PLUG-IN IMAGE QUALITY CONTROL FOR FEW-STEP POSTERIOR DIFFUSION SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based super-resolution (SR) has made remarkable progress, mainly through diffusion-prior approaches that rely on assumed degradation likelihoods or generative priors. Posterior diffusion SR instead learns directly from LR–HR pairs, but plug-in mechanisms for controlling the image quality of pretrained models remain less explored. To the best of our knowledge, we propose the first plug-in module for few-step posterior diffusion SR that enables image quality control without retraining the core model. Motivated by high-order diffusion solvers, we formulate a derivative-based correction targeting the per-step posterior-mean discrepancy, which may include the effects of accumulated discretization error. This discrepancy is related to a covariance-weighted KL gradient, motivating an LR-based image-guidance surrogate whose effective scale is calibrated for each denoising step. The sign of the guidance selects between a fidelity-oriented mode and a sharpness-oriented mode. In blind SR, the fidelity-oriented mode improves fidelity metrics, while the sharpness-oriented mode achieves higher scores on non-reference perceptual metrics such as MANIQA and CLIPIQA surpassing Stable Diffusion-based methods on RealSR. Our module provides continuous access to a range of image-quality operating points within a single pretrained posterior model.

1 INTRODUCTION

Image super-resolution (SR) is a fundamental problem in low-level vision, aiming to recover high-resolution (HR) images from their low-resolution (LR) counterparts. Despite substantial progress of regression or GAN based methods, these methods still struggle with the intrinsic ill-posedness of SR.

Diffusion models Ho et al. (2020); Song et al. (2021b) have recently emerged as a powerful generative framework for addressing this ambiguity. Diffusion models can effectively model complex image distributions and recover high-frequency details in ill-posed restoration problems. Although diffusion models were initially criticized for their high computational cost, subsequent advances in efficient sampling and diffusion formulations Song et al. (2021a); Yue et al. (2023) and distillation-based acceleration Wu et al. (2024a); Dong et al. (2024); Chen et al. (2025); Yue et al. (2025) have reduced inference to one or a few sampling steps, substantially narrowing the runtime gap with regression-based SR. In parallel, Stable Diffusion-based methods Lin et al. (2024); Wu et al. (2024b); Yang et al. (2024); Chen et al. (2024c) exploit large-scale generative and semantic priors to further improve perceptual quality.

Diffusion-based SR can be broadly viewed from two complementary perspectives. *Diffusion-prior* approaches first exploit a pretrained or separately learned image prior $p(x)$ and incorporate the LR observation through a degradation likelihood, restoration-specific conditioning, or auxiliary control mechanism Kawar et al. (2022); Wang et al. (2023b); Lin et al. (2024). Such approaches benefit from powerful generative priors and provide considerable flexibility for introducing additional guidance at inference time. However, connecting the observation to the generative prior requires an additional degradation or conditioning mechanism, and explicitly modeling real-world degradation can be difficult because practical imaging pipelines involve complex combinations of blur, noise, compression, and camera processing.

In contrast, *posterior learning* approaches directly learn a conditional restoration distribution $p(x|y)$ from paired LR–HR observations Saharia et al. (2021); Li et al. (2022); Yue et al. (2023); Zhang et al. (2025). In this work, we use the term *posterior learning* to refer specifically to this paired

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.

Table 1: Comparison between diffusion-prior and posterior-learning paradigms for image super-resolution.

Aspect	Diffusion Prior SR	Posterior Learning SR
Modeling	Learn or reuse image prior $p(x)$; incorporate y via likelihood, conditioning, or guidance	Directly learn $p(x\|y)$ from paired LR–HR data
Degradation	Explicit degradation model or learned observation-conditioning module	**Implicitly represented by the training dataset**
Pretraining	Large-scale generative prior can be directly exploited	Typically trained directly for the restoration task
Post-Control	Guidance / auxiliary conditioning (e.g., CLIP, ControlNet, restoration guidance)	**Our plug-in guidance**

conditional diffusion paradigm. The degradation characteristics represented in the training data are therefore implicitly incorporated into the learned conditional distribution, avoiding the need to specify an explicit measurement likelihood at inference time. This formulation naturally extends the supervised learning paradigm widely adopted in regression-based SR. However, post-training control of pretrained posterior diffusion models remains considerably less explored: unlike generative-prior frameworks, established plug-in mechanisms for continuously adjusting their perceptual operating point are still lacking.

For example, DiffBIR Lin et al. (2024) introduces region-adaptive restoration guidance that modifies the denoising process at inference time without retraining the pretrained model, allowing users to balance perceptual quality and reconstruction fidelity through a tunable scale. Such inference-time control transforms a fixed model from a single quality setting into a family of achievable operating points. Consequently, a pretrained model can produce outputs with different image-quality characteristics without retraining, providing greater flexibility when adapting the model to different application requirements. This property is particularly attractive in practical deployment, where the preferred balance among fidelity, sharpness, perceptual quality, and other image characteristics may vary across applications.

Comparable control mechanisms, however, have not been systematically developed for posterior diffusion SR. The output characteristics of a posterior model are typically coupled with its learned transition process and sampling procedure. Consequently, moving a pretrained model toward a different quality operating point may require fine-tuning, retraining, or model-specific novel sampling method. Furthermore, evaluating a sampling method only at a fixed operating point can make it difficult to distinguish improvements originating from the sampling formulation itself from changes in the output characteristics induced by a particular sampling configuration. Providing continuous post-training control therefore offers not only practical flexibility but also a characterization of the achievable image-quality range of a pretrained posterior sampler.

From a numerical analysis perspective, posterior diffusion sampling can be interpreted as a sequence of discrete transitions along an LR-to-HR generative trajectory. Low-order local approximations are computationally efficient, but their discretization errors can accumulate and become increasingly important as the number of sampling steps decreases. Higher-order diffusion solvers Lu et al. (2022; 2025); Zheng et al. (2023) explicitly address such numerical errors for conventional diffusion trajectories. For general image generation, Li & van der Schaar (2024) analyzed cumulative errors between forward and backward processes, but the role of discretization error as a controllable correction mechanism in posterior SR remains largely unexplored.

Our contributions. The contributions of this work are summarized as follows. **First**, we propose the first *plug-in quality-control module* for few-step posterior diffusion SR, enabling continuous control of pretrained models through marginal calibration without retraining the core diffusion model. **Second**, we introduce an *image-based guidance formulation* for posterior diffusion and establish its formal equivalence to a second-order correction term, thereby connecting guidance-based conditioning with numerical error correction. **Third**, by analyzing discretization errors in posterior diffusion sampling, we theoretically derive a principled pair of guidance scales, enabling stable and continuous control of the resulting plug-in guidance.

2 Related Works

Image Super-Resolution Since the emergence of deep neural network-based image super-resolution, significant advancements have been made in this field. Generative adversarial networks (GANs) Goodfellow et al. (2014) have played a crucial role in enhancing details in super-resolution outputs Wang et al. (2021; 2018). By leveraging adversarial learning, GANs improve texture sharpness and realism in super-resolved images. Recently, diffusion models Saharia et al. (2021); Li et al. (2022); Kawar et al. (2022); Wang et al. (2023b) have emerged as a powerful alternative for image super-resolution. Originally developed for novel image synthesis, these models have demonstrated exceptional capabilities in texture restoration and fine-detail reconstruction. Unlike GANs, diffusion models progressively refine images from noise, providing a more stable and controlled SR framework. Diffusion-based SR outperforms GAN-based approachesRombach et al. (2022). Furthermore, developments in VQGAN Esser et al. (2021) and stable diffusion models Rombach et al. (2022) have significantly influenced image synthesis Saharia et al. (2022), expanding beyond SR into text-conditioned image generation. Pretrained Stable Diffusion models have become widely adopted as diffusion priors, leading to a rapid increase in research on diffusion prior–based super-resolution. With methods such as ControlNet Zhang et al. (2023) and CLIP Radford et al. (2021a), it has become possible to steer image generation or restoration toward desired semantic attributes, resulting in significant improvements in perceptual quality Lin et al. (2024); Wu et al. (2024b); Yang et al. (2024).
Performance Enhancement Diffusion models for image super-resolution face an inherent challenge of high computational overhead Saharia et al. (2021); Ho et al. (2020). Various efforts have been made to address this issue Song et al. (2021a); Lu et al. (2022; 2025); Zheng et al. (2023); Wu et al. (2024a); Wang et al. (2024). Starting with DDIM, subsequent works on DPM solvers have significantly reduced the number of iteration steps from thousands to just tens. Residual-based approaches in image synthesis Liu et al. (2024) and image super-resolution Yue et al. (2023) have demonstrated stable performance even with fewer than ten steps. More recently, techniques such as model distillation Wang et al. (2024) and single-step SR Wu et al. (2024a) have emerged, further pushing the boundaries of efficiency in diffusion-based SR.
Error Estimation Reducing the number of iteration steps inherently increases step intervals in finite difference methods, making diffusion models more susceptible to numerical errors Strang (2007); Peter & Eckhard (1992); Hochbruck & Ostermann (2005). Consequently, research has been conducted to analyze and mitigate numerical errors and their effects Li & van der Schaar (2024); Li et al. (2024). Trajectory analysis Chen et al. (2024b;a) is also effective for detailed analysis of numerical errors. In Li & van der Schaar (2024), the authors investigated errors in pretrained diffusion models by analyzing discrepancies between forward and reverse processes.

3 Plug-in Module for Few-Step Posterior Diffusion Model

3.1 Background for Inverse Problem Solver

Super-resolution (SR) is an ill-posed inverse problem, where the goal is to recover the high-resolution (HR) image $\mathbf{x}$ from its degraded low-resolution (LR) observation $\mathbf{y}$. The forward measurement process can be expressed as $\mathbf{y} = \boldsymbol{A}(\mathbf{x}) + \mathbf{n}, \quad \mathbf{y}, \mathbf{n} \in \mathbb{R}^n,\ \mathbf{x} \in \mathbb{R}^d$ where $\boldsymbol{A}(\cdot) : \mathbb{R}^d \to \mathbb{R}^n$ is the degradation operator (e.g., bicubic downsampling, blur, or camera pipeline), and $\mathbf{n}$ denotes measurement noise. From a Bayesian perspective, the posterior distribution is given by $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x})/p(\mathbf{y})$ where $p(\mathbf{x})$ the prior over natural images, $p(\mathbf{y}|\mathbf{x})$ the likelihood for degradation, $p(\mathbf{x}|\mathbf{y})$ the posterior. Conditioning the diffusion process on $\mathbf{y}$ yields a posterior-guided reverse ODE:

$$d\mathbf{x} = \Big[f(t)\,\mathbf{x} - \tfrac{1}{2} g^2(t)\, \nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y}) \Big]\, dt, \tag{1}$$

where the score term $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y})$ incorporates the assumed likelihood.

In **Diffusion Prior** methods Saharia et al. (2021); Li et al. (2022); Wang et al. (2023b); Wu et al. (2024b), The likelihood $p(\mathbf{y}|\mathbf{x})$ is typically assumed to be a Gaussian degradation model: $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; A(\mathbf{x}), \sigma^2 I)$ and the diffusion model serves as the image prior $p(\mathbf{x})$.

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x})$$

Thus, the accuracy of prior-based SR crucially depends on the correctness of the degradation model.
Posterior learning approaches Yue et al. (2023); Zhang et al. (2025); Yue et al. (2024); Luo et al.

(2023) directly learn the conditional distribution $p(\mathbf{x}|\mathbf{y})$ from paired LR–HR data, bypassing the need for an explicit likelihood model. In this case, the degradation process $\boldsymbol{A}(\cdot)$ does not need to be analytically specified; it is implicitly encoded through the training dataset. The training objective is to maximize the conditional log-likelihood: $\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\big[\log p_\theta(\mathbf{x}|\mathbf{y})\big]$. This makes posterior approaches attractive for real-world SR, where the degradation is complex or unknown. However, posterior SR is sensitive to discretization errors during sampling and currently lacks plug-in quality control mechanisms, which are more natural in prior-based frameworks.

3.2 Background for High Order Diffusion ODE solvers

The sampling procedure of DPMs gradually denoises from pure Gaussian random variables to produce meaningful data, which can be viewed as discretizing the diffusion SDEs. From Lu et al. (2022; 2025); Zheng et al. (2023), Dedicated high-order diffusion ODE solvers can generate high-quality samples in 10 to 20 steps for sampling. However, the effectiveness of high order diffusion ODE solvers has not been carefully examined in super-resolution. Different from text-to-image generation, super-resolution typically have correct answers in high resolution images, which requires us to control diffusion process very carefully. Following equations from Lu et al. (2025) are 1st order and 2nd order diffusion ODE solvers. Details about this equation are available in appendix A .

SDE-DPM-Solver++2 Let $\mathbf{z}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Assume we have a previous solution $\mathbf{x}_r$ with its model output $\mathbf{x}_\theta(\mathbf{x}_r, r)$ at time $r < t$. Denote $r_1 = \frac{\lambda_r - \lambda_s}{h}$. By assuming $\mathbf{x}_\theta(\mathbf{x}_\lambda, \lambda) \approx \mathbf{x}_\theta(\mathbf{x}_s, s) + \frac{\lambda-\lambda_s}{r_1 h}(\mathbf{x}_\theta(\mathbf{x}_r, r) - \mathbf{x}_\theta(\mathbf{x}_s, s))$, we have

$$\mathbf{x}_t = \frac{\sigma_t}{\sigma_s}e^{-h}\mathbf{x}_s + \alpha_t(1-e^{-2h})\mathbf{x}_\theta(\mathbf{x}_s, s) + \frac{\alpha_t(1-e^{-2h})}{2}\left(\frac{\mathbf{x}_\theta(\mathbf{x}_r, r) - \mathbf{x}_\theta(\mathbf{x}_s, s)}{r_1}\right) + \sigma_t\sqrt{1-e^{-2h}}\mathbf{z}_s. \quad (2)$$

3.2.1 High order ODE solver for Posterior learning method

Motivated by section 3.2, we generate high order ODE solver from 1st-order Posterior learning method Yue et al. (2023). Details of derivation is available in appendix C.1 . Now, we obtain the **2nd-order sampler of Yue et al. (2023)**:

$$\mathbf{x}_{t-1} = \frac{\eta_{t-1}}{\eta_t}\mathbf{x}_t + \frac{\alpha_t}{\eta_t}f_{\boldsymbol{\theta}}^{(t)} + \frac{\alpha_t}{2\eta_t}\left(\frac{f_{\boldsymbol{\theta}}^{(t)} - f_{\boldsymbol{\theta}}^{(t+1)}}{r_1}\right) + \kappa\sqrt{\frac{\eta_{t-1}\,\alpha_t}{\eta_t}}\,\mathbf{z}, \quad (3)$$

where $f_{\boldsymbol{\theta}}^{(t)} \coloneqq f_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}_0, t)$, $f_{\boldsymbol{\theta}}^{(t+1)} \coloneqq f_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}, \mathbf{y}_0, t+1)$, $r_1 = \frac{\lambda_t - \lambda_{t+1}}{h} = \frac{\log(\eta_{t+1}/\eta_t)}{\log(\eta_t/\eta_{t-1})}$, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Results of applying the 2nd-order correction term within the posterior learning framework are presented in Figure 2. In DPM-Solver, the perceptual quality change induced by the 2nd-order term exhibits considerable variance across different images. In ResShift, while the 2nd-order term does provide a degree of image control, suppressing the artifacts it readily introduces requires an additional scalar weight applied to the term. Although a floating-point scale in the range [0.0,1.0] can mitigate these artifacts, there is no principled criterion for forming scale value pairs across multiple denoising steps, and this approach does not yield continuous, well-behaved image control. Furthermore, the fidelity improvement achievable through this scheme is limited, making it difficult to analyze how the vector direction of the 2nd-order term itself influences output quality. To address these limitations, the following section proposes a theoretical framework that forgoes the conventional Runge–Kutta formulation for computing the 2nd-order term and instead derives the correction directly via gradient computation. This reformulation enables a systematic analysis of the role that the direction of the additive trajectory vector plays in determining output quality.

3.2.2 Image Quality Control by the 2nd-Order Term

As shown in the previous sections, a higher-order sampler is obtained by adding a 2nd-order correction term to the first-order update. When computed accurately, this term reduces the discrepancy between the SR estimate and the HR target, thereby improving fidelity-oriented metrics such as PSNR.

However, improved distortion metrics do not necessarily imply better perceptual quality. Since SR is ill-posed, the MSE-optimal estimator is the posterior mean $\mathbb{E}[\mathbf{x}|\mathbf{y}]$, which averages plausible HR

solutions and can suppress high-frequency details Ledig et al. (2017); Blau & Michaeli (2020). Pixel-wise errors also penalize perceptually plausible but misaligned details more strongly than smooth predictions, producing the well-known perception–distortion trade-off Blau & Michaeli (2020).

The benefit of the 2nd-order term also depends on its weight. Let $\boldsymbol{d}$ denote the correction direction and w its weight. The error $|\mathbf{x} - \hat{\mathbf{x}} - w\boldsymbol{d}|_2^2$ is minimized at $w^\star = \langle \mathbf{x} - \hat{\mathbf{x}}, \boldsymbol{d}\rangle / |\boldsymbol{d}|_2^2$. Thus, increasing w initially reduces the error, whereas $w > w^\star$ overshoots the target and degrades PSNR. A large weight can additionally amplify estimation errors contained in the finite difference of network predictions.

Meanwhile, no-reference perceptual metrics such as MANIQA Yang et al. (2022) and CLIPIQA Wang et al. (2023a) are sensitive to perceptually relevant high-frequency characteristics. We observe that the 2nd-order term behaves similarly to unsharp masking: because it is proportional to the difference between consecutive $\mathbf{x}_0$ predictions, it emphasizes newly recovered high-frequency components. Its sign and magnitude can therefore be used to enhance image acuity through sharpening(Appendix C.2). Excessive amplification, however, introduces halos, overshoot, and noise amplification, eventually degrading perceptual quality Trentacoste et al. (2012); Hu et al. (2026).

Accordingly, we exploit the 2nd-order term to control image quality toward either improved fidelity or improved no-reference perceptual quality. In diffusion-prior methods, DPM-Solver provides a well-defined higher-order correction. In posterior learning, however, the correction derived in the previous section is less accurate, and there is no principled rule for selecting its scale at each step of a multi-step sampler. As shown in Figure 9, arbitrary scale combinations can readily introduce artifacts, making accurate scale selection critical.

In the next section, we derive the derivative-based 2nd-order correction and its associated parameters for few-step posterior models. This converts a pretrained model with a fixed operating point into a continuously controllable model and exposes a range of attainable image-quality metric values.

3.3 Our Proposal

We propose a framework for continuous image quality control in few-step posterior diffusion, motivated by the artifacts observed with the multistep correction in Section 3.2.1. We first relate the discrepancy between learned and HR-referenced posterior means to a covariance-weighted KL gradient, then introduce an LR-based auxiliary guidance potential and calibrate its per-step scale. Following the discrepancy definitions in Appendix C.3, we distinguish this oracle comparison from numerical discretization error alone. The Gaussian calculations below fix the observation $\mathbf{y}_0$ and require positive-definite covariances independent of $\mathbf{x}_t$ at each nondegenerate step. We use the ResShift Yue et al. (2023) and UPSR Zhang et al. (2025) formulations in Appendix B, with covariance conventions and proofs in Appendices C.3 and C.4.

Posterior-mean discrepancy and a KL gradient. Let $\boldsymbol{\mu}_\theta(\mathbf{x}_t) = \frac{\eta_{t-1}}{\eta_t}\mathbf{x}_t + \frac{\alpha_t}{\eta_t} f_\theta(\mathbf{x}_t, \mathbf{y}_0, t)$ be the learned reverse mean and $\boldsymbol{\mu}_q(\mathbf{x}_t) = \frac{\eta_{t-1}}{\eta_t}\mathbf{x}_t + \frac{\alpha_t}{\eta_t}\mathbf{x}_0$ the oracle posterior mean conditioned on the HR image $\mathbf{x}_0$. Define $\mathcal{P}_t = \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_t)$ and $\mathcal{Q}_t = \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_t)$. Here $v_t = \kappa^2 \frac{\eta_{t-1}}{\eta_t}\alpha_t$, with $\boldsymbol{\Sigma}_t = v_t\mathbf{I}$ for ResShift and $\boldsymbol{\Sigma}_t = v_t \mathbf{W}_u(\mathbf{y}_0)^2$ for UPSR, where $\mathbf{W}_u$ contains fixed, positive spatial noise weights. Proposition C.1 gives

$$\boldsymbol{\delta}_t := \boldsymbol{\mu}_q - \boldsymbol{\mu}_\theta = \frac{\alpha_t}{\eta_t}\big(\mathbf{x}_0 - f_\theta(\mathbf{x}_t, \mathbf{y}_0, t)\big) = \boldsymbol{A}_t \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(\mathcal{P}_t \| \mathcal{Q}_t) + \boldsymbol{r}_t, \tag{4}$$

Specifically, writing $a_t = \eta_{t-1}/\eta_t$ and $\boldsymbol{J}_t = \partial(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)/\partial \mathbf{x}_t = -a_t\mathbf{I} + \boldsymbol{R}_t$, the exact identity is $\nabla_{\mathbf{x}_t} D_{\mathrm{KL}} = -\boldsymbol{J}_t^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\delta}_t$, and the vector residual is $\boldsymbol{r}_t = a_t^{-1}\boldsymbol{\Sigma}_t \boldsymbol{R}_t^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\delta}_t$. If this residual is negligible, then

$$\tilde{\boldsymbol{\mu}}_\theta := \boldsymbol{\mu}_\theta + \boldsymbol{\delta}_t \approx \boldsymbol{\mu}_\theta + \boldsymbol{A}_t \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(\mathcal{P}_t \| \mathcal{Q}_t). \tag{5}$$

The approximation $\boldsymbol{J}_t \approx -a_t\mathbf{I}$ is equivalent to $\partial f_\theta / \partial \mathbf{x}_t \approx -(\eta_{t-1}/\alpha_t)\mathbf{I}$. It is an additional assumption, not a consequence of Gaussianity or a Taylor expansion alone. The discrepancy $\boldsymbol{\delta}_t$ may reflect the effects of accumulated discretization error through the sampled state $\mathbf{x}_t$, alongside prediction error and conditional uncertainty. Our correction therefore targets an aggregate HR-referenced posterior-mean discrepancy rather than an isolated numerical discretization error. This interpretation preserves the numerical-error motivation without identifying the entire discrepancy with truncation error.

For other affine reverse means, including GOU-based models Luo et al. (2023); Yue et al. (2024), shared state and endpoint terms similarly cancel, but the prediction coefficient and Jacobian approximation must be adapted (Remark C.4). Inverse-covariance statements exclude deterministic terminal steps.

Sensitivity to the reference image. HR references are unavailable at inference time. To examine the effect of replacing a reference, fix $\mathbf{y}_0$ and define forward-reference Gaussians $Q_t^{\mathrm{HR}} = \mathcal{N}(m_t\mathbf{x}_0 + \boldsymbol{b}_t, \boldsymbol{S}_t)$ and $Q_t^{\mathrm{LR}} = \mathcal{N}(m_t\mathbf{c} + \boldsymbol{b}_t, \boldsymbol{S}_t)$, where m_t, $\boldsymbol{b}_t$, and $\boldsymbol{S}_t$ remain fixed during replacement. For $P_t = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{C}_t)$ with fixed positive-definite covariance, let $\boldsymbol{g}_t^{\mathrm{HR}} = \nabla_{\boldsymbol{\mu}_p} D_{\mathrm{KL}}(P_t \| Q_t^{\mathrm{HR}})$ and define $\boldsymbol{g}_t^{\mathrm{LR}}$ analogously. Proposition C.2 gives

$$\|\boldsymbol{g}_t^{\mathrm{HR}} - \boldsymbol{g}_t^{\mathrm{LR}}\|_2 \le L_t W_2(Q_t^{\mathrm{HR}}, Q_t^{\mathrm{LR}}) = |m_t| L_t \|\mathbf{x}_0 - \mathbf{c}\|_2, \qquad L_t = \|\boldsymbol{S}_t^{-1}\|_2. \tag{6}$$

For ResShift, $m_t = 1 - \eta_t$ and $\boldsymbol{S}_t = \kappa^2\eta_t\mathbf{I}$; for UPSR, $\boldsymbol{S}_t = \kappa^2\eta_t\mathbf{W}_u(\mathbf{y}_0)^2$, with the same fixed positive weights for both references. The bound becomes small at early denoising steps where $\eta_t \approx 1$, indicating reduced sensitivity to replacing the HR reference with the LR condition. As denoising proceeds, the bound becomes less restrictive, motivating timestep-dependent calibration of the LR-based surrogate. This provides a reference-sensitivity argument for LR substitution, while its use in reverse-process guidance is evaluated empirically.

LR-based image guidance. We use the LR condition $\mathbf{c}$, represented in the same latent space as $\mathbf{x}_t$, to construct an image-guidance surrogate. Let P_t^{aux} and Q_t^{aux} denote the Gaussian surrogates centered at $\mathbf{x}_t$ and $\mathbf{c}$, respectively, with covariance factors implicit in the model-specific KL evaluation. We write $\ell_t(\mathbf{x}_t; \mathbf{c})$ for the control potential approximated by the negative of this LR-based KL discrepancy. The guidance computation is fixed for each backbone; only its effective per-step scale is calibrated. With guided transition kernels $K_{s,t}$ defined below, the controlled process is

$$p_s(\mathbf{x}_0 \mid \mathbf{y}_0; \mathbf{c}) = \int p_\theta(\mathbf{x}_T \mid \mathbf{y}_0) \prod_{t=1}^{T} K_{s,t}(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0; \mathbf{c})\, \mathrm{d}\mathbf{x}_{1:T}. \tag{7}$$

A first-order expansion of a log-likelihood or control potential yields the Gaussian mean-shift algebra of classifier guidance Dhariwal & Nichol (2021) (Proposition C.3). Evaluating the potential gradient at $\mathbf{x}_t$, as an additional approximation to evaluation at the reverse mean, motivates the implemented kernel using the surrogate gradient in Eq. 9

$$K_{s,t} := \mathcal{N}\big(\boldsymbol{\mu}_\theta - s\alpha_t^* \Sigma_{\theta,t} \boldsymbol{G}_t,\ \Sigma_{\theta,t}\big). \tag{8}$$

Here $\Sigma_{\theta,t}$ denotes the base model's reverse covariance. The Gaussian image-guidance interpretation in Appendix C.4 motivates the approximation

$$\nabla_{\mathbf{x}_t} \ell_t(\mathbf{x}_t; \mathbf{c}) \approx -\nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(P_t^{\mathrm{aux}} \| Q_t^{\mathrm{aux}}) = -\boldsymbol{G}_t. \tag{9}$$

Here $\boldsymbol{G}_t$ denotes the evaluated LR-based KL gradient, including the fixed covariance factors used in its computation. Any common per-timestep scalar normalization of this gradient is absorbed into the calibrated coefficient α_t^*, while spatial weighting remains part of the fixed gradient computation. Thus, no separate auxiliary-covariance parameter is estimated. When $\mathbf{c}$ is a deterministic function of the already observed $\mathbf{y}_0$, it supplies no additional Bayesian information; ℓ_t is a designed control potential. Its gradient is not identified with the oracle reverse-kernel gradient up to a scalar. Instead, it provides a practical control direction whose scale is calibrated empirically.

Reference region. When HR references are available, we fit scalar coefficients to the oracle mean discrepancy:

$$\Delta_t = \boldsymbol{\mu}_q - \boldsymbol{\mu}_\theta, \tag{10}$$

$$(\widehat{A}_t, \widehat{B}_t) = \arg\min_{a_t, b_t \in \mathbb{R}} \big\| \Delta_t - \big(a_t \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(\mathcal{P}_t \| \mathcal{Q}_t) + b_t \mathbf{1}\big) \big\|_2^2, \tag{11}$$

$$\boldsymbol{\mu}_{\mathrm{new}} = \boldsymbol{\mu}_\theta + \widehat{A}_t \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(\mathcal{P}_t \| \mathcal{Q}_t) + \widehat{B}_t \mathbf{1}. \tag{12}$$

These are empirical regression coefficients: $\widehat{A}_t$ is distinct from the theoretical matrix $\boldsymbol{A}_t$, and the broadcast scalar intercept $\widehat{B}_t\mathbf{1}$ is not the vector residual $\boldsymbol{r}_t$. Figure 6 evaluates this reference-based correction empirically;

Algorithm 1 Calibrate Guidance Scale

Require: LR–HR pairs $\{(\mathbf{c}_i, \mathbf{x}_{0,i})\}_{i=1}^{n}$, model parameters θ, candidate scales $\{\alpha_t\}$
Ensure: Calibrated guidance scales $\{\alpha_t^*\}_{t=1}^{T}$

for $i = 1$ to n **do**
  $\mathbf{x}_T \sim p_\theta(\mathbf{x}_T \mid \mathbf{y}_0)$ with $\mathbf{y}_0 = \mathbf{c}_i$
  for $t = T$ down to 1 **do**
    $(\boldsymbol{\mu}_\theta, \Sigma_{\theta,t}) \leftarrow$ model output at $\mathbf{x}_t$
    $\Delta_{i,t} \leftarrow \boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_{0,i}) - \boldsymbol{\mu}_\theta$
    $\boldsymbol{G}_{i,t} \leftarrow \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(P_{i,t}^{\mathrm{aux}} \| Q_{i,t}^{\mathrm{aux}})$
    $\mathbf{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \Sigma_{\theta,t})$
  end for
end for
for $t = T$ down to 1 **do**
  $\alpha_t^* \leftarrow \arg\min_{\alpha_t} \sum_{i=1}^{n} \big(\|\Delta_{i,t}\| - \alpha_t \|\boldsymbol{G}_{i,t}\|\big)^2$
end for
return $\{\alpha_t^*\}_{t=1}^{T}$

Algorithm 2 Inference with LR Guidance

Require: LR condition $\mathbf{c}$, calibrated scales $\{\alpha_t^*\}$, control scalar $s \in [-1, 1]$, model parameters θ
Ensure: Reconstructed SR latent $\mathbf{x}_0$

$\mathbf{x}_T \sim p_\theta(\mathbf{x}_T \mid \mathbf{y}_0)$ with $\mathbf{y}_0 = \mathbf{c}$
for $t = T$ down to 1 **do**
  $(\boldsymbol{\mu}_\theta, \Sigma_{\theta,t}) \leftarrow$ model output at $\mathbf{x}_t$
  $\boldsymbol{G}_t \leftarrow \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(P_t^{\mathrm{aux}} \| Q_t^{\mathrm{aux}})$
  $\mathbf{x}_{t-1} \sim \mathcal{N}\big(\boldsymbol{\mu}_\theta - s\, \alpha_t^*\, \Sigma_{\theta,t}\, \boldsymbol{G}_t,\ \Sigma_{\theta,t}\big)$
end for
return $\mathbf{x}_0$

Non-reference region. In real-world blind SR, neither Eq. 10 nor Eq. 11 is available. We calibrate on LR–HR pairs and transfer the scales to LR-only inputs, using $\boldsymbol{G}_t = \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(P_t^{\mathrm{aux}} \| Q_t^{\mathrm{aux}})$. The LR-based and HR-based directions need not align, as illustrated in Fig. 3. Moreover, regression over multiple images does not yield stable estimates of both $\widehat{A}_t$ and $\widehat{B}_t$. We therefore omit the fitted intercept and retain the per-step magnitude calibration over n pairs:

$$\alpha_t^* = \arg\min_{\alpha_t} \sum_{i=1}^{n} \Big(\big\|\Delta_{i,t}\big\| - \alpha_t \big\|\boldsymbol{G}_{i,t}\big\| \Big)^2. \tag{13}$$

This empirical objective matches the magnitude of the oracle discrepancy to that of the scaled auxiliary gradient; it does not establish equality of their directions. At inference, the scale is used in Eq. 8:

$$\mathbf{x}_{t-1} \sim \mathcal{N}\big(\boldsymbol{\mu}_\theta - s\alpha_t^* \Sigma_{\theta,t} \boldsymbol{G}_t,\ \Sigma_{\theta,t}\big), \qquad s \in [-1, 1]. \tag{14}$$

Algorithms 1 and 2 summarize the procedures. The same gradient computation is used during calibration and inference. At a zero-variance terminal step, Eq. 14 is interpreted as the deterministic base-model update; no inverse of the reverse covariance is evaluated.

Direction and strength of control. The scalar s controls the signed mean shift $-s\alpha_t^* \Sigma_{\theta,t} \boldsymbol{G}_t$. For nonnegative calibrated scales, changing the sign of s reverses the guidance contribution, giving the fidelity-oriented LR^+ and sharpness-oriented LR^- modes. Under a residual-like approximation to the guidance direction, negative guidance amplifies differences between the current state and the LR condition. If $\mathbf{c} \approx \mathrm{Smooth}(\mathbf{x}_t)$, this resembles an unsharp-mask residual. This is an interpretation rather than an exact equivalence: the residual is added to $\boldsymbol{\mu}_\theta$, and $\mathbf{x}_t - \mathbf{c}$ may also contain noise and reconstruction error (Appendix C.2). The fidelity and sharpness tendencies are evaluated empirically. Varying $|s|$ within $[0, 1]$ provides continuous control from the base model ($s = 0$) to the calibrated operating point.

4 Experiment

Testing environment. We conduct experiments on the DIV2K validation set Agustsson & Timofte (2017), the RealSR dataset Cai et al. (2019), and the Flickr30k dataset Young et al. (2014). DIV2K and RealSR are used for the quantitative comparison in Table 2. Since the images vary in size, we crop them into 512×512 HR patches, and the corresponding 128×128 LR images are used for $\times 4$ upscaling. The guidance scales are calibrated on ImageNet Deng et al. (2009), which is used to train the pretrained model of Yue et al. (2023). All experiments are conducted on an NVIDIA A100 GPU.

Metrics. We employ both full-reference and no-reference metrics: PSNR and SSIM Wang et al. (2004) for reference-based fidelity; LPIPS Zhang et al. (2018) and DISTS Ding et al. (2020) for reference-based perceptual quality; NIQE Zhang et al. (2015), MANIQA Yang et al. (2022), and CLIPIQA Wang et al. (2023a) for no-reference quality, which do not require HR images; and FID Heusel et al. (2017) for the distributional distance between restored and ground-truth images.

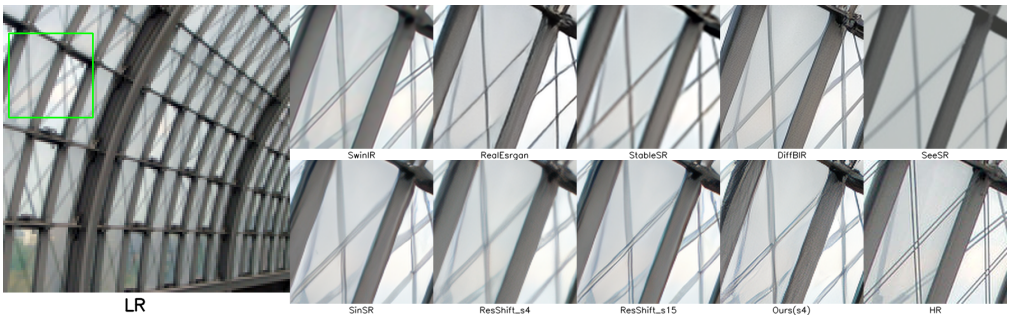

Figure 1: **Non-reference region.** Qualitative comparisons of blind SR. LR guidance produces sharper edges and separates the two adjacent wires that appear merged in the baseline result. Appendix C.2).

Compared methods and backbones. We compare with regression-based (SwinIR Liang et al. (2021)), GAN-based (Real-ESRGAN Wang et al. (2021)), and diffusion-based methods: StableSR Wang et al. (2023b), DiffBIR Lin et al. (2024), PASD Yang et al. (2024), ResShift Yue et al. (2023), SinSR Wang et al. (2024), OSEDiff Wu et al. (2024a), InvSR Yue et al. (2025), and UPSR Zhang et al. (2025). In Table 2, the best and second-best results among diffusion-based methods are highlighted in red and blue, respectively; regression- and GAN-based methods are excluded from the ranking. We apply our plug-in module, without retraining, to the two few-step posterior diffusion models, ResShift (4 steps) and UPSR (5 steps). We do not apply it to diffusion-prior methods built on pretrained latent diffusion models Rombach et al. (2022), such as StableSR and DiffBIR, since their additional networks (e.g., ControlNet Zhang et al. (2023) and CLIP Radford et al. (2021b)) could overlap with the effect of our correction.

Validation of Eqs. 11 and 12. Figure 6 shows the results of correcting SR images with Eqs. 11 and 12 in the reference region, where HR information is simultaneously available (e.g., the overlapping field of view between the wide and tele cameras of a smartphone). Although this is a reference-based rather than blind SR setting, it demonstrates that our derived formulation, which computes the 2nd-order correction term accurately, is valid.

Qualitative analysis. Figure 1 presents blind SR results obtained with Algorithms 1 and 2, where the guidance scales are calibrated on training samples and transferred to the non-reference region. Due to discretization error, the 4-step ResShift inherently produces lower-quality outputs than the 15-step version. In the HR ground truth, all wires follow a dual-wire structure, and among the SR results only SwinIR, ResShift (s15), and ours reconstruct this structure. ***Remarkably, despite operating with only four steps, our method reconstructs the dual-wire structure*** while also sharpening the edges of the original image. Additional experiment result is also available in Figure 7 and Figure 8. Please note that enhanced sharpness in our result also improves acuity of the image.

Quantitative analysis. Table 2 evaluates the two control modes of our module in the blind SR setting, where only LR-based guidance is available. LR^+ denotes the fidelity-oriented mode and LR^- the sharpness-oriented, sign-flipped mode. As discussed in Section 3.2.2, the parameter of the 2nd-order term determines a range of attainable image quality. In Table 2, LR^+ and LR^- correspond to the two ends of the image quality range that our method can produce.

LR^+ consistently improves all reference-based metrics, i.e., PSNR, SSIM, LPIPS, and DISTS, on both datasets and both backbones. The same trend holds on DIV2K. Since the reference-based perceptual metrics (LPIPS and DISTS) also improve, the gain is not merely due to smoothing but reflects the correction of the discretization error. This validates that the LR-based surrogate in Proposition C.2 carries part of the correction direction toward HR although their gain is not large. In exchange, the no-reference metrics degrade. LR^- consistently improves MANIQA, CLIPIQA, and FID. LR^- relatively achieve large enhancement compared to LR^+. In exchange, PSNR and SSIM decrease, which means that sharpness enhancement degrade fidelity metrics. This trend consistently holds for our method based on UPSR Zhang et al. (2025).

Behavior of NIQE. NIQE behaves differently from the other no-reference metrics. LR^+ leaves it nearly unchanged, whereas LR^- consistently degrades it even though MANIQA and CLIPIQA

Table 2: Quantitative comparison with contemporary algorithms for non-overlapped region. 's' : the number of diffusion steps. The best and second-best of each metric are highlighted in red and blue, respectively, with Ours (ResShift) and Ours (UPSR) ranked separately against the baselines.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	CLIPIQA↑	FID↓
RealSR	HR	inf	1.000	0.000	0.000	4.463	0.4890	0.6314	0.0001250
	Bicubic	25.93	0.7571	0.2874	0.2380	6.764	0.4082	0.5084	171.1
	SwinIR	26.40	0.8226	0.1988	0.2128	5.533	0.4822	0.5205	85.79
	Realesrgan	24.82	0.6908	0.2286	0.2112	5.554	0.5155	0.6021	171.3
	StableSR(s200)	23.14	0.7004	0.2512	0.2184	5.993	0.4458	0.6193	165.8
	DiffBIR(s50)	24.04	0.6807	0.2740	0.2217	5.538	0.6186	0.7076	154.2
	PASD(s20)	24.60	0.7128	0.2429	0.2093	4.795	0.5401	0.6250	132.0
	Resshift(s15)	24.93	0.7340	0.1981	0.1875	5.319	0.5574	0.6776	116.8
	Resshift(s4)	25.26	0.7298	0.2075	0.1932	5.930	0.4812	0.6468	118.8
	SinSR(s1)	24.81	0.7321	0.2642	0.2045	5.189	0.5045	0.6397	124.1
	OSEDiff(s1)	23.22	0.6584	0.2793	0.2210	5.464	0.4776	0.6618	172.7
	InvSR(s1)	23.16	0.6600	0.2749	0.2263	4.827	0.5883	0.7048	114.5
	UPSR(s5)	25.61	0.7389	0.1846	0.1731	4.856	0.5311	0.6320	118.3
	Ours(Res, s4,LR^-)	24.53	0.7305	0.2018	0.1984	6.366	0.6493	0.7577	110.5
	Ours(Res, s4,LR^+)	25.69	0.7846	0.1721	0.1871	**5.911**	0.4134	0.5037	142.0
	Ours(upsr, s5,LR^-)	24.58	0.7260	0.1916	0.1840	5.514	0.6284	0.7030	116.0
	Ours(upsr, s5,LR^+)	26.90	0.7657	0.1582	0.1528	**4.851**	0.4087	0.5744	128.3
DIV2k	HR	inf	1.000	0.0000	0.0000	4.879	0.3081	0.4587	0.0000
	Bicubic	25.67	0.7113	0.3438	0.3201	7.239	0.1900	0.3128	228.4
	SwinIR	25.52	0.7563	0.1963	0.2534	6.574	0.2361	0.3691	165.0
	Realesrgan	25.01	0.7292	0.2089	0.2310	4.377	0.3230	0.4401	160.7
	StableSR(s200)	25.08	0.7139	0.2305	0.2412	6.143	0.3341	0.4028	158.0
	DiffBIR(s50)	24.95	0.6645	0.2458	0.2228	5.170	0.5196	0.6948	160.8
	PASD(s20)	25.26	0.7300	0.2395	0.2220	5.840	0.4093	0.5033	166.5
	Resshift(s15)	25.19	0.6823	0.2365	0.2308	6.443	0.3676	0.5650	173.4
	Resshift(s4)	25.19	0.6824	0.2109	0.2364	6.325	0.3243	0.5262	170.4
	SinSR(s1)	25.32	0.7002	0.2268	0.2382	6.067	0.3594	0.5571	219.8
	OSEDiff(s1)	24.79	0.7149	0.2194	0.2138	5.859	0.4278	0.6597	169.5
	InvSR(s1)	24.14	0.6705	0.2415	0.2538	5.981	0.4711	0.6911	185.3
	UPSR(s5)	25.54	0.6908	0.2074	0.2052	5.879	0.3579	0.5341	165.7
	Ours(Res, s4, LR^-)	24.91	0.6823	0.2155	0.2358	6.724	0.4750	0.6874	**164.3**
	Ours(Res, s4, LR^+)	25.43	**0.7013**	0.1995	0.2107	**6.221**	0.2518	0.4086	170.7
	Ours(upsr, s5,LR^-)	24.51	0.6789	0.1947	0.2006	6.628	0.4763	0.6750	**162.1**
	Ours(upsr, s5,LR^+)	26.83	0.7160	0.1608	0.1869	**5.882**	0.2754	0.3046	184.4

improve. MANIQA and CLIPIQA are learned from human opinion scores and tend to reward sharpness and contrast Hu et al. (2026), whereas NIQE measures the deviation of local natural scene statistics from a model fitted to pristine natural images. We conjecture that the halos and amplified high-frequency noise introduced by sharpening Trentacoste et al. (2012) shift these statistics away from those of natural images, so NIQE penalizes the very sharpening that the learned metrics reward.

Bounds of image quality. The two modes applied to a single pretrained model span a wide range of each metric, and intermediate operating points are obtained by reducing $|s|$ in Eq. 14. Rather than sliding along the perception–distortion trade-off through a better sampler, our module exposes the attainable upper and lower bounds of each metric for a given pretrained model. This suggests that when a pretrained model is revised, e.g., by retraining, it should be evaluated by how these ranges shift rather than by the level shift of each metric at a single operating point.

5 Conclusion

In this work, we introduced the first plug-in image quality control framework for posterior diffusion SR. Theoretically, we showed that the discretization error of first-order posterior sampling admits a second-order characterization that can be approximated by the gradient of a KL divergence between latent distributions, and that this correction is equivalent to image-based classifier guidance. We calibrate step-wise guidance scales on paired LR–HR data and apply them at inference without fully retraining the diffusion model. Although exact correction requires HR references, LR-based guidance serves as a practical surrogate in blind SR; its sign selects between fidelity-oriented and sharpness-oriented control, and the calibrated scale bounds artifact-free operation. Experiments on ResShift and UPSR show that our module moves posterior SR models between a fidelity-oriented and a sharpness-oriented operating point, improving reference-based metrics or learned no-reference metrics and FID, respectively, with results competitive with state-of-the-art prior-based methods.

AI USE STATEMENT

We used ChatGPT to assist with grammar check, LaTeX revision, and the review of mathematical formulations, and result interpretation. The authors reviewed and verified the retained AI-assisted content and take responsibility for the final manuscript.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on Single Image Super-Resolution: Dataset and study. In *CVPRW*, July 2017.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *arXiv preprint arXiv:1711.06077*, 2020.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world Single Image Super-Resolution: A new benchmark and a new model. In *ICCV*, pp. 3086–3095, 2019.

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression (adc) for real-world isr. In *CVPR*, 2025.

Defang Chen, Zhenyu Zhou, Jian-Ping Mei, Chunhua Shen, Chun Chen, and Can Wang. A geometric perspective on diffusion models. In *ICLR*, May 2024a.

Defang Chen, Zhenyu Zhou, Can Wang, Chunhua Shen, and Siwei Lyu. On the trajectory regularity of ode-based diffusion sampling. In *ICML*, July 2024b.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Faithful and controllable image super-resolution with diffusion models. In *arXiv preprint arXiv:2411.18824*, 2024c.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pp. 8780–8794, 2021.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity, 2020.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In *arXiv preprint arXiv:2411.18263*, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021.

Matthias Gelbrich. On a formula for the wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147:185–203, 1990.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, volume 2, pp. 2672–2680, 2014.

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. In *NeurIPS*, 2017.

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

M. Hochbruck and A. Ostermann. Explicit exponential Runge-Kutta methods for semilinear parabolic problems, 2005.

Jinfan Hu, Fanghua Yu, Zhiyuan You, Xiang Yin, Hongyu An, Xinqi Lin, Chao Dong, and Jinjin Gu. Position: Evaluation of visual processing should be human-centered, not metric-centered. In *arXiv preprint arXiv:2603.00643*, 2026.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic Single Image Super-Resolution using a generative adversarial network. In *CVPR*, 2017.

Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. SRDiff: Single Image Super-Resolution with diffusion probabilistic models. In *Neurocomputing*, volume 479, pp. 47–59, 2022.

Mingxiao Li, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *ICLR*, May 2024.

Yangming Li and Mihaela van der Schaar. On error propagation of diffusion models. In *ICLR*, May 2024.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV-W*, pp. 1833–1844, 2021.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior, 2024.

Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *CVPR*, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. In *arXiv preprint arXiv:2211.01095*, 2025.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Image restoration with mean-reverting stochastic differential equations. In *arXiv preprint arXiv:2301.11699*, 2023.

E. Kloeden Peter and Platen Eckhard. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Image super-resolution via iterative refinement. In *arXiv preprint arXiv:2104.07636*, 2021.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH Conference*, pp. 1–10, 2022.

Joram Soch. Statproofbook/statproofbook.github.io: The book of statistical proofs. 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

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*, 2021b.

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*, 2021c.

Gilbert Strang. *Computational Science and Engineering*. Wellesley-Cambridge Press, 2007.

Matthew Trentacoste, Ratal Mantiuk, Wolfgang Heidrich, and Florian Dufrot. Unsharp masking, countershading and halos: Enhancements or artifacts? In *Proceedings of Eurographics*, 2012.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023a.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023b.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCV-W*, 2018.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. REAL-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV-W*, pp. 1905–1914, 2021.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. SinSR: Diffusion-based Image Super-Resolution in a single step. In *CVPR*, 2024.

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

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world Image Super-Resolution. In *arXiv preprint arXiv:2406.08177*, 2024a.

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, 2024b.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, pp. 1191–1200, 2022.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic Image Super-Resolution and personalized stylization. In *ECCV*, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, volume 2, pp. 67–78, 2014.

Conghan Yue, Zhengwei Peng, Junlong Ma, Shiyan Du, Pengxu Wei, and Dongyu Zhang. Image restoration through generalized ornstein-uhlenbeck bridge. In *arXiv preprint arXiv:2312.10299*, 2024.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for Image Super-Resolution by residual shifting. In *NeurIPS*, 2023.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *CVPR*, 2025.

Leheng Zhang, Weiyi You, Kexuan Shi, and Shuhang Gu. Uncertainty-guided perturbation for image super-resolution diffusion model. In *arXiv preprint arXiv:2503.18512*, 2025.

Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator, 2015.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

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.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. DPM-Solver-v3: Improved diffusion ode solver with empirical model statistics. In *NeurIPS*, 2023.

CONTENTS

1 Introduction **1**

2 Related Works **3**

3 Plug-in Module for Few-Step Posterior Diffusion Model **3**

3.1 Background for Inverse Problem Solver 3

3.2 Background for High Order Diffusion ODE solvers 4

3.2.1 High order ODE solver for Posterior learning method 4

3.2.2 Image Quality Control by the 2nd-Order Term 4

3.3 Our Proposal 5

4 Experiment **7**

5 Conclusion **9**

A Background for Analytical DPM solver **21**

A.1 Diffusion (Probability Flow) ODEs 21

A.2 Designing Solvers by Data Prediction Model 21

B Background for Posterior Learning Models **22**

C Mathematical Details for Section 3.3 **23**

C.1 High-Order Correction for the Posterior Learning Model (Section 3.2.1) 23

C.2 Image Quality Control by the Guidance Term (Section 3.2.2) 24

C.3 Error Measures and Covariance Conventions for Section 3.3 24

C.4 Propositions for Section 3.3 25

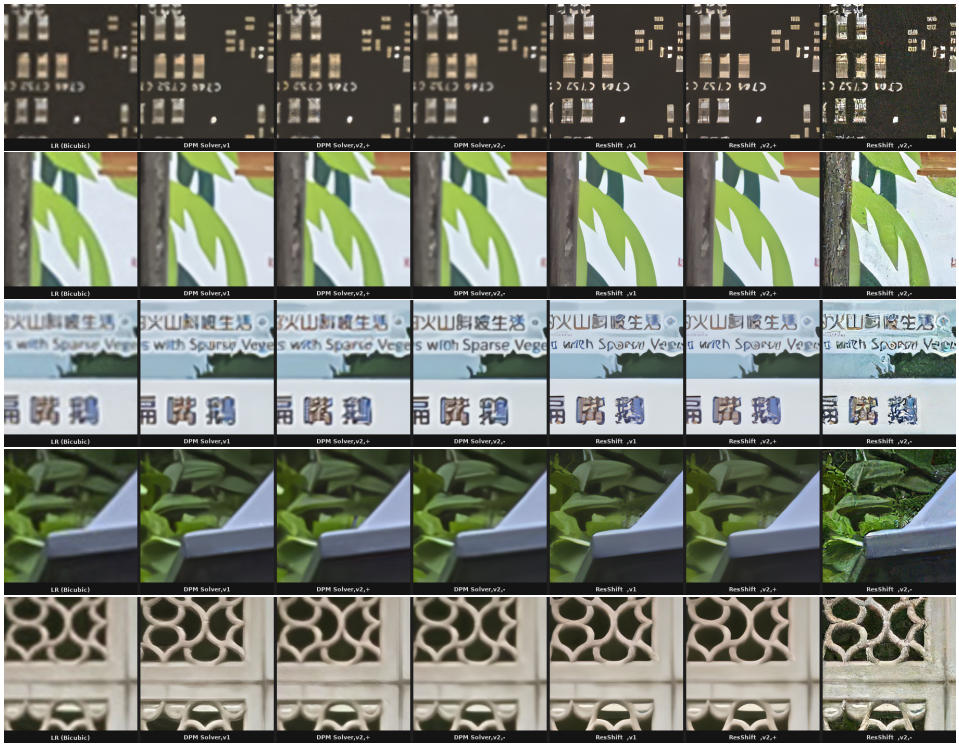

Figure 2: Models from von Platen et al. (2022) based on Stable Diffusion Rombach et al. (2022) uses 15 steps and Modes based on Yue et al. (2023) uses 4 steps. A dataset from the RealSR dataset Cai et al. (2019) is used. For DPM-Solver variants, v1 denotes SDE-DPM-Solver++1, `(v2,+)` denotes SDE-DPM-Solver++2, and `(v2,-)` denotes SDE-DPM-Solver++2 with the sign of the 2nd-order correction term flipped. For ResShift variants, v1 denotes the original model, `(v2,+)` incorporates the 2nd-order correction term introduced in Section 3.2.1, and `(v2,-)` flips its sign. In DPM-Solver, the 2nd-order term has a limited impact on perceptual quality, meaning that without aggressive parameter scaling, the visual difference remains marginal. Moreover, no principled method yet exists for tuning these parameters on a per-image basis under a fixed 15-step budget. In ResShift, applying the 2nd-order term with a positive sign yields only modest quality changes, whereas flipping its sign to negative causes a sharp increase in visible artifacts. These observations highlight the need for a principled methodology to reliably control output quality through 2nd-order correction.

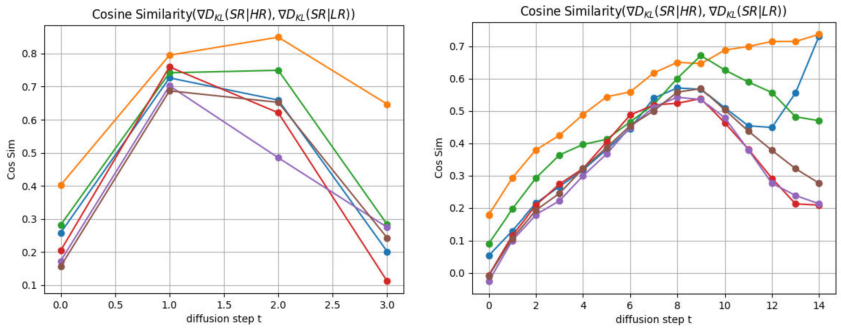

Figure 3: (a) cosine similarity between the trajectory between $\nabla D_{KL}(p_{SR}|p_{HR})$ and $\nabla D_{KL}(p_{SR}|p_{LR})$ for step 4 and (b) for step 15 pretrained models of Yue et al. (2023)

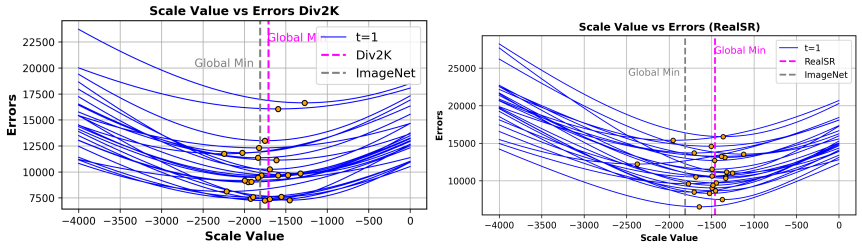

Figure 4: **Numerical error versus calibration scale.** Each curve corresponds to one LR–HR tile pair. Image-specific calibration locates the minimum of each curve accurately, whereas dataset-level calibration (e.g., on ImageNet) averages the minima of many curves into a single global scale that deviates from the image-specific optima.

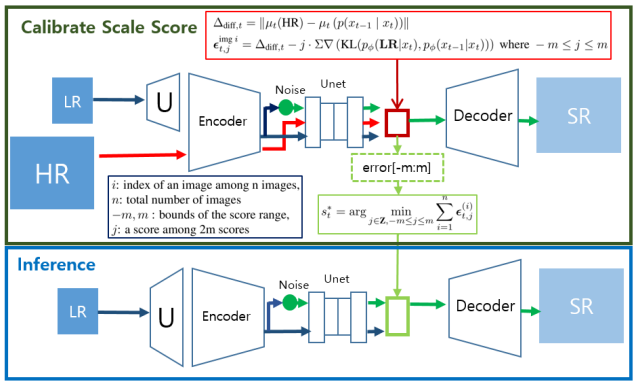

Figure 5: (a) Block diagram for inferring the non-overlapped region using optimized parameters extracted from the overlapped region. A score vector $\mathbf{S}^*$ means the scale score calibrated in Algorithm 1 stage. Calibrated score vector is delivered to Inference pipeline described in Algorithm 2.



Figure 6: Experiment for Reference region. Qualitative comparison on the Flickr30k dataset Young et al. (2014). Ours (s4) and Ours (s15) represent ResShift (s4 and s15) with the proposed restoration in Eq. 12. Please note that most of the collapsed patterns and text that failed to be reconstructed in the baseline have been successfully restored.



Figure 7: Experiment for Non-Reference region. Qualitative comparisons of image quality across various algorithms on the DIV2K dataset Agustsson & Timofte (2017). Our method demonstrates outstanding performance in enhancing the restoration of unstructured objects that lack clear semantic information. Note the significant perceptual quality enhancement compared to the baseline ResShift network.



Figure 8: Non-reference region. Visual comparison against state-of-the-art SR methods. Our method produces more details than other competing ones.

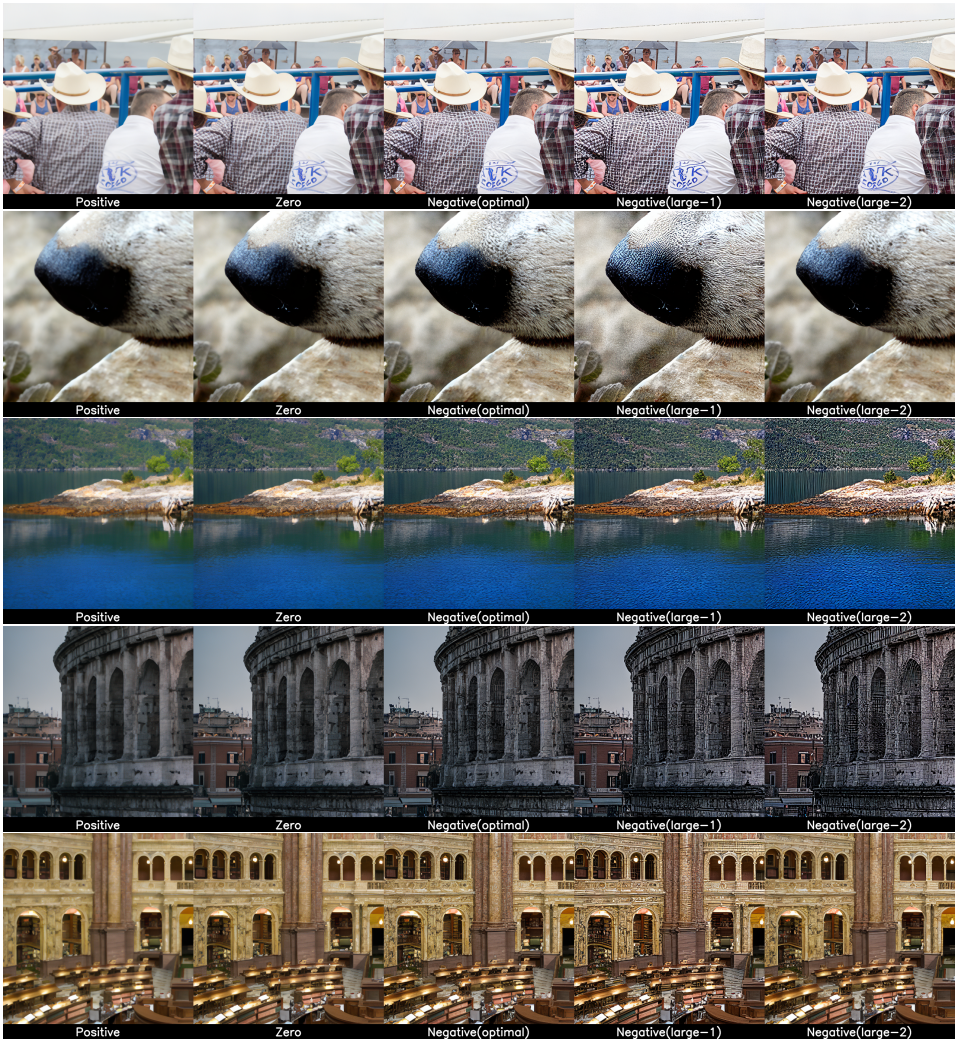

Figure 9: Each parameter set is listed in the table below. The first three columns illustrate the effect of positive, zero, and moderately negative values, while the last two use larger negative magnitudes that lead to distinct artifact patterns. This comparison highlights the sensitivity of the correction module to the sign and absolute scale of the parameters.

Positive(optimal)	Zero	Negative	Negative(large-1)	Negative(large-2)
$(1700, 70, 60)$	$(0, 0, 0)$	$(-1700, -70, -60)$	$(-3000, -120, -120)$	$(-2500, -200, -100)$

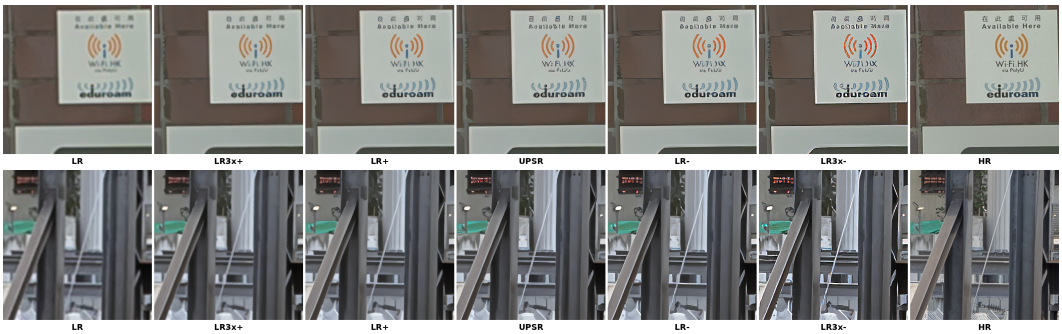

Figure 10: **Effect of the guidance sign and scale.** From left to right: LR input, LR^{+} with $3\times$ the calibrated scale, LR^{+} with the calibrated scale, UPSRZhang et al. (2025) result of the base model without guidance, LR^{-} with the calibrated scale, LR^{-} with $3\times$ the calibrated scale, and HR ground truth. With the calibrated scale, LR^{+} suppresses noisy textures and yields a smoother output, whereas LR^{-} strengthens edges and text contrast. The quality thus moves continuously in opposite directions from the base SR result according to the sign and magnitude of the guidance. When the scale exceeds the calibrated value ($3\times$), LR^{+} over-smooths the image toward the LR input, and LR^{-} over-sharpens it, producing halos, color saturation around the orange arcs, and distorted strokes in the text. This indicates that the calibrated scale serves as a practical bound for reliable quality control.

A BACKGROUND FOR ANALYTICAL DPM SOLVER

For the convenience of the reader, we provides details of section 3.2 Lu et al. (2022; 2025) to build analytical solver of Diffusion Probabilistic Models (DPMs). Any reader with such backgrounds can skip this section. Notations are partially revised.

A.1 DIFFUSION (PROBABILITY FLOW) ODES

Song et al. (2021c) proved that the reverse process ODE is

$$\frac{d\mathbf{x}_t}{dt} = f(t)\mathbf{x}_t - \frac{1}{2}g^2(t)\nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t), \quad \mathbf{x}_T \sim q_T(\mathbf{x}_T), \tag{15}$$

where the marginal distribution of $\mathbf{x}_t$ is also $q_t(\mathbf{x}_t)$. By replacing the score function with the noise prediction model, Song et al. (2021c) defined the following parameterized ODE (*diffusion ODE*):

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{h}_\theta(\mathbf{x}_t, t) \coloneqq f(t)\mathbf{x}_t + \frac{g^2(t)}{2\sigma_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t), \quad \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2\mathbf{I}). \tag{16}$$

From Lu et al. (2022), Exact solution of Eq. 16 is written as

$$\mathbf{x}_t = \frac{\alpha_t}{\alpha_s}\mathbf{x}_s - \alpha_t \int_s^t \left(\frac{\mathrm{d}\lambda_\tau}{\mathrm{d}\tau}\right) \frac{\sigma_\tau}{\alpha_\tau}\boldsymbol{\epsilon}_\theta(\mathbf{x}_\tau, \tau)\mathrm{d}\tau = \frac{\alpha_t}{\alpha_s}\mathbf{x}_s - \alpha_t \int_{\lambda_s}^{\lambda_t} e^{-\lambda}\hat{\boldsymbol{\epsilon}}_\theta(\hat{\mathbf{x}}_\lambda, \lambda)\mathrm{d}\lambda. \tag{17}$$

where an initial value $\mathbf{x}_s$ at time $s > 0$, it is the solution $\mathbf{x}_t$ at time $t \in [0, s]$ of diffusion ODEs.

A.2 DESIGNING SOLVERS BY DATA PREDICTION MODEL

We parameterize the denoising model in terms of a *noise prediction model* $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t; t)$:

$$\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_t) = (\mathbf{x}_t - \sigma_t\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t))/\alpha_t, \tag{18}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t)$ is parameterized as a neural network.

The equivalent diffusion ODE w.r.t. the data prediction model $\mathbf{x}_\theta$ is

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \left(f(t) + \frac{g^2(t)}{2\sigma_t^2}\right)\mathbf{x}_t - \frac{\alpha_t g^2(t)}{2\sigma_t^2}\mathbf{x}_\theta(\mathbf{x}_t, t), \quad \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2\mathbf{I}), \tag{19}$$

where the coefficients $f(t) = \frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}$, $g^2(t) = \frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t} - 2\frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}\sigma_t^2$ (Kingma et al., 2021).

For solving the diffusion ODE w.r.t. $\mathbf{x}_\theta$ in Eq. 19, a simplified formulation of the exact solution of diffusion ODEs w.r.t. $\mathbf{x}_\theta$ is written as below.

Exact solution of diffusion ODEs of $\mathbf{x}_\theta$ Given an initial value $\mathbf{x}_s$ at time $s > 0$, the solution $\mathbf{x}_t$ at time $t \in [0, s]$ of diffusion ODEs in Eq. 19 is:

$$\mathbf{x}_t = \frac{\sigma_t}{\sigma_s}\mathbf{x}_s + \sigma_t \int_{\lambda_s}^{\lambda_t} e^{\lambda}\hat{\mathbf{x}}_\theta(\hat{\mathbf{x}}_\lambda, \lambda)\mathrm{d}\lambda. \tag{20}$$

The aim of this solver is to approximate the exact solution at time t_i. Denote $\mathbf{x}_\theta^{(n)}(\lambda) \coloneqq \frac{\mathrm{d}^n\mathbf{x}_\theta(\mathbf{x}_\lambda, \lambda)}{\mathrm{d}\lambda^n}$ as the n-th order total derivatives of $\mathbf{x}_\theta$ w.r.t. logSNR λ. For $k \geq 1$, taking the $(k-1)$-th Taylor expansion at $\lambda_{t_{i-1}}$ for $\mathbf{x}_\theta$ w.r.t. $\lambda \in [\lambda_{t_{i-1}}, \lambda_{t_i}]$ and substitute it into Eq. 20 with $s = t_{i-1}$ and $t = t_i$, we have

$$\tilde{\mathbf{x}}_{t_i} = \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}}\tilde{\mathbf{x}}_{t_{i-1}} + \sigma_{t_i}\sum_{n=0}^{k-1} \underbrace{\mathbf{x}_\theta^{(n)}(\hat{\mathbf{x}}_{\lambda_{t_{i-1}}}, \lambda_{t_{i-1}})}_{\text{estimated}} \underbrace{\int_{\lambda_{t_{i-1}}}^{\lambda_{t_i}} e^{\lambda}\frac{(\lambda - \lambda_{t_{i-1}})^n}{n!}\mathrm{d}\lambda}_{\text{analytically computed}} + \underbrace{\mathcal{O}(h_i^{k+1})}_{\text{omitted}}, \tag{21}$$

The simplified version of Eq. 21 is written as follows:

SDE-DPM-Solver++1 Let $\mathbf{z}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. By assuming $\mathbf{x}_\theta(\mathbf{x}_\lambda, \lambda) \approx \mathbf{x}_\theta(\mathbf{x}_s, s)$, we have

$$\mathbf{x}_t = \frac{\sigma_t}{\sigma_s}e^{-h}\mathbf{x}_s + \alpha_t(1 - e^{-2h})\mathbf{x}_\theta(\mathbf{x}_s, s) + \sigma_t\sqrt{1 - e^{-2h}}\mathbf{z}_s. \tag{22}$$

SDE-DPM-Solver++2 Let $\mathbf{z}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Assume we have a previous solution $\mathbf{x}_r$ with its model output $\mathbf{x}_\theta(\mathbf{x}_r, r)$ at time $r < t$. Denote $r_1 = \frac{\lambda_r - \lambda_s}{h}$. By assuming $\mathbf{x}_\theta(\mathbf{x}_\lambda, \lambda) \approx \mathbf{x}_\theta(\mathbf{x}_s, s) + \frac{\lambda - \lambda_s}{r_1 h}(\mathbf{x}_\theta(\mathbf{x}_r, r) - \mathbf{x}_\theta(\mathbf{x}_s, s))$, we have

$$\mathbf{x}_t = \frac{\sigma_t}{\sigma_s} e^{-h} \mathbf{x}_s + \alpha_t (1 - e^{-2h}) \mathbf{x}_\theta(\mathbf{x}_s, s) + \frac{\alpha_t (1 - e^{-2h})}{2} \left(\frac{\mathbf{x}_\theta(\mathbf{x}_r, r) - \mathbf{x}_\theta(\mathbf{x}_s, s)}{r_1} \right) + \sigma_t \sqrt{1 - e^{-2h}} \mathbf{z}_s. \quad (23)$$

B Background for Posterior Learning Models

Posterior learning models for image restoration are commonly built upon the generalized Ornstein–Uhlenbeck (GOU) process, a mean-reverting SDE that drives the state $\boldsymbol{x}$ toward a target mean $\boldsymbol{\mu}$ (typically the degraded observation):

$$\mathrm{d}\boldsymbol{x} = \theta_t(\boldsymbol{\mu} - \boldsymbol{x})\,\mathrm{d}t + \sigma_t\,\mathrm{d}\boldsymbol{w}, \quad (24)$$

where θ_t controls the speed of mean reversion, σ_t is the diffusion coefficient, and $\boldsymbol{w}$ is a standard Wiener process. The corresponding reverse-time SDE is given by

$$\mathrm{d}\boldsymbol{x} = \left[\theta_t(\boldsymbol{\mu} - \boldsymbol{x}) - \sigma_t^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right] \mathrm{d}t + \sigma_t\,\mathrm{d}\bar{\boldsymbol{w}}, \quad (25)$$

where $\bar{\boldsymbol{w}}$ is a reverse-time Wiener process and $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ is the score function. By constraining the coefficients as $\sigma_t^2/\theta_t = 2\lambda^2$ with a constant stationary variance λ^2, the marginal distribution becomes analytically tractable, *i.e.*, $q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\big(\boldsymbol{\mu} + (\boldsymbol{x}_0 - \boldsymbol{\mu})e^{-\bar{\theta}_t},\ \lambda^2(1 - e^{-2\bar{\theta}_t})\boldsymbol{I}\big)$ with $\bar{\theta}_t = \int_0^t \theta_z\,\mathrm{d}z$ Luo et al. (2023). Yue et al. (2024) further applies Doob's h-transform to remove the terminal variance, yielding a bridge that maps exactly between the HQ and LQ images. However, these methods target general image restoration rather than super-resolution specifically, and they typically require around 100 sampling steps, which is prohibitive in practice. In contrast, Yue et al. (2023) and Zhang et al. (2025) are tailored to super-resolution and operate with five or fewer sampling steps. Since Zhang et al. (2025) is built upon the formulation of Yue et al. (2023), we briefly summarize Yue et al. (2023) below for the convenience of the reader, with the notation partially revised. Readers familiar with this background may skip this section.

Forward Process. Rather than simulating a continuous-time SDE, Yue et al. (2023) constructs a discrete Markov chain that gradually shifts the HR image toward the LR image by transferring their residual. Let $\boldsymbol{e}_0 = \boldsymbol{y}_0 - \boldsymbol{x}_0$ denote the residual between the LR image $\boldsymbol{y}_0$ and the HR image $\boldsymbol{x}_0$. A shifting sequence $\{\eta_t\}_{t=1}^T$ is introduced, which monotonically increases with the timestep t and satisfies $\eta_1 \to 0$ and $\eta_T \to 1$. The transition distribution is then defined as

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}_{t-1} + \alpha_t \boldsymbol{e}_0, \kappa^2 \alpha_t \boldsymbol{I}), \quad t = 1, 2, \cdots, T, \quad (26)$$

where $\alpha_t = \eta_t - \eta_{t-1}$ for $t > 1$, $\alpha_1 = \eta_1$, and κ is a hyper-parameter controlling the noise variance. The marginal distribution at any timestep t is analytically tractable:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}_0 + \eta_t \boldsymbol{e}_0, \kappa^2 \eta_t \boldsymbol{I}), \quad t = 1, 2, \cdots, T. \quad (27)$$

Hence, $\boldsymbol{x}_t$ starts near the HR image at $t = 1$ and approaches a noisy version of the LR image at $t = T$.

Reverse Process. The reverse process aims to estimate the posterior distribution $p(\boldsymbol{x}_0|\boldsymbol{y}_0)$ via

$$p(\boldsymbol{x}_0|\boldsymbol{y}_0) = \int p(\boldsymbol{x}_T|\boldsymbol{y}_0) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y}_0)\,\mathrm{d}\boldsymbol{x}_{1:T}, \quad (28)$$

where $p(\boldsymbol{x}_T|\boldsymbol{y}_0) \approx \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{y}_0, \kappa^2 \boldsymbol{I})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y}_0)$ is the reverse transition kernel with learnable parameters $\boldsymbol{\theta}$. Modeling this kernel as a Gaussian, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t))$, the parameters $\boldsymbol{\theta}$ are optimized by

$$\min_{\boldsymbol{\theta}} \sum_t D_{\mathrm{KL}}\left[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}_0) \,\|\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y}_0)\right]. \quad (29)$$

The target distribution in Eq. 29 has the closed form

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \frac{\eta_{t-1}}{\eta_t}\boldsymbol{x}_t + \frac{\alpha_t}{\eta_t}\boldsymbol{x}_0,\ \kappa^2 \frac{\eta_{t-1}}{\eta_t}\alpha_t \boldsymbol{I}\right). \quad (30)$$

Since this variance does not depend on $\boldsymbol{x}_t$ or $\boldsymbol{y}_0$, the model variance is fixed to it, *i.e.*, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t) = \kappa^2 \frac{\eta_{t-1}}{\eta_t}\alpha_t \boldsymbol{I}$. The mean is reparameterized as

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t) = \frac{\eta_{t-1}}{\eta_t}\boldsymbol{x}_t + \frac{\alpha_t}{\eta_t} f_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t), \quad (31)$$

where $f_{\boldsymbol{\theta}}$ is a deep neural network that predicts $\boldsymbol{x}_0$. Under this parameterization, the KL objective in Eq. 29 reduces to the weighted regression loss

$$\min_{\boldsymbol{\theta}} \sum\nolimits_t w_t \|f_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t) - \boldsymbol{x}_0\|_2^2, \quad (32)$$

where w_t is a timestep-dependent weight.

C Mathematical Details for Section 3.3

C.1 High-Order Correction for the Posterior Learning Model (Section 3.2.1)

In this section, we construct a multistep correction for the ResShift update by matching its coefficients to those of SDE-DPM-Solver++ and reusing consecutive data predictions. The derivation is specific to the ResShift kernel and does not assume spatial isotropy for all posterior SR models. Following A and B, the ResShift reverse step, whose posterior mean is reparameterized by the network $f_{\boldsymbol{\theta}}$, reads

$$\mathbf{x}_{t-1} = \frac{\eta_{t-1}}{\eta_t}\mathbf{x}_t + \frac{\alpha_t}{\eta_t} f_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}_0, t) + \kappa\sqrt{\frac{\eta_{t-1}\,\alpha_t}{\eta_t}}\,\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{33}$$

where $\alpha_t = \eta_t - \eta_{t-1}$ and κ is the noise-variance hyperparameter. The variance of the stochastic term, $\kappa^2\eta_{t-1}\alpha_t/\eta_t$, matches the closed-form posterior variance of ResShift (Eq. 30). The first-order SDE-DPM-Solver++ update from time s to time t reads

$$\mathbf{x}_t = \frac{\sigma_t^{\mathrm{dpm}}}{\sigma_s^{\mathrm{dpm}}} e^{-h}\,\mathbf{x}_s + \alpha_t^{\mathrm{dpm}}\big(1 - e^{-2h}\big)\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_s, s) + \sigma_t^{\mathrm{dpm}}\sqrt{1 - e^{-2h}}\,\mathbf{z}_s, \tag{34}$$

where $h = \lambda_t - \lambda_s$, $\lambda_t = \log(\alpha_t^{\mathrm{dpm}}/\sigma_t^{\mathrm{dpm}})$ is the log-SNR, and $\mathbf{x}_{\boldsymbol{\theta}}$ is the data-prediction model. Below, we map the DPM denoising direction $s \to t$ to the ResShift direction $t \to t-1$.

Parameter identification. We use the substitution $\alpha_t^{\mathrm{dpm}} = 1$ and $\sigma_t^{\mathrm{dpm}} = \kappa\sqrt{\eta_t}$, under which the log-SNR becomes

$$\lambda_t = \log\frac{1}{\kappa\sqrt{\eta_t}} = -\log\kappa - \tfrac{1}{2}\log\eta_t. \tag{35}$$

In the ResShift direction $t \to t-1$ (i.e., $s \leftrightarrow t$ and $t \leftrightarrow t-1$ in the DPM notation),

$$h = \lambda_{t-1} - \lambda_t = \Big(-\log\kappa - \tfrac{1}{2}\log\eta_{t-1}\Big) - \Big(-\log\kappa - \tfrac{1}{2}\log\eta_t\Big) = \tfrac{1}{2}\log\frac{\eta_t}{\eta_{t-1}} > 0, \tag{36}$$

since $\eta_t > \eta_{t-1}$. From Eq. 36,

$$e^{-h} = \sqrt{\frac{\eta_{t-1}}{\eta_t}}, \qquad e^{-2h} = \frac{\eta_{t-1}}{\eta_t}. \tag{37}$$

We verify the correspondence coefficient by coefficient:

$$\frac{\sigma_{t-1}^{\mathrm{dpm}}}{\sigma_t^{\mathrm{dpm}}}\,e^{-h} = \frac{\kappa\sqrt{\eta_{t-1}}}{\kappa\sqrt{\eta_t}} \cdot \sqrt{\frac{\eta_{t-1}}{\eta_t}} = \frac{\eta_{t-1}}{\eta_t}, \tag{38}$$

$$\alpha_{t-1}^{\mathrm{dpm}}\big(1 - e^{-2h}\big) = 1 - \frac{\eta_{t-1}}{\eta_t} = \frac{\alpha_t}{\eta_t}, \tag{39}$$

$$\sigma_{t-1}^{\mathrm{dpm}}\sqrt{1 - e^{-2h}} = \kappa\sqrt{\eta_{t-1}} \cdot \sqrt{\frac{\alpha_t}{\eta_t}} = \kappa\sqrt{\frac{\eta_{t-1}\,\alpha_t}{\eta_t}}. \tag{40}$$

Hence, the two first-order updates have matching algebraic coefficients under this substitution. This correspondence motivates the multistep correction below; by itself, it does not establish equivalence of the underlying continuous-time processes or a convergence-order guarantee for the pretrained ResShift model.

Extension to the 2nd-order correction. The 2nd-order multistep variant (SDE-DPM-Solver++(2M)) linearly extrapolates the data prediction in λ using the model output $\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_r, r)$ at the preceding time r:

$$\mathbf{x}_t = \tfrac{\sigma_t^{\mathrm{dpm}}}{\sigma_s^{\mathrm{dpm}}}e^{-h}\mathbf{x}_s + \alpha_t^{\mathrm{dpm}}\big(1 - e^{-2h}\big)\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_s, s) + \tfrac{\alpha_t^{\mathrm{dpm}}\left(1-e^{-2h}\right)}{2r_1}\big(\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_s, s) - \mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_r, r)\big) + \sigma_t^{\mathrm{dpm}}\sqrt{1 - e^{-2h}}\,\mathbf{z}_s, \tag{41}$$

where $r_1 = (\lambda_s - \lambda_r)/h > 0$ is the ratio of the previous step size to the current one. Substituting $s \leftrightarrow t$, $t \leftrightarrow t-1$, and $r \leftrightarrow t+1$, we obtain the following **multistep correction**:

$$\boxed{\mathbf{x}_{t-1} = \frac{\eta_{t-1}}{\eta_t}\mathbf{x}_t + \frac{\alpha_t}{\eta_t} f_{\boldsymbol{\theta}}^{(t)} + \frac{\alpha_t}{2\eta_t\, r_1}\big(f_{\boldsymbol{\theta}}^{(t)} - f_{\boldsymbol{\theta}}^{(t+1)}\big) + \kappa\sqrt{\frac{\eta_{t-1}\,\alpha_t}{\eta_t}}\,\mathbf{z},} \tag{42}$$

where $f_{\boldsymbol{\theta}}^{(t)} \coloneqq f_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}_0, t)$, $f_{\boldsymbol{\theta}}^{(t+1)} \coloneqq f_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}, \mathbf{y}_0, t+1)$, and

$$r_1 = \frac{\lambda_t - \lambda_{t+1}}{h} = \frac{\log(\eta_{t+1}/\eta_t)}{\log(\eta_t/\eta_{t-1})} > 0. \tag{43}$$

Eq. 42 is the ResShift counterpart of SDE-DPM-Solver++(2M), which introduces a solver-inspired correction by recycling the previously computed network output $f_{\boldsymbol{\theta}}^{(t+1)}$. The correction term is proportional to the difference between the current and previous data predictions, which is the term analyzed in Section 3.2.2.

The multistep formula is used only when a previous model output is available and $\eta_{t-1} > 0$. The first sampling step uses the first-order update. A deterministic terminal transition is handled separately rather than evaluating the log-SNR ratio at $\eta_0 = 0$.

C.2 Image Quality Control by the Guidance Term (Section 3.2.2)

Connection to classical Laplacian sharpening. For an image $\mathbf{I}$, Laplacian sharpening is written as

$$\mathbf{I}_{\text{sharp}} = \mathbf{I} - c\nabla^2\mathbf{I}, \qquad c > 0. \tag{44}$$

For a low-pass filter whose residual approximates a scaled Laplacian, this has the unsharp-mask form

$$\mathbf{I}_{\text{sharp}} \approx \mathbf{I} + c'\big(\mathbf{I} - \text{Smooth}(\mathbf{I})\big), \qquad c' > 0. \tag{45}$$

The high-frequency residual is amplified. Below, we use this as an interpretation of image guidance, rather than identifying image guidance with the consecutive-prediction correction in Eq. 42.

Guidance gradient. Let $\mathbf{c}$ denote the LR-derived condition represented in the same coordinate space as $\mathbf{x}_t$. At each guided timestep, define

$$P_t^{\text{aux}} = \mathcal{N}(\mathbf{x}_t, \mathbf{\Gamma}_t), \qquad Q_t^{\text{aux}} = \mathcal{N}(\mathbf{c}, \mathbf{\Gamma}_t), \tag{46}$$

where $\mathbf{\Gamma}_t$ is positive definite and independent of $\mathbf{x}_t$ for fixed observation $\mathbf{y}_0$. It may depend on the timestep and the observation and need not be spatially isotropic. These auxiliary Gaussians define an energy; they are not identified with the actual marginal distribution of the diffusion state.

$$E_t(\mathbf{x}_t, \mathbf{c}) \coloneqq D_{\text{KL}}(P_t^{\text{aux}} \| Q_t^{\text{aux}}) = \frac{1}{2}(\mathbf{x}_t - \mathbf{c})^\top \mathbf{\Gamma}_t^{-1}(\mathbf{x}_t - \mathbf{c}). \tag{47}$$

Lemma 1 in Appendix C.4 gives

$$\boldsymbol{G}_t = \nabla_{\mathbf{x}_t} E_t(\mathbf{x}_t, \mathbf{c}) = \mathbf{\Gamma}_t^{-1}(\mathbf{x}_t - \mathbf{c}). \tag{48}$$

Writing the reverse-transition covariance as $\Sigma_{\theta,t}$, the guided mean in Eq. 14 is

$$\boldsymbol{\mu}_{\text{new}} = \boldsymbol{\mu}_\theta - s\alpha_t^* \Sigma_{\theta,t} \mathbf{\Gamma}_t^{-1}(\mathbf{x}_t - \mathbf{c}). \tag{49}$$

Matched-covariance guidance. In practice, $\mathbf{\Gamma}_t^{-1}$ does not need to be measured separately. It is measured together within α_t^*, the expression simplifies to

$$\boldsymbol{\mu}_{\text{new}} = \boldsymbol{\mu}_\theta - s\alpha_t^*(\mathbf{x}_t - \mathbf{c}). \tag{50}$$

This cancellation requires matching covariances, not spatial isotropy. The following interpretation concerns this case and assumes $\alpha_t^* \geq 0$.

Fidelity-oriented mode (LR$^+$, $s > 0$). With $\beta = s\alpha_t^* > 0$, the update is

$$\boldsymbol{\mu}_{\text{new}} = \boldsymbol{\mu}_\theta + \beta(\mathbf{c} - \mathbf{x}_t). \tag{51}$$

The added vector points from the current state toward the LR-derived condition. When $\mathbf{x}_t - \mathbf{c}$ is dominated by excess detail or noise, this update can suppress these components. We call this setting fidelity-oriented because it improves the full-reference metrics in our experiments; this is not a general guarantee of PSNR or SSIM improvement.

Sharpness-oriented mode (LR$^-$, $s < 0$). With $\gamma = |s|\alpha_t^* > 0$, the update is

$$\boldsymbol{\mu}_{\text{new}} = \boldsymbol{\mu}_\theta + \gamma(\mathbf{x}_t - \mathbf{c}). \tag{52}$$

If $\mathbf{c}$ approximately represents a low-pass version of the current state in the relevant coordinate space, then

$$\mathbf{x}_t - \mathbf{c} \approx \mathbf{x}_t - \text{Smooth}(\mathbf{x}_t), \tag{53}$$

and the added residual has a high-pass character analogous to Eq. 45. This is a sharpening interpretation rather than an exact equivalence: the residual is computed from $\mathbf{x}_t$ but added to $\boldsymbol{\mu}_\theta$. Moreover, it can contain diffusion noise, low-frequency discrepancies, and reconstruction artifacts in addition to useful details. For latent-space models, the decoded effect also depends on the decoder. Sign-flipped guidance can therefore enhance edge contrast, while excessive amplification can introduce halos and noise amplification Trentacoste et al. (2012); Hu et al. (2026).

C.3 Error Measures and Covariance Conventions for Section 3.3

Definition of error along the diffusion path. Following the modular and cumulative comparison in Li & van der Schaar (2024), fix the observation $\mathbf{y}_0$ and consider model and reference reverse kernels. Their modular discrepancy is

$$\mathbf{E}_t^{\text{modular}} = \mathbb{E}_{\mathbf{x}_t \sim \nu_t}\left[D_{\text{KL}}\left(p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0) \,\|\, q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0)\right)\right], \tag{54}$$

where ν_t is the explicitly chosen state distribution used for the comparison. The marginal discrepancy is

$$\mathbf{E}_t^{\text{cumu}} = D_{\text{KL}}\left(p_\theta(\mathbf{x}_{t-1} \mid \mathbf{y}_0) \,\|\, q(\mathbf{x}_{t-1} \mid \mathbf{y}_0)\right). \tag{55}$$

Here the reference reverse kernel is induced by the reference forward process, and the model marginal is obtained by sequential sampling. These definitions do not identify every model-reference discrepancy with numerical discretization error. In the proposition below, the reference is instead explicitly conditioned on an HR image $\mathbf{x}_0$, which defines a sample-specific oracle comparison.

Covariance conventions. In the Gaussian KL calculations, the observation $\mathbf{y}_0$ is fixed. At each nondegenerate timestep, the covariance matrices used below are positive definite and independent of $\mathbf{x}_t$. They may vary with time and spatial location. Define

$$v_t = \kappa^2 \frac{\eta_{t-1}}{\eta_t} \alpha_t. \tag{56}$$

The reverse-transition covariance of ResShift is

$$\mathbf{\Sigma}_t^{\mathrm{ResShift}} = v_t \mathbf{I}. \tag{57}$$

For UPSR, uncertainty-guided noise weighting gives

$$\mathbf{\Sigma}_t^{\mathrm{UPSR}}(\mathbf{y}_0) = v_t \mathbf{W}_u(\mathbf{y}_0)^2, \tag{58}$$

where $\mathbf{W}_u(\mathbf{y}_0)$ is a diagonal matrix of spatially varying noise weights Zhang et al. (2025). Its covariance is generally not isotropic. Both models use the affine reverse-mean form

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0) = \mathcal{N}\left(\frac{\eta_{t-1}}{\eta_t} \mathbf{x}_t + \frac{\alpha_t}{\eta_t} f_\theta(\mathbf{x}_t, \mathbf{y}_0, t), \mathbf{\Sigma}_t(\mathbf{y}_0) \right). \tag{59}$$

For UPSR, the notation $f_\theta(\mathbf{x}_t, \mathbf{y}_0, t)$ also includes any deterministic conditioning features computed from $\mathbf{y}_0$. Positive definiteness requires $v_t > 0$ and, for UPSR, strictly positive diagonal weights. A deterministic terminal transition must be treated separately; inverse-covariance formulas do not apply to a zero-variance kernel.

C.4 Propositions for Section 3.3

Lemma 1 (Gaussian KL gradient). Let $\mathcal{P} = \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and $\mathcal{Q} = \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ on $\mathbb{R}^d$, with positive-definite covariances held fixed when differentiating with respect to the means. Then

$$\nabla_{\boldsymbol{\mu}_1} D_{\mathrm{KL}}(\mathcal{P} \| \mathcal{Q}) = \mathbf{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{60}$$

Proof. The Gaussian KL divergence Soch (2024) is

$$D_{\mathrm{KL}}(\mathcal{P} \| \mathcal{Q}) = \frac{1}{2}\Big[\log \frac{\det \mathbf{\Sigma}_2}{\det \mathbf{\Sigma}_1} - d + \mathrm{tr}(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \mathbf{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \Big]. \tag{61}$$

Differentiating the last term proves Eq. 60. If the covariances equal $\mathbf{\Sigma}$, this reduces to

$$D_{\mathrm{KL}}(\mathcal{P} \| \mathcal{Q}) = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{62}$$

□

Definition 1 (Per-step posterior-mean discrepancy). At timestep t, let $\boldsymbol{\mu}_\theta$ be the learned reverse mean and $\boldsymbol{\mu}_q$ the HR-conditioned reference mean from $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0, \mathbf{y}_0)$ (Eq. 30):

$$\begin{aligned} \boldsymbol{\mu}_\theta(\mathbf{x}_t) &= \frac{\eta_{t-1}}{\eta_t}\mathbf{x}_t + \frac{\alpha_t}{\eta_t} f_\theta(\mathbf{x}_t, \mathbf{y}_0, t), \\ \boldsymbol{\mu}_q(\mathbf{x}_t) &= \frac{\eta_{t-1}}{\eta_t}\mathbf{x}_t + \frac{\alpha_t}{\eta_t}\mathbf{x}_0. \end{aligned} \tag{63}$$

Define the discrepancy vector

$$\boldsymbol{\delta}_t \coloneqq \boldsymbol{\mu}_q - \boldsymbol{\mu}_\theta = \frac{\alpha_t}{\eta_t}\big(\mathbf{x}_0 - f_\theta(\mathbf{x}_t, \mathbf{y}_0, t)\big) \in \mathbb{R}^d. \tag{64}$$

This is an oracle posterior-mean discrepancy; the definition alone does not isolate numerical truncation error from prediction error.

Proposition C.1 (Posterior-mean correction via a KL gradient). *Fix $\mathbf{y}_0$ and $\mathbf{x}_0$. At a nondegenerate timestep, let*

$$\mathcal{P}_t = \mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{\Sigma}_t), \qquad \mathcal{Q}_t = \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{\Sigma}_t), \tag{65}$$

where $\mathbf{\Sigma}_t$ is positive definite and independent of $\mathbf{x}_t$ for fixed $\mathbf{y}_0$. Define

$$a_t = \frac{\eta_{t-1}}{\eta_t} > 0, \qquad \boldsymbol{J}_t = \frac{\partial(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)}{\partial \mathbf{x}_t}. \tag{66}$$

Then

$$\nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(\mathcal{P}_t \| \mathcal{Q}_t) = -\boldsymbol{J}_t^\top \mathbf{\Sigma}_t^{-1} \boldsymbol{\delta}_t. \tag{67}$$

Writing $\boldsymbol{J}_t = -a_t \mathbf{I} + \boldsymbol{R}_t$ gives

$$\boldsymbol{\delta}_t = \boldsymbol{A}_t \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(\mathcal{P}_t \| \mathcal{Q}_t) + \boldsymbol{r}_t, \qquad \boldsymbol{A}_t = \frac{1}{a_t}\mathbf{\Sigma}_t, \tag{68}$$

where

$$\boldsymbol{r}_t = \frac{1}{a_t}\mathbf{\Sigma}_t \boldsymbol{R}_t^\top \mathbf{\Sigma}_t^{-1} \boldsymbol{\delta}_t. \tag{69}$$

When this residual is negligible, the corrected mean satisfies

$$\tilde{\boldsymbol{\mu}}_\theta = \boldsymbol{\mu}_\theta + \boldsymbol{\delta}_t \approx \boldsymbol{\mu}_\theta + \frac{\eta_t}{\eta_{t-1}} \mathbf{\Sigma}_t \nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(\mathcal{P}_t \| \mathcal{Q}_t). \tag{70}$$

Proof. The common covariance is independent of $\mathbf{x}_t$, so Lemma 1 yields

$$D_{\mathrm{KL}}(\mathcal{P}_t\|\mathcal{Q}_t) = \frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)^\top \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q). \tag{71}$$

The shared state terms in Eq. 63 cancel, giving

$$\boldsymbol{J}_t = \frac{\alpha_t}{\eta_t}\frac{\partial f_\theta(\mathbf{x}_t, \mathbf{y}_0, t)}{\partial \mathbf{x}_t}. \tag{72}$$

The chain rule proves Eq. 67. Substituting $\boldsymbol{J}_t = -a_t\mathbf{I} + \boldsymbol{R}_t$ gives

$$\nabla_{\mathbf{x}_t} D_{\mathrm{KL}}(\mathcal{P}_t\|\mathcal{Q}_t) = a_t\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\delta}_t - \boldsymbol{R}_t^\top\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\delta}_t. \tag{73}$$

Multiplying by $\boldsymbol{\Sigma}_t/a_t$ and rearranging proves Eqs. 68 and 69. □

Scope of the approximation. The approximation $\boldsymbol{J}_t \approx -a_t\mathbf{I}$ corresponds to

$$\frac{\partial f_\theta}{\partial \mathbf{x}_t} \approx -\frac{\eta_{t-1}}{\alpha_t}\mathbf{I}. \tag{74}$$

It is an additional Jacobian approximation, not a consequence of Gaussianity or a first-order Taylor expansion alone. Its residual obeys

$$\|\boldsymbol{r}_t\|_2 \le \frac{\|\boldsymbol{\Sigma}_t\|_2\|\boldsymbol{\Sigma}_t^{-1}\|_2\|\boldsymbol{R}_t\|_2}{a_t}\|\boldsymbol{\delta}_t\|_2. \tag{75}$$

For isotropic $\boldsymbol{\Sigma}_t = v_t\mathbf{I}$, $\boldsymbol{A}_t = (\eta_t v_t/\eta_{t-1})\mathbf{I}$, recovering the scalar-times-gradient form. For general covariance, $\boldsymbol{A}_t$ is a matrix. A fitted scalar intercept is an empirical regression parameter and is not identified with the vector residual $\boldsymbol{r}_t$.

Remark (Generality beyond ResShift and UPSR). The preceding calculation uses discrete reverse kernels rather than the continuous reverse ODE in Eq. 1. For any model whose learned and HR-conditioned means have the form

$$\boldsymbol{\mu}_\theta = a_t\mathbf{x}_t + c_t\mathbf{y}_0 + b_t f_\theta, \qquad \boldsymbol{\mu}_q = a_t\mathbf{x}_t + c_t\mathbf{y}_0 + b_t\mathbf{x}_0, \tag{76}$$

with schedule-dependent coefficients, their difference is

$$\boldsymbol{\delta}_t = b_t(\mathbf{x}_0 - f_\theta). \tag{77}$$

The shared state and endpoint terms cancel. This applies to GOU-based formulations with such affine means, including retained endpoint terms Luo et al. (2023); Yue et al. (2024). For equal covariances independent of $\mathbf{x}_t$,

$$\nabla_{\mathbf{x}_t} D_{\mathrm{KL}} = -\boldsymbol{J}_t^\top\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\delta}_t, \qquad \boldsymbol{J}_t = b_t\frac{\partial f_\theta}{\partial \mathbf{x}_t}. \tag{78}$$

A scalar-times-covariance correction additionally requires a model-specific approximation $\boldsymbol{J}_t \approx -\rho_t\mathbf{I}$, $\rho_t \neq 0$, yielding $\boldsymbol{\delta}_t \approx \rho_t^{-1}\boldsymbol{\Sigma}_t\nabla_{\mathbf{x}_t} D_{\mathrm{KL}}$. The affine mean structure alone does not establish this approximation. ResShift has $a_t = \eta_{t-1}/\eta_t$, $b_t = \alpha_t/\eta_t$, and $c_t = 0$.

Proposition C.2 (Sensitivity of the KL gradient to the reference). *Fix the observation $\mathbf{y}_0$ and define*

$$Q_t^{\mathrm{HR}} = \mathcal{N}(m_t\mathbf{x}_0 + \boldsymbol{b}_t, \boldsymbol{S}_t), \tag{79}$$

$$Q_t^{\mathrm{LR}} = \mathcal{N}(m_t\mathbf{c} + \boldsymbol{b}_t, \boldsymbol{S}_t), \tag{80}$$

where $\boldsymbol{S}_t$ is positive definite. The quantities m_t, $\boldsymbol{b}_t$, and $\boldsymbol{S}_t$ remain fixed when replacing $\mathbf{x}_0$ by $\mathbf{c}$. Let $P_t = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{C}_t)$ be a Gaussian with fixed positive-definite covariance, and define

$$\boldsymbol{g}_t^{\mathrm{HR}} = \nabla_{\boldsymbol{\mu}_p} D_{\mathrm{KL}}(P_t\|Q_t^{\mathrm{HR}}), \qquad \boldsymbol{g}_t^{\mathrm{LR}} = \nabla_{\boldsymbol{\mu}_p} D_{\mathrm{KL}}(P_t\|Q_t^{\mathrm{LR}}). \tag{81}$$

Then

$$\boldsymbol{g}_t^{\mathrm{HR}} - \boldsymbol{g}_t^{\mathrm{LR}} = m_t\boldsymbol{S}_t^{-1}(\mathbf{c} - \mathbf{x}_0), \tag{82}$$

and

$$\begin{aligned}\|\boldsymbol{g}_t^{\mathrm{HR}} - \boldsymbol{g}_t^{\mathrm{LR}}\|_2 &\le L_t W_2(Q_t^{\mathrm{HR}}, Q_t^{\mathrm{LR}}) \\ &= |m_t|L_t\|\mathbf{x}_0 - \mathbf{c}\|_2, \qquad L_t = \|\boldsymbol{S}_t^{-1}\|_2.\end{aligned} \tag{83}$$

Proof. Lemma 1 gives

$$\boldsymbol{g}_t^{\mathrm{HR}} = \boldsymbol{S}_t^{-1}(\boldsymbol{\mu}_p - m_t\mathbf{x}_0 - \boldsymbol{b}_t), \tag{84}$$

$$\boldsymbol{g}_t^{\mathrm{LR}} = \boldsymbol{S}_t^{-1}(\boldsymbol{\mu}_p - m_t\mathbf{c} - \boldsymbol{b}_t). \tag{85}$$

Subtracting yields the claimed gradient difference. The operator-norm inequality gives

$$\|m_t\boldsymbol{S}_t^{-1}(\mathbf{c} - \mathbf{x}_0)\|_2 \le |m_t|\|\boldsymbol{S}_t^{-1}\|_2\|\mathbf{c} - \mathbf{x}_0\|_2. \tag{86}$$

For Gaussians with equal covariance, the Wasserstein-2 distance Gelbrich (1990) is the distance between means:

$$W_2(Q_t^{\mathrm{HR}}, Q_t^{\mathrm{LR}}) = |m_t|\|\mathbf{x}_0 - \mathbf{c}\|_2. \tag{87}$$

This proves Eq. 83. □

For DDPM, $m_t = \sqrt{\bar{\alpha}_t^{\mathrm{ddpm}}}$, $\boldsymbol{b}_t = \mathbf{0}$, and $\boldsymbol{S}_t = (1 - \bar{\alpha}_t^{\mathrm{ddpm}})\mathbf{I}$. For ResShift, Eq. 27 gives $m_t = 1 - \eta_t$, $\boldsymbol{b}_t = \eta_t\mathbf{y}_0$, and $\boldsymbol{S}_t = \kappa^2\eta_t\mathbf{I}$. For UPSR, the same affine means use $\boldsymbol{S}_t = \kappa^2\eta_t\mathbf{W}_u(\mathbf{y}_0)^2$ with the uncertainty map held fixed for both references. The bound controls the absolute gradient difference. For ResShift it vanishes as $\eta_t \to 1$ for fixed reference discrepancy; for UPSR this also requires the noise weights to remain bounded away from zero. Small absolute difference alone does not guarantee small relative error or close directional alignment. This proposition concerns forward-reference Gaussians and does not by itself identify their gradients with the reverse-kernel gradient in Proposition C.1.

Proposition C.3 (Guidance for auxiliary conditioning). *Suppose an auxiliary observation* $\mathbf{c}$ *admits the likelihood factorization in Eq. 91, and the base reverse kernel is* $\mathcal{N}(\boldsymbol{\mu}_\theta, \Sigma_\theta)$*. A first-order expansion of its log-likelihood around* $\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta$ *gives*

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0, \mathbf{c}) \approx \mathcal{N}(\boldsymbol{\mu}_\theta + \Sigma_\theta\boldsymbol{g}, \Sigma_\theta), \tag{88}$$

where $\boldsymbol{g} = \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{c} \mid \mathbf{x}_{t-1}, \mathbf{y}_0)\big|_{\mathbf{x}_{t-1}=\boldsymbol{\mu}_\theta}$*. Replacing this gradient by its current-state approximation yields*

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0, \mathbf{c}) \approx \mathcal{N}\left(\boldsymbol{\mu}_\theta + \Sigma_\theta\nabla_{\mathbf{x}_t} \log p(\mathbf{c} \mid \mathbf{x}_t, \mathbf{y}_0), \Sigma_\theta\right). \tag{89}$$

Here Σ_θ *is positive definite and fixed with respect to the transition variable* $\mathbf{x}_{t-1}$*; isotropy is not required.*

Proof. For a compatible conditional reverse Markov model,

$$p(\mathbf{x}_0 \mid \mathbf{y}_0, \mathbf{c}) = \int p(\mathbf{x}_T \mid \mathbf{y}_0, \mathbf{c}) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0, \mathbf{c})\, \mathrm{d}\mathbf{x}_{1:T}. \tag{90}$$

Under the conditional-independence assumption $p(\mathbf{c} \mid \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_0) = p(\mathbf{c} \mid \mathbf{x}_{t-1}, \mathbf{y}_0)$, Bayes' rule gives

$$\begin{aligned} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0, \mathbf{c}) &= \frac{p(\mathbf{c} \mid \mathbf{x}_{t-1}, \mathbf{y}_0)p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0)}{p(\mathbf{c} \mid \mathbf{x}_t, \mathbf{y}_0)} \\ &= Z\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}_0)p(\mathbf{c} \mid \mathbf{x}_{t-1}, \mathbf{y}_0), \end{aligned} \tag{91}$$

where Z is independent of $\mathbf{x}_{t-1}$. Following the local expansion used for classifier guidance Dhariwal & Nichol (2021),

$$\log p(\mathbf{c} \mid \mathbf{x}_{t-1}, \mathbf{y}_0) \approx \log p(\mathbf{c} \mid \boldsymbol{\mu}_\theta, \mathbf{y}_0) + (\mathbf{x}_{t-1} - \boldsymbol{\mu}_\theta)^\top\boldsymbol{g}. \tag{92}$$

Let $\boldsymbol{u} = \mathbf{x}_{t-1} - \boldsymbol{\mu}_\theta$. Completing the square gives

$$\begin{aligned} -\tfrac{1}{2}\boldsymbol{u}^\top\Sigma_\theta^{-1}\boldsymbol{u} + \boldsymbol{u}^\top\boldsymbol{g} = -\tfrac{1}{2}(\boldsymbol{u} - \Sigma_\theta\boldsymbol{g})^\top\Sigma_\theta^{-1}(\boldsymbol{u} - \Sigma_\theta\boldsymbol{g}) \\ + \tfrac{1}{2}\boldsymbol{g}^\top\Sigma_\theta\boldsymbol{g}. \end{aligned} \tag{93}$$

The last term is constant with respect to $\mathbf{x}_{t-1}$, yielding the shifted Gaussian. Evaluating the guidance at $\mathbf{x}_t$ instead of $\boldsymbol{\mu}_\theta$ is an additional approximation, giving Eq. 89. □

Remark (Gaussian image guidance). For LR-based control, define the auxiliary potential

$$\ell_t(\mathbf{x}_t; \mathbf{c}) = -\tfrac{1}{2}(\mathbf{x}_t - \mathbf{c})^\top\boldsymbol{\Gamma}_t^{-1}(\mathbf{x}_t - \mathbf{c}), \tag{94}$$

where $\boldsymbol{\Gamma}_t$ is positive definite and independent of $\mathbf{x}_t$ for fixed observation. Using the auxiliary Gaussians in Appendix C.2,

$$\nabla_{\mathbf{x}_t}\ell_t(\mathbf{x}_t; \mathbf{c}) = -\nabla_{\mathbf{x}_t}D_{\mathrm{KL}}(P_t^{\mathrm{aux}}\|Q_t^{\mathrm{aux}}) = -\boldsymbol{\Gamma}_t^{-1}(\mathbf{x}_t - \mathbf{c}). \tag{95}$$

The signed mean shift is therefore

$$\boldsymbol{\mu}_{\mathrm{new}} = \boldsymbol{\mu}_\theta - s\alpha_t^*\Sigma_{\theta,t}\boldsymbol{\Gamma}_t^{-1}(\mathbf{x}_t - \mathbf{c}). \tag{96}$$

In practice, $\boldsymbol{\Gamma}_t^{-1}$ does not need to be measured separately. It is measured together within α_t^*, this reduces to Eq. 50.

Proof. Differentiate the quadratic potential and apply Lemma 1 to the two auxiliary Gaussians. □

When $\mathbf{c}$ is a deterministic function of the already observed $\mathbf{y}_0$, it supplies no additional probabilistic information given $\mathbf{y}_0$. In that case, ℓ_t is a designed control potential, rather than an independent observation likelihood. Its linearization produces the same Gaussian mean-shift algebra as auxiliary conditioning. For either sign of the control, the implemented shifted Gaussian is well-defined. This does not assert normalizability of an unlinearized, negatively weighted quadratic energy tilt.

Equation 95 identifies the image-guidance potential with an auxiliary Gaussian KL gradient. It does not by itself identify that gradient with the reverse-kernel KL gradient in Proposition C.1. Such an identification requires an additional approximation relating the two gradients.