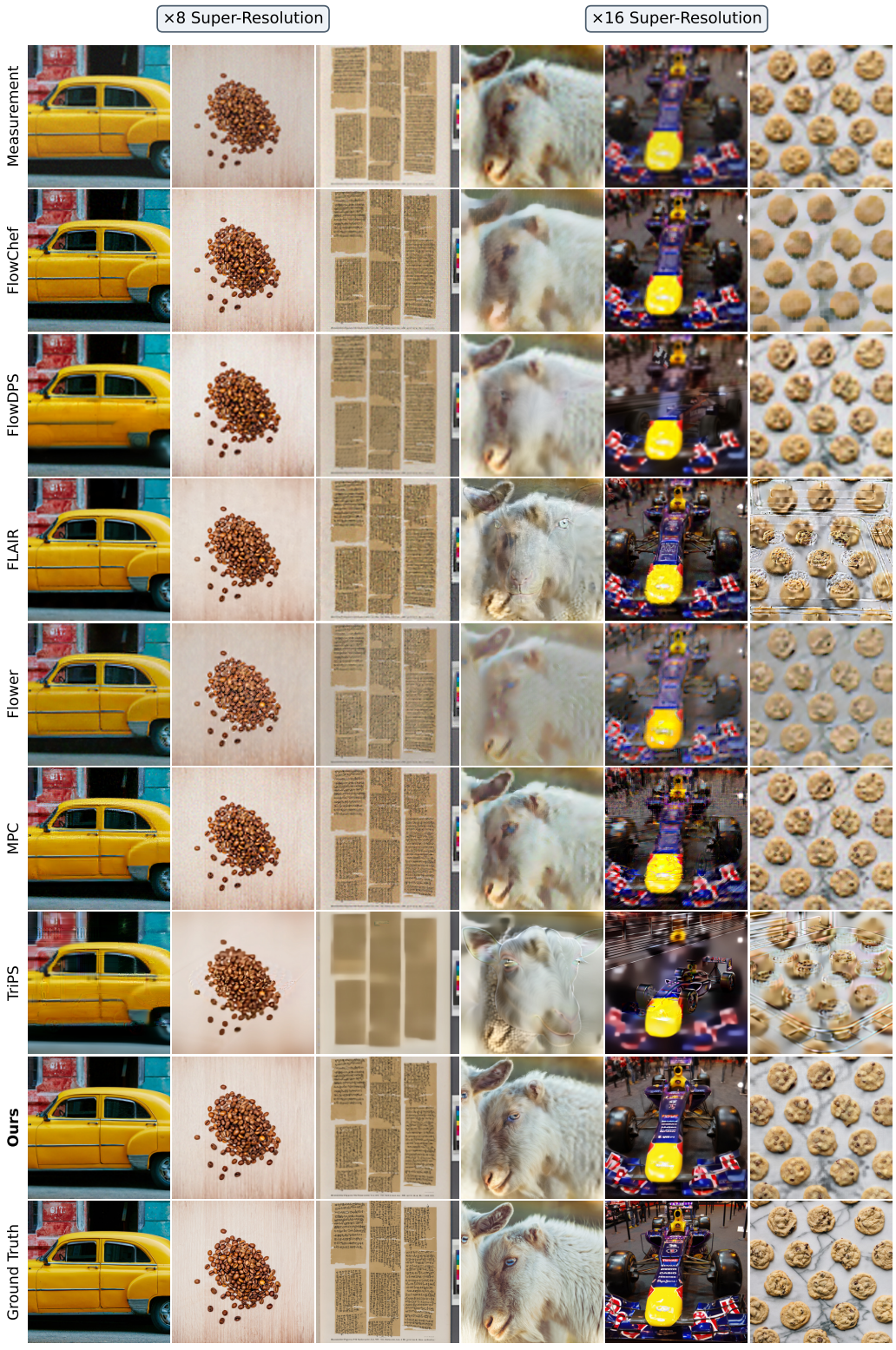

Figure A11: Additional qualitative comparisons on DIV2K for super-resolution ($\times 8$, left; $\times 16$, right).

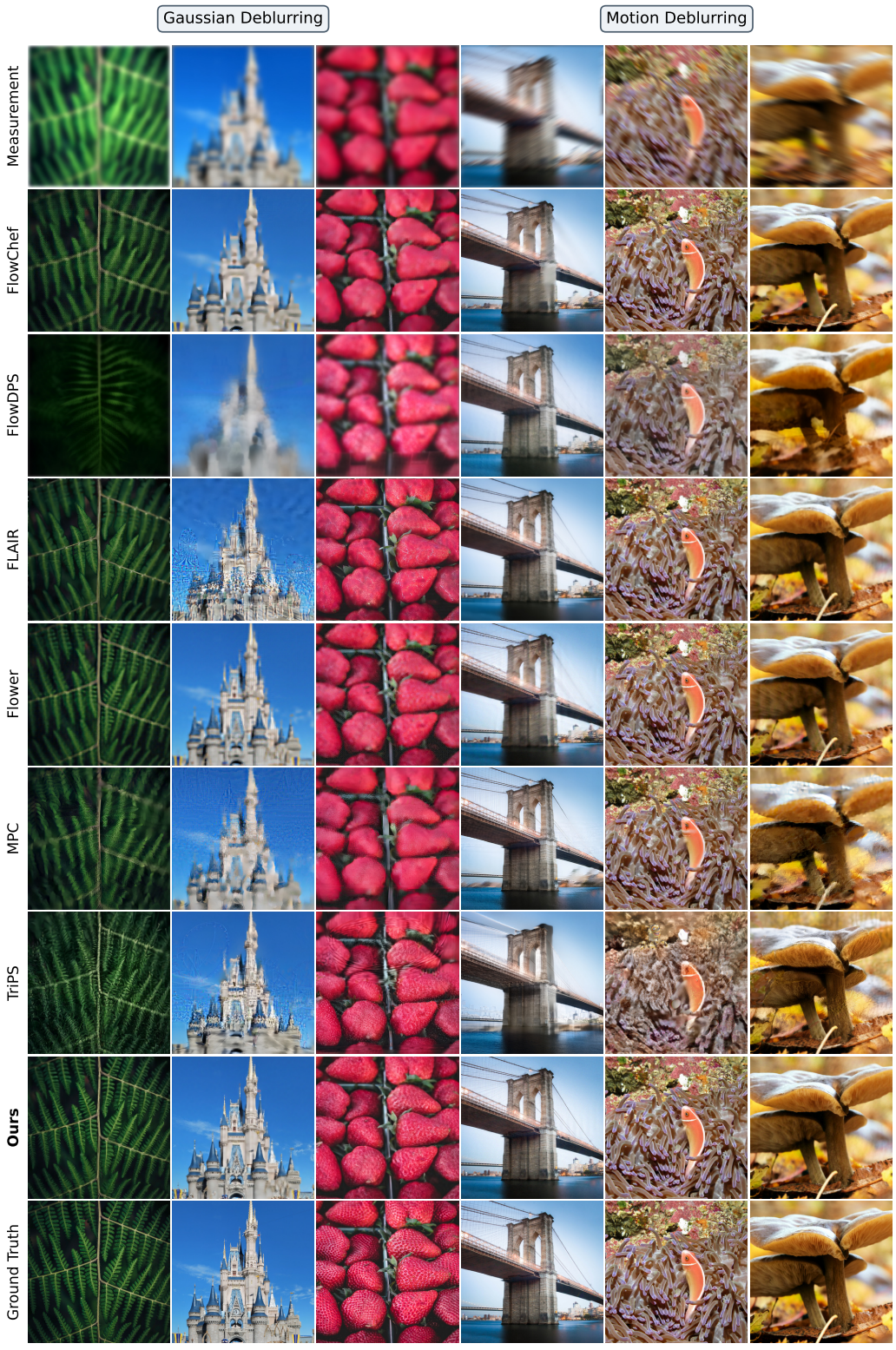

Figure A12: Additional qualitative comparisons on DIV2K for Gaussian deblurring (left) and motion deblurring (right).

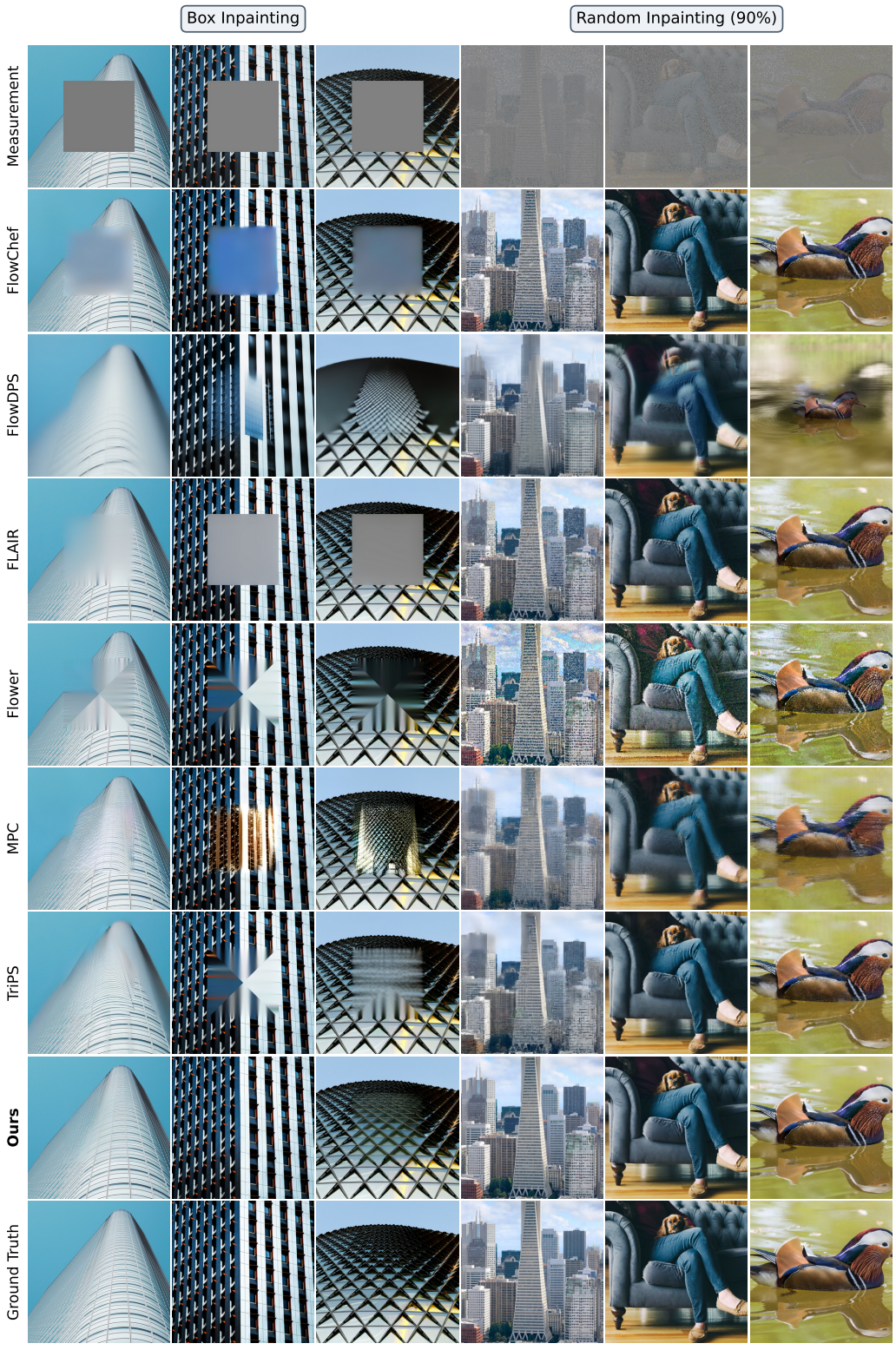

Figure A13: Additional qualitative comparisons on DIV2K for box inpainting (left) and random inpainting with 90% missing pixels (right).

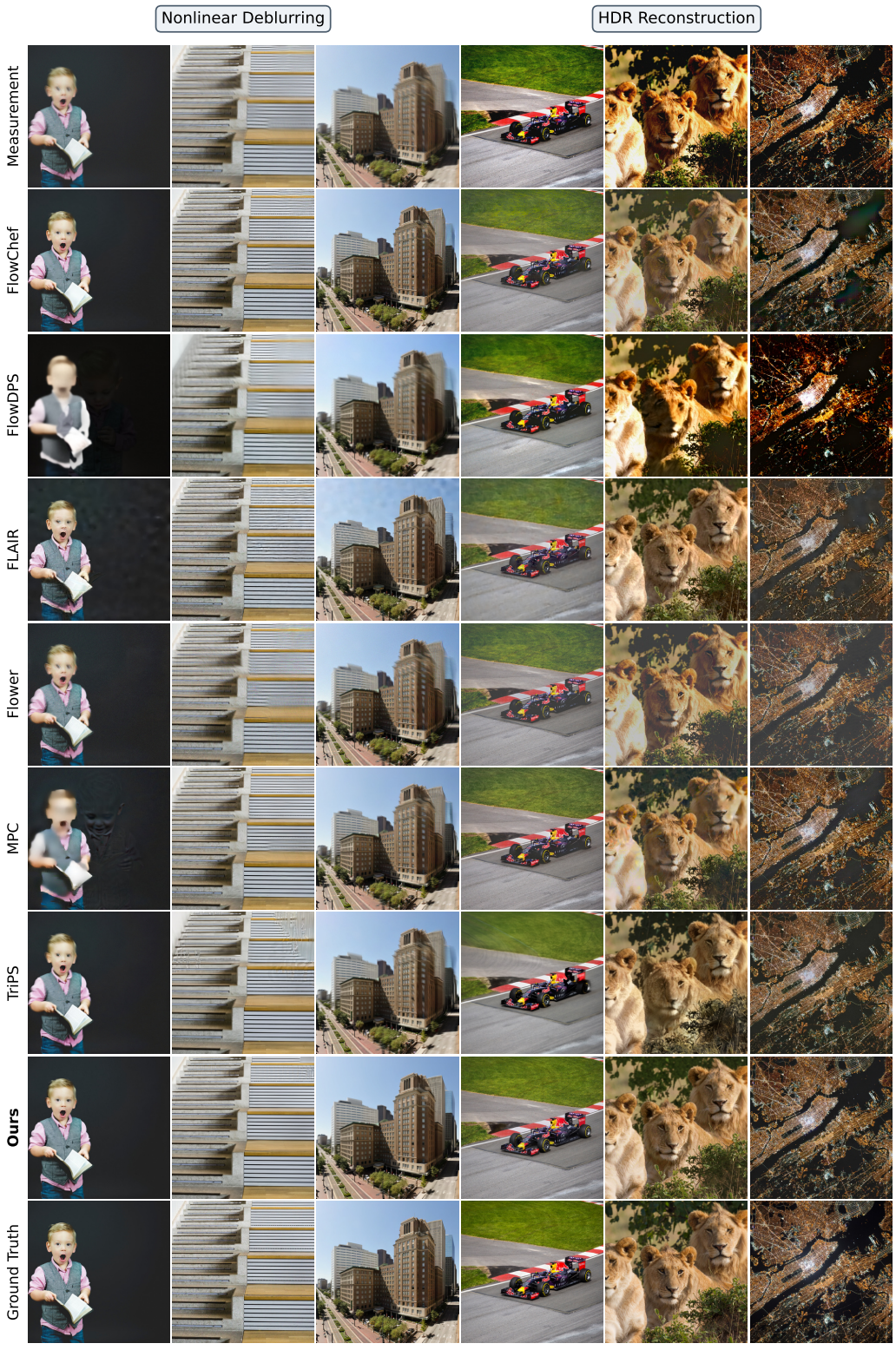

Figure A14: Additional qualitative comparisons on DIV2K for nonlinear deblurring (left) and HDR reconstruction (right).



Figure A15: Additional qualitative comparisons on FFHQ for super-resolution ($\times 8$, left; $\times 16$, right).

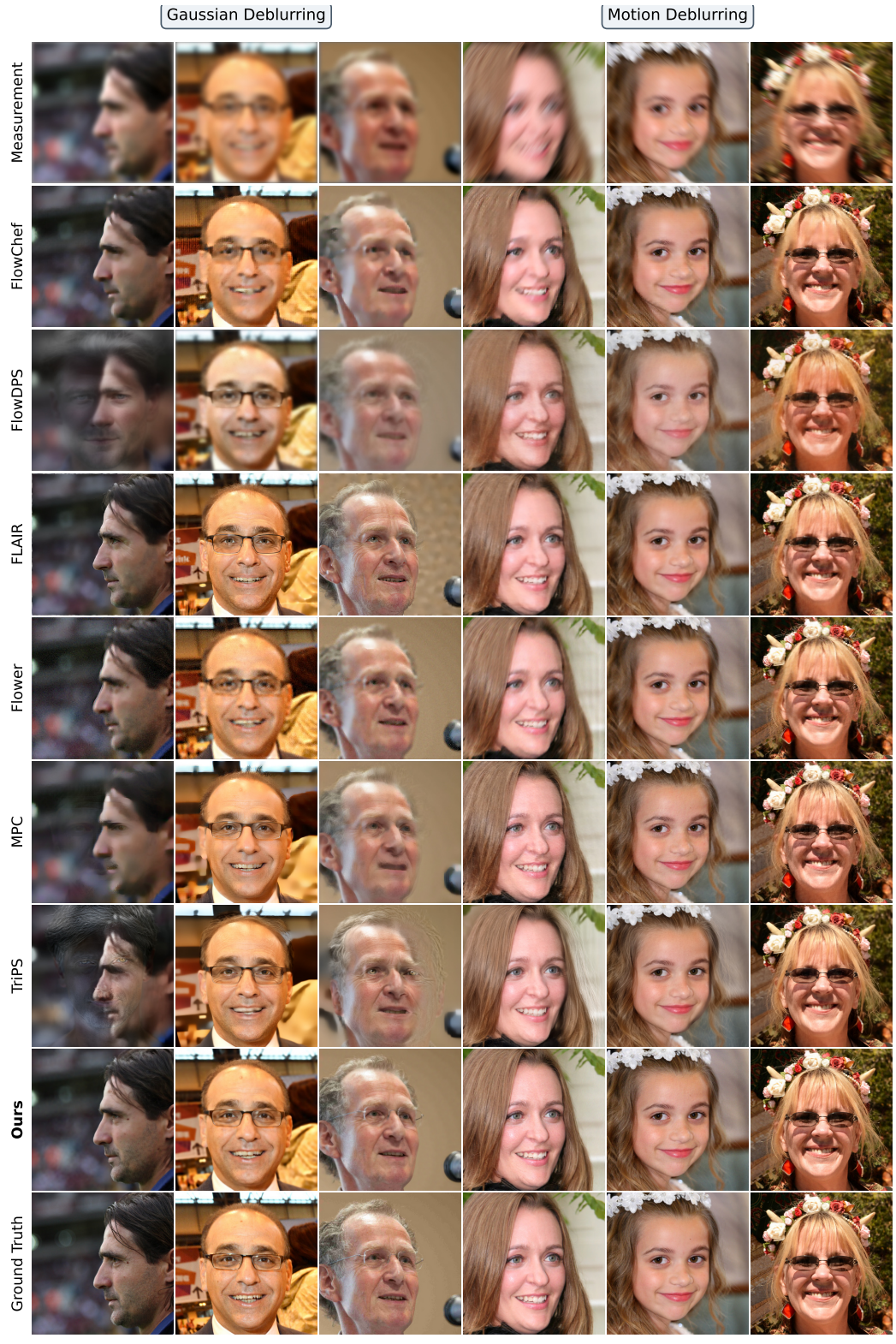

Figure A16: Additional qualitative comparisons on FFHQ for Gaussian deblurring (left) and motion deblurring (right).

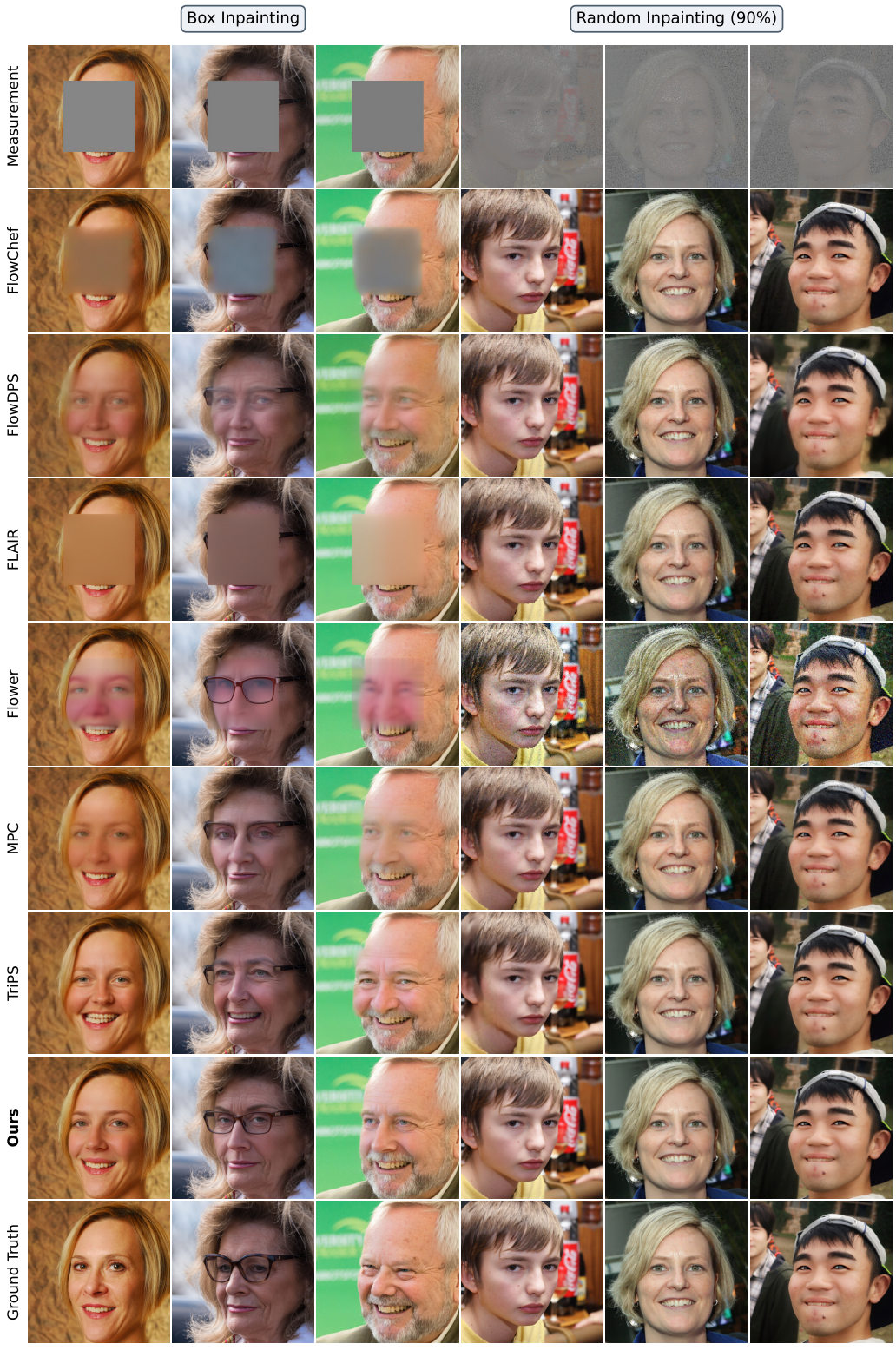

Figure A17: Additional qualitative comparisons on FFHQ for box inpainting (left) and random inpainting with 90% missing pixels (right).

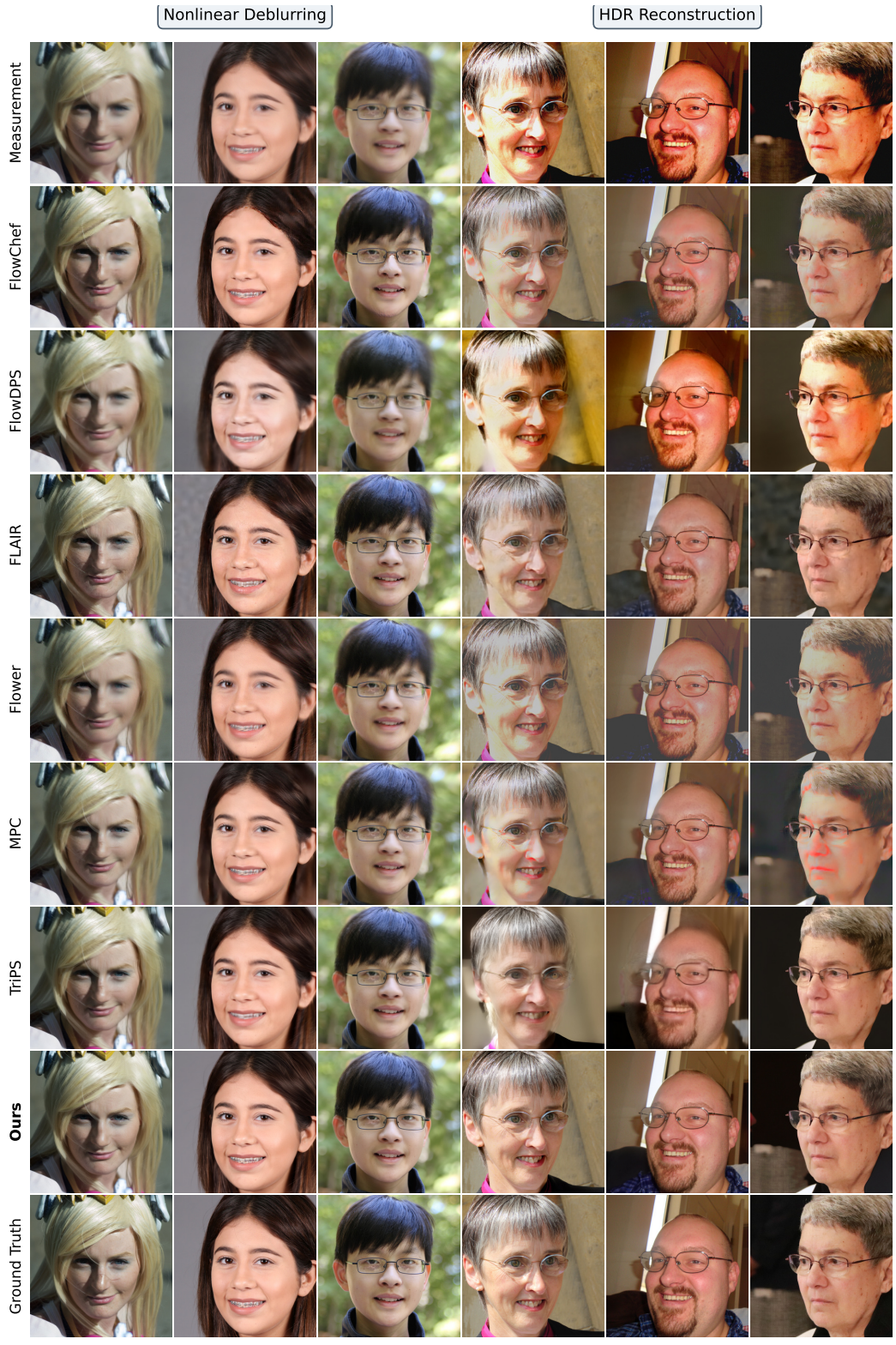

Figure A18: Additional qualitative comparisons on FFHQ for nonlinear deblurring (left) and HDR reconstruction (right).

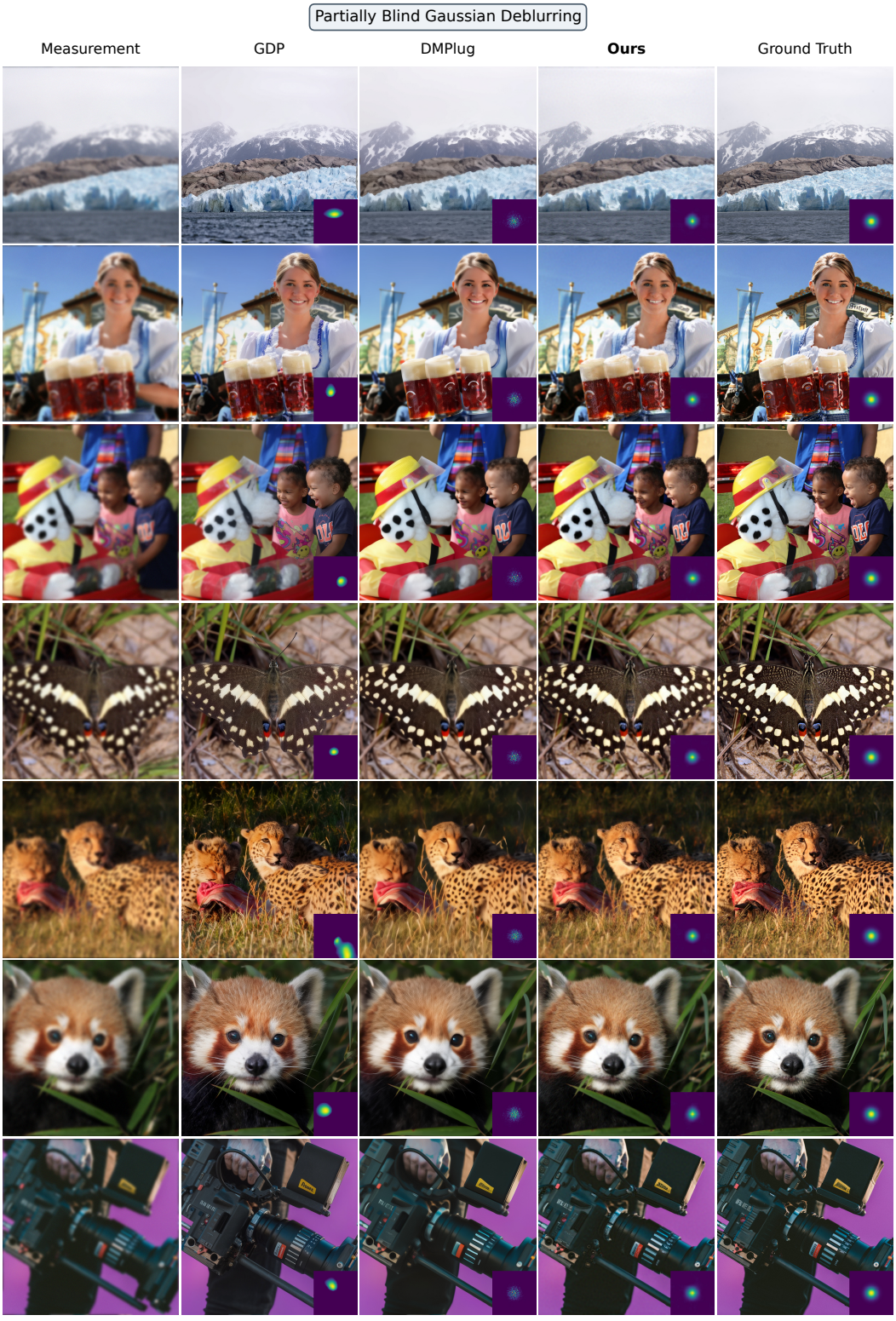

Figure A19: Qualitative comparisons on DIV2K for partially blind Gaussian deblurring.

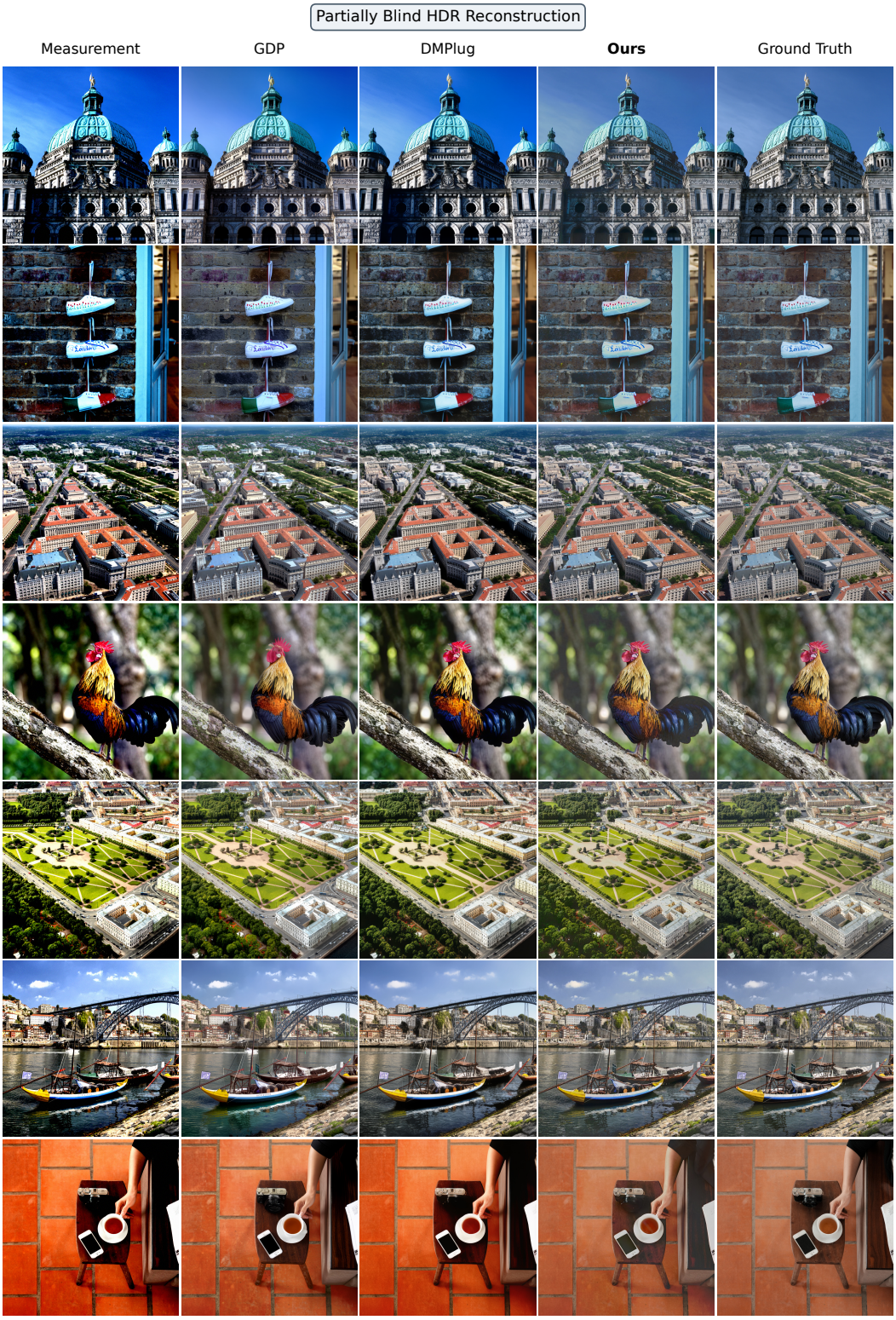

Figure A20: Qualitative comparisons on DIV2K for partially blind HDR reconstruction.