FAPE-IR: Frequency-Aware Planning and Execution Framework for All-in-One Image Restoration

Supplementary Material

7. Distortions and Over-Generation in AR/FM and Adversarial Training for Low-Level Restoration

7.1. Notation and Standing Assumptions

Let $\mathcal{X} = \mathbb{R}^d$ denote image space with Fourier variable $\omega \in \mathbb{R}^d$. Let $p \equiv p_{\text{data}}$ be the ground-truth image law and q_θ the model law of generator G_θ. For a smoothing schedule $\sigma_t \geq 0$, define $p_t := p * \mathcal{N}(0, \sigma_t^2 I)$ and write $\widehat{f}$ for the Fourier transform. Let $\Phi : \mathcal{X} \to \mathbb{R}^m$ be a fixed perceptual map (LPIPS), frozen during training. Expectations are finite.

We consider a standard linear degradation model for low-level restoration:

$$c = Hx + \eta, \qquad x \sim p,\ \eta \sim \mathcal{N}(0, \sigma_\eta^2 I), \tag{6}$$

with known operator $H : \mathcal{X} \to \mathcal{X}$ (e.g., blur+downsample for SR, convolutional blur for deblurring, masking for inpainting). Denote the conditional posterior by $p(x \mid c)$ and its smoothed versions $p_t(x \mid c) := p(\cdot \mid c) * \mathcal{N}(0, \sigma_t^2 I)$.

Assumption 1 (Spectral regularity of natural images)**.** *The power spectrum of natural images satisfies:*

$$S_{xx}(\omega) := \mathbb{E}|\widehat{x}(\omega)|^2 \asymp \|\omega\|^{-\kappa}, \quad \kappa > 0, \tag{7}$$

so that energy is mainly concentrated at low frequencies and decays polynomially with frequency [81]. This reflects the empirical smoothness of natural images and excludes degenerate high-frequency dominated signals [6, 19, 24, 67, 69, 81].

Assumption 2 (Forward operator)**.** *H is linear, bounded, and shift-invariant on the domain of interest with frequency response $\widehat{H}(\omega)$. For blur/downsample/inpainting, $|\widehat{H}(\omega)|$ decays (or vanishes) for large $\|\omega\|$ or outside the pass-band [29, 32, 80].*

Assumption 3 (Regularized flow path)**.** *The conditional path is given by Gaussian smoothing [43, 71, 94]:*

$$p_t(\cdot \mid c) = p(\cdot \mid c) * K_{\sigma_t}, \tag{8}$$

with σ_t smooth in t. Then p_t is C^1 in t, and there exists a drift $v_t^\star(\cdot \mid c)$ such that $\partial_t p_t + \nabla \cdot (v_t^\star p_t) = 0$. For $p_t = p * \mathcal{N}(0, \sigma_t^2 I)$ one can take the explicit choice [65, 82]:

$$v_t^\star(x, c) = -\tfrac{\dot{\sigma}_t^2}{2} \nabla \log p_t(x \mid c). \tag{9}$$

Assumption 4 (Lipschitz drift and locally bounded conditional density)**.** *Each $v_t^\star(\cdot \mid c)$ is L-Lipschitz. Moreover, for each t there exists a compact set $K_t \subset \mathcal{X}$ such that the conditional densities are locally bounded and bounded away from zero on K_t:*

$$0 < m_t \leq p_t(x \mid c) \leq M_t < \infty \quad \textit{for all } x \in K_t, \tag{10}$$

uniformly (in the sense appropriate to the expectations we consider). All expectations in this paper are taken over x restricted to K_t (or via truncation/limit arguments), which is natural in low-level restoration where reconstructions lie in a plausible compact region.

Assumption 5 (IPM critic: attainment and compact parameterization)**.** *We assume $d_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} (\mathbb{E}_p f - \mathbb{E}_q f)$ with $\mathcal{F} = \{f_\psi : \psi \in \Psi\}$ parameterized by a compact set Ψ, such that the supremum is attained at some $\psi^\star \in \Psi$. Then Danskin's theorem applies to yield:*

$$\nabla_\theta d_{\mathcal{F}}(p, q_\theta) \ = \ -\mathbb{E}\big[J_{G_\theta}^\top \nabla_x f_{\psi^\star}(G_\theta)\big]. \tag{11}$$

(Alternatively, one may work with subgradients if $\mathcal{F}$ is non-compact, e.g. the 1-Lipschitz ball for W_1.) See [2, 3, 72].

7.2. Objectives (Restoration Setting)

Autoregression (AR). Conditioned on c, the chain factorization is $q_\theta(x \mid c) = \prod_{t=1}^T q_\theta(x_t \mid x_{<t}, c)$, and

$$\theta_{\text{AR}}^\star \in \arg\min_\theta \mathbb{E}_{c \sim p(c)} \, \text{KL}\big(p(\cdot \mid c) \,\|\, q_\theta(\cdot \mid c)\big). \tag{12}$$

Flow Matching (FM). FM is a general framework for distribution transport, with Rectified Flow as the linear-path special case; related score-based views connect to diffusion/SDE modeling and score matching [23, 45, 46, 70]. Here we adopt the basic form based on conditional scores of smoothed posteriors:

$$\begin{aligned}\theta_{\text{FM}}^\star \in \arg\min_\theta \int_0^1 w(t) \, \mathbb{E}_{c \sim p(c)} \, \mathbb{E}_{x_t \sim p_t(\cdot|c)} \\ \big\| s_\theta(x_t, c, t) - \nabla \log p_t(x_t \mid c) \big\|_2^2 \, dt.\end{aligned} \tag{13}$$

Adversarial Training. Given pairs (c, x), augment the composite objective with a measurement term:

$$\begin{aligned}\mathcal{J}(\theta) = \lambda \, d_{\mathcal{F}}\big(p, q_\theta\big) + \alpha \, \mathbb{E} \, \|G_\theta(c) - x\|_2^2 \\ + \beta \, \mathbb{E} \, \|\Phi(G_\theta(c)) - \Phi(x)\|_2^2, \quad \lambda, \alpha, \beta \geq 0,\end{aligned} \tag{14}$$

with $d_{\mathcal{F}}$ often chosen as W_1 (WGAN) [2, 3, 72].

7.3. AR Causes Artifacts and Over-Generation

Modeling assumptions for AR proofs. We make explicit the family of admissible AR conditionals and the inference rule used at test time.

Assumption 6 (Single-mode AR conditionals or deterministic limit)**.** *For each t, conditionals $q_\theta(x_t \mid x_{<t}, c)$ belong to a centered, symmetric, log-concave location family:*

$$\mathcal{Q} := \{\, x_t \mapsto f_\sigma(x_t - \mu_t) \;:\; \mu_t \in \mathbb{R}^{d_t},\ \sigma \in (0, \sigma_{\max}] \,\}, \tag{15}$$

where f_σ is even, strictly log-concave, smooth in σ, and the deterministic limit $\sigma \to 0$*, i.e. $q_\theta(\cdot \mid x_{<t}, c) \Rightarrow \delta_{\mu_t(x_{<t},c)}$.*

Assumption 7 (Greedy or vanishing-temperature decoding)**.** *At test time, decoding is either greedy ($x_t = \arg\max q_\theta(\cdot \mid x_{<t}, c)$) or stochastic with temperature $\tau \downarrow 0$ so that $x_t \Rightarrow \arg\max q_\theta(\cdot \mid x_{<t}, c)$ in probability.*

Definition 1 (Posterior $\ker(H)$-ambiguity at step t)**.** *Let $E_t : \mathbb{R}^{d_t} \to \mathcal{X}$ be the (fixed) embedding that maps the local step-t variable into image space. We say that the step-t posterior exhibits a symmetric $\ker(H)$-ambiguity with gap $u_t \in \mathbb{R}^{d_t}$ if for a measurable set $A_t \subseteq \{(x_{<t}, c)\}$ with $p(A_t) \geq \rho_t > 0$,*

$$\begin{aligned} p(x_t \mid x_{<t}, c) &= \tfrac{1}{2}\,\delta_{x_t^{(0)}(x_{<t},c)+u_t} \\ &\quad + \tfrac{1}{2}\,\delta_{x_t^{(0)}(x_{<t},c)-u_t}, \\ H\,E_t u_t &= 0. \end{aligned} \tag{16}$$

Here $x_t^{(0)}$ is the posterior midpoint in the local coordinates; the corresponding image-space ambiguity direction is $E_t u_t$.

Theorem 1 (AR deterministic collapse along $\ker(H)$)**.** *Under Assumptions 6–7 and Definition 1, the maximum-likelihood solution of Equation 12 satisfies*

$$\mu_t^\star(x_{<t}, c) = x_t^{(0)}(x_{<t}, c) \text{ for } p\text{-a.e. } (x_{<t}, c) \in A_t, \tag{17}$$

and the decoded token obeys $x_t = \mu_t^\star(x_{<t}, c)$ a.s. in the greedy limit. Consequently, the image-space component of x_t along the ambiguity direction $\mathrm{span}\{E_t u_t\}$ is collapsed to zero mean, removing genuine $\pm E_t u_t$ variability and inducing structured artifacts aligned with $\ker(H)$.

Proof. Fix $(x_{<t}, c) \in A_t$. By Assumption 6 the conditional log-likelihood for center μ is $\ell(\mu) = \log f_\sigma(x_t^{(0)} + u_t - \mu) + \log f_\sigma(x_t^{(0)} - u_t - \mu)$. Since f_σ is even and strictly log-concave, ℓ is strictly concave and achieves its unique maximum at the midpoint $\mu = x_t^{(0)}$. Teacher forcing MLE therefore yields $\mu_t^\star = x_t^{(0)}$ on A_t, and by Assumption 7 the decoded value equals $\mu_t^\star$ almost surely in the deterministic limit. Because $H E_t u_t = 0$, the collapsed direction is exactly unobservable in c, hence constitutes artifact-prone freedom. □

Remark 1 (When collapse can be avoided). If the conditional family admits *explicit multimodality* (e.g. mixture components with separate centers and nontrivial sampling at test time), Theorem 1 need not hold. This motivates latent-variable AR or distributional heads when H is severely rank-deficient.

From local collapse to global artifact patterns. We connect stepwise ambiguity to image-level artifacts via an additive geometric functional.

Definition 2 (Artifact *deficit*)**.** *Let $\Pi : \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a convex seminorm measuring nullspace-structured deviations (e.g. $\Pi(x) = \|P_{\ker(H)} x\|_2$ or a high-pass seminorm restricted to $\ker(H)$). For a sample $\tilde{x} \sim q_\theta(\cdot \mid c)$ define the* artifact deficit*:*

$$\mathcal{D}(\tilde{x}; c) \;:=\; \mathbb{E}_{p(x|c)}\Pi(x) \;-\; \Pi(\tilde{x}). \tag{18}$$

By Jensen's inequality and Theorem 1, when the posterior on a direction is symmetric and decoding collapses to the midpoint, $\mathcal{D}$ is typically nonnegative; its positivity quantifies the loss of genuine posterior variability along $\ker(H)$.

Proposition 1 (TV-based upper bound on deficit over ambiguity sets)**.** *Suppose Definition 1 holds for indices $t \in \mathcal{T} \subseteq \{1, \ldots, T\}$ with probabilities $\{\rho_t\}$ and ambiguity vectors $\{u_t\}$. Let Π be any convex seminorm that is strictly convex on $\ker(H)$. For each $R > 0$, define the truncated seminorm $\Pi_R := \min\{\Pi, R\}$ to ensure boundedness on K_t (Assumption 4). Define on A_t:*

$$\bar{\varepsilon}_t^{(A)} := \operatorname*{ess\,sup}_{(x_{<t},c)\in A_t} \mathrm{TV}\big(p(\cdot \mid x_{<t}, c),\, q_\theta(\cdot \mid x_{<t}, c)\big). \tag{19}$$

Let

$$\begin{aligned} \Delta_{\Pi_R}^{\max} := \sup_{t\leq T} \sup_{(x_{<t},c)\in A_t} \tfrac{1}{2}\Big(&\Pi_R(x_t^{(0)}+u_t) \\ &+ \Pi_R(x_t^{(0)}-u_t) - 2\Pi_R(x_t^{(0)})\Big). \end{aligned} \tag{20}$$

Then

$$\begin{aligned} \mathbb{E}\big[\mathcal{D}_R(\tilde{x}; c)\big] \;\leq\; \Delta_{\Pi_R}^{\max} \sum_{t=1}^{T} \rho_t\, \bar{\varepsilon}_t^{(A)} \\ +\; \mathbb{E}\big[\mathcal{D}_R(\tilde{x}; c)\, \mathbf{1}_{(\cup_t A_t)^\complement}\big], \end{aligned} \tag{21}$$

where $\mathcal{D}_R$ is defined by replacing Π with Π_R in Definition 2. Passing to the limit $R \to \infty$ by monotone convergence yields the same bound for $\mathcal{D}$ provided the moments in Assumption 4 hold. The complement term can be further bounded once additional structure outside $\cup_t A_t$ is specified (e.g. no nullspace ambiguity or separate curvature bounds).

Proof. On each A_t, for bounded Π_R we have $|\mathbb{E}_p \Pi_R - \mathbb{E}_q \Pi_R| \leq \|\Pi_R\|_\infty \mathrm{TV}(p, q)$. Summing over $t \in \mathcal{T}$ and using the definition of $\Delta_{\Pi_R}^{\max}$ yields Equation 21. The monotone limit follows by Assumption 4. □

Quantifying Exposure Amplification in Rollout.

Assumption 8 (Lipschitz artifact seminorm)**.** *There exists $L_\Pi > 0$ such that $|\Pi(x) - \Pi(y)| \le L_\Pi \|P_{\ker(H)}(x-y)\|_2$ for all x, y. When needed, we also use the truncated seminorm $\Pi_R := \min\{\Pi, R\}$ to guarantee integrability; all bounds pass to the limit $R \to \infty$ by monotone convergence under our moment assumptions.*

Proposition 2 (Positive artifact deficit under repeated ambiguity)**.** *Suppose Definition 1 holds for indices $t \in \mathcal{T} \subseteq \{1, \dots, T\}$ with probabilities $\{\rho_t\}$ and ambiguity vectors $\{u_t\}$. Let Π be any convex seminorm that is strictly convex on $\ker(H)$. Then:*

$$\mathbb{E}\big[\mathcal{D}(\tilde{x}; c)\big] \;\ge\; \sum_{t\in\mathcal{T}} \rho_t \Big[\tfrac{1}{2}\Pi\big(x_t^{(0)}+u_t\big) + \tfrac{1}{2}\Pi\big(x_t^{(0)}-u_t\big)\Big] - \sum_{t\in\mathcal{T}} \rho_t\, \Pi\big(x_t^{(0)}\big). \tag{22}$$

Proof. Fix $(x_{<t}, c) \in A_t$. Under p, Π sees two symmetric values around $x_t^{(0)}$; by convexity $\frac{1}{2}\Pi(x_t^{(0)}+u_t) + \frac{1}{2}\Pi(x_t^{(0)}-u_t) \ge \Pi(x_t^{(0)})$, with strict inequality by strict convexity if $u_t \neq 0$. Under q_θ, Theorem 1 yields the degenerate value $\Pi(x_t^{(0)})$. Taking differences and averaging over A_t gives the bound; summing over indices $t \in \mathcal{T}$ yields the stated inequality. A matching lower/upper bound in general needs additional independence/orthogonality assumptions and is omitted here. □

Exposure bias and accumulation. During training, loss is minimized with teacher forcing; at test time the model uses its own predictions. Define the per-step deviation:

$$\delta_t(x_{<t}, c) := \mathrm{TV}\big(p(\cdot \mid x_{<t}, c),\, q_\theta(\cdot \mid x_{<t}, c)\big), \qquad \varepsilon_t := \mathbb{E}_{x_{<t}\sim p(\cdot|c)}\, \delta_t(x_{<t}, c). \tag{23}$$

We also write the *worst-case* per-step deviation $\bar{\varepsilon}_t := \operatorname{ess\,sup}_{x_{<t},c} \delta_t(x_{<t}, c)$.

Lemma 1 (Accumulation of local deviations)**.** *For the AR factorizations of p and q_θ using consistent measurable versions of the conditionals, the following bounds hold:*

$$\mathrm{TV}\big(p(\cdot \mid c), q_\theta(\cdot \mid c)\big) \;\le\; 1 - \prod_{t=1}^{T}(1-\bar{\varepsilon}_t) \;\le\; \sum_{t=1}^{T} \bar{\varepsilon}_t, \tag{24}$$

and, independently,

$$\mathrm{TV}\big(p(\cdot \mid c), q_\theta(\cdot \mid c)\big) \;\le\; \sum_{t=1}^{T} \varepsilon_t. \tag{25}$$

In particular, for small $\{\alpha_t\}$ with $\alpha_t \in \{\bar{\varepsilon}_t\}$,

$$1 - \prod_{t=1}^{T}(1-\alpha_t) = \sum_{t=1}^{T} \alpha_t \;+\; \mathcal{O}\Big(\sum_{s<t} \alpha_s \alpha_t\Big). \tag{26}$$

Chain rule and conditional optimization. By standard decomposition,

$$\mathrm{KL}\big(p(\cdot \mid c) \,\|\, q_\theta(\cdot \mid c)\big) = \sum_{t=1}^{T} \mathbb{E}_{x\sim p(\cdot|c)} \Big[\mathrm{KL}\big(p(\cdot \mid x_{<t}, c) \,\|\, q_\theta(\cdot \mid x_{<t}, c)\big)\Big]. \tag{27}$$

Thus training matches each conditional distribution separately. However, if $p(x_t \mid x_{<t}, c)$ admits multiple plausible continuations (e.g. along weakly observed directions of H), then any surrogate that collapses to a single conditional predictor necessarily *destroys* these alternatives. This mismatch does not show up as blur in practice, samples remain sharp, but it manifests as systematic artifacts where the model enforces spurious "averaged" structures.

Proof of Equation 27. Write $p(x \mid c) = \prod_{t=1}^{T} p(x_t \mid x_{<t}, c)$ and $q_\theta(x \mid c) = \prod_{t=1}^{T} q_\theta(x_t \mid x_{<t}, c)$. Then

$$\begin{aligned}\mathrm{KL}\big(p \| q_\theta\big) &= \mathbb{E}_{x\sim p}\left[\sum_{t=1}^{T} \log \frac{p(x_t \mid x_{<t}, c)}{q_\theta(x_t \mid x_{<t}, c)}\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{x_{<t}\sim p}\, \mathrm{KL}\big(p(\cdot \mid x_{<t}, c) \,\|\, q_\theta(\cdot \mid x_{<t}, c)\big).\end{aligned} \tag{28}$$

Hence Equation 27 holds, formalizing that AR training aligns each conditional factor separately. □

Proposition 3 (Nullspace ambiguity induces artificial averaging)**.** *Let H in Equation 6 have nontrivial nullspace and suppose the posterior $p(x \mid c)$ assigns equal mass to $x_\pm = x_0 \pm u$ with $Hu = 0$. Then the Bayes point estimator is:*

$$\widehat{x} = \mathbb{E}[x \mid c] = x_0, \tag{29}$$

which cancels *the true variability $\pm u$. Thus any deterministic AR predictor removes detail along u, creating structured artifacts not present in the true posterior (while stochastic decoding that breaks symmetry may* inconsistently *realize one of the modes).*

Proof. Under squared loss, the Bayes estimator equals the conditional mean:

$$\widehat{x} = \mathbb{E}[x \mid c] = \tfrac{1}{2}(x_0 + u) + \tfrac{1}{2}(x_0 - u) = x_0. \tag{30}$$

Hence the genuine posterior variability $\pm u$ (unobservable since $Hu = 0$) is averaged out.

Moreover, consider an AR conditional family restricted to a symmetric log-concave location family (e.g. $x_t \sim \mathcal{N}(\mu_t, \sigma^2 I)$ with fixed σ, or the $\sigma \to 0$ deterministic limit). Given a symmetric two-point posterior $\frac{1}{2}\delta_{x_0+u} + \frac{1}{2}\delta_{x_0-u}$, the conditional MLE for the center is the midpoint x_0 by symmetry and concavity: the log-likelihood takes the form

$\ell(\mu) = \log f(x_0 + u - \mu) + \log f(x_0 - u - \mu)$ with f log-concave and even; ℓ is concave and maximized at $\mu = x_0$ (where the two arguments are opposite), yielding the same collapse to x_0. Thus single-mode/deterministic conditionals enforce artificial averaging along nullspace directions. □

Takeaways. AR provably enforces *deterministic resolutions* of measurement ambiguities, averages away legitimate alternatives along nullspace directions, and accumulates local deviations into global structural errors. Unlike blur, which is not observed in practice, these mechanisms explain the clear yet artifact-laden outputs: geometric kinks, inconsistent fills, and over-generation of non-existent structures. Exposure/rollout issues are closely related to scheduled sampling, sequence-level training, and imitation-learning reductions [5, 62, 66].

Remark on bounds. The product-form term in Equation 24 uses worst-case per-step deviations $\bar{\varepsilon}_t$ and is tight for sequential maximal couplings [80]. A nontrivial *lower* bound on $\mathrm{TV}(p, q_\theta)$ or on artifact deficits in terms of per-step quantities generally requires additional independence/decoupling assumptions; without them we provide only the robust upper bounds above (cf. Proposition 1).

7.4. Why Distortions and Over-generation Appear in Flow Matching Optimization

A precise local sandwich. We refine the conditional spectral sandwich (Theorem 3, stated later) by (i) focusing on the small-perturbation regime and (ii) replacing $|\widehat{H}(\omega)|^2$ with an *information transfer coefficient* $\Xi_t(\omega)$ capturing the H-noise-prior interplay [80].

Assumption 9 (Small-perturbation regime and bounded densities)**.** *Let $p_t(\cdot \mid c) = p(\cdot \mid c) * K_{\sigma_t}$ and assume that $q_{\theta,t}(\cdot \mid c)$ admits the representation:*

$$q_{\theta,t}(\cdot \mid c) = q_\theta(\cdot \mid c) * K_{\sigma_t}, p_t(\cdot \mid c) = p(\cdot \mid c) * K_{\sigma_t}. \quad (31)$$

Define the unsmoothed *discrepancy $\delta(\cdot \mid c) := q_\theta(\cdot \mid c) - p(\cdot \mid c)$ and its smoothed version*

$$\Delta_t(\cdot \mid c) := \delta(\cdot \mid c) * K_{\sigma_t}, \quad (32)$$

so that $q_{\theta,t} = p_t + \Delta_t$ and $\int \Delta_t(x \mid c)\, dx = 0$. Assume $\|\Delta_t\|_{L^\infty} \le \eta_t m_t/2$ with some $0 < \eta_t < 1$ and $\|\nabla \Delta_t\|_{L^2} < \infty$, where $0 < m_t \le p_t \le M_t < \infty$ are as in Assumption 4.

Definition 3 (Information transfer coefficient)**.** *In the linear–Gaussian setting with $x \sim \mathcal{N}(0, S_{xx})$ and $c = Hx + \eta$, define the frequency-wise* posterior information weight

$$\Xi_{\mathrm{LG}}(\omega) := \frac{|\widehat{H}(\omega)|^2}{\sigma_\eta^2 + |\widehat{H}(\omega)|^2 S_{xx}(\omega)} \in [0, 1/\sigma_\eta^2], \quad (33)$$

which vanishes where $\widehat{H}(\omega) = 0$ and increases with the local posterior SNR. In the general small-perturbation regime of Assumption 9, there exist constants $0 < c_t^{(1)} \le c_t^{(2)} < \infty$ (depending only on (m_t, M_t) and low-order moments of $p_t(\cdot \mid c)$, but not on ω) such that:

$$c_t^{(1)}\, \Xi_{\mathrm{LG}}(\omega) \;\le\; \Xi_t(\omega) \;\le\; c_t^{(2)}\, \Xi_{\mathrm{LG}}(\omega). \quad (34)$$

Lemma 2 (Pythagorean decomposition in $L^2(p_t)$)**.** *Let $s_\theta = \nabla \log q_{\theta,t} + r_{\theta,t}$ with $r_{\theta,t} \in L^2(p_t)$. Then*

$$\mathbb{E}_{p_t} \|s_\theta - \nabla \log p_t\|_2^2 = D_F(p_t \| q_{\theta,t}) + \|r_{\theta,t}\|_{L^2(p_t)}^2 + 2\langle r_{\theta,t},\, \nabla \log q_{\theta,t} - \nabla \log p_t \rangle_{L^2(p_t)}. \quad (35)$$

If $r_{\theta,t}$ is the $L^2(p_t)$-orthogonal residual of projecting $\nabla \log p_t$ onto the model class, the cross term vanishes and the training loss splits into a Fisher part plus an approximation error (this orthogonality is an idealized projection property and need not hold for general parameterizations).

Theorem 2 (Local Fisher sandwich with spectral weights)**.** *Under Assumptions 2, 4, and 9, there exist finite constants $0 < c_t \le C_t < \infty$ such that the (conditional) Fisher divergence satisfies:*

$$\begin{aligned} &\int_0^1 w(t)\, c_t \int_{\mathbb{R}^d} \underbrace{\|\omega\|^2 e^{-\sigma_t^2 \|\omega\|^2}\, \Xi_t(\omega)}_{=:\widetilde{W}_t(\omega)} \mathbb{E}_c \big|\widehat{\delta}(\omega \mid c)\big|^2 \, d\omega\, dt \\ &\le \int_0^1 w(t)\, D_F\big(p_t(\cdot \mid c) \,\|\, q_{\theta,t}(\cdot \mid c)\big)\, dt \\ &\le \int_0^1 w(t)\, C_t \int_{\mathbb{R}^d} \widetilde{W}_t(\omega)\, \mathbb{E}_c \big|\widehat{\delta}(\omega \mid c)\big|^2 \, d\omega\, dt + R, \end{aligned} \quad (36)$$

with a quadratic remainder $0 \le R \le \int_0^1 w(t)\, \kappa_t \|\Delta_t\|_{L^\infty}^2\, dt$, where κ_t depends on (m_t, M_t) and the Lipschitz constants in Assumption 4. As $\max_t \|\Delta_t\|_{L^\infty} \to 0$, we have $R \to 0$, and Equation 36 becomes an equality up to factors c_t, C_t. Moreover, on any time window where $\sigma_t \in [\sigma_{\min}, \sigma_{\max}]$ is bounded, the constants c_t, C_t can be chosen uniformly bounded with respect to t on that window.

Corollary 1 (From Fisher to the FM objective)**.** *Under the setting of Lemma 2, the FM loss in Equation 13 decomposes into the Fisher term bounded by Equation 36 plus a nonnegative approximation error $\|r_{\theta,t}\|_{L^2(p_t)}^2$ (when the cross term vanishes). Hence all conclusions drawn from Equation 36 for D_F transfer to $\mathcal{L}_{\mathrm{FM}}$ up to adding this nonnegative term.*

Corollary 2 (High-frequency underweighting and nullspace gaps)**.** *Under Theorem 2, for any set $\Omega \subset \mathbb{R}^d$:*

1. *If $\inf_{\omega \in \Omega} \|\omega\| \to \infty$, then $\sup_{\omega \in \Omega} \widetilde{W}_t(\omega) \to 0$ exponentially in $\|\omega\|$, hence discrepancies concentrated in Ω contribute arbitrarily little to $\mathcal{L}_{\mathrm{FM}}$.*

2. *If $|\widehat{H}(\omega)| = 0$ on Ω, then $\Xi_t(\omega) = 0$ on Ω (for the linear–Gaussian definition) and thus $\widetilde{W}_t(\omega) = 0$; this establishes exact blindness on $\ker(H)$ bands. When $|\widehat{H}(\omega)|$ is nonzero but very small, the weight remains near-blind.*

Early-time forgetting bound.

Proposition 4 (Quantified forgetting under strong smoothing (expectation version))**.** *Let $p_t(\cdot \mid c) = p(\cdot \mid c) * \mathcal{N}(0, \sigma_t^2 I)$ and assume $\mathbb{E}\|Hx\| < \infty$. Then there exists a constant $C_d > 0$ depending only on the dimension such that:*

$$\mathbb{E}_c \,\mathrm{TV}\big(p_t(\cdot \mid c),\, p_t(\cdot)\big) \le C_d \times \min\left\{1,\ \frac{1}{\sigma_t}\,\mathbb{E}_c\, W_1\big(p(\cdot \mid c), p(\cdot)\big)\right\} \xrightarrow[\sigma_t \to \infty]{} 0. \tag{37}$$

Proof sketch. *Convolution with K_{σ_t} smooths test functions by shrinking their effective Lipschitz seminorm by $\|\nabla K_{\sigma_t}\|_{L^1} = \Theta(1/\sigma_t)$; apply the Kantorovich–Rubinstein dual for TV/Wasserstein comparison on smoothed measures and integrate over c.*

Remark 2 (Uniform version under additional boundedness). If, in addition, c is restricted to a compact set (or one assumes $\sup_c W_1(p(\cdot \mid c), p(\cdot)) < \infty$), then the same argument yields:

$$\sup_c \,\mathrm{TV}\big(p_t(\cdot \mid c), p_t(\cdot)\big) \;\le\; \frac{C_d}{\sigma_t} \times \sigma_t \sup_c W_1\big(p(\cdot \mid c), p(\cdot)\big) \xrightarrow[\sigma_t \to \infty]{} 0. \tag{38}$$

Large perceptual error with small FM loss.

Proposition 5 (Loss–distortion gap with quantitative construction)**.** *Let Φ be L_Φ-Lipschitz. Fix a measurable $\Omega \subset \mathbb{R}^d$ with $\sup_{\omega\in\Omega} \widetilde{W}_t(\omega) \le \epsilon_W$. Assume further that Φ is* co-Lipschitz *on Ω: there exist a seminorm Π_Ω supported on Ω and a constant $\kappa_\Omega > 0$ such that for all u with $\operatorname{supp}\widehat{u} \subseteq \Omega$,*

$$\Pi_\Omega(u) \;\le\; \kappa_\Omega^{-1}\, \|\Phi(u)\|_2. \tag{39}$$

Then there is a constant $C > 0$ (depending on $\Omega, \{w(t), \sigma_t\}$) and a perturbation family $\{\Delta_t\}_t$ with uniformly small $\|\Delta_t\|_\infty$ such that

$$\int_0^1 w(t) \int \widetilde{W}_t\, \mathbb{E}_c |\widehat{\delta}|^2 \le \epsilon_W, \mathbb{E}\|\Phi(x) - \Phi(\tilde{x})\|_2^2 \ge C\,\kappa_\Omega^2. \tag{40}$$

Specifically, take $\widehat{\delta}(\omega \mid c) = a_t\, \mathbf{1}_\Omega(\omega)\, e^{\mathrm{i}\varphi(c,\omega)}$ with phases chosen so that $\|\Delta_t\|_\infty \le \|\widehat{\Delta_t}\|_{L^1} \le |a_t|\,|\Omega|$, and scale a_t so that the weighted quadratic form equals ϵ_W. Hausdorff–Young and Plancherel give the stated controls; the co-Lipschitz property transfers spectral mass on Ω to a nontrivial feature deviation.

A weighted spectral view. Let $\Delta_t(x \mid c) := q_{\theta,t}(x \mid c) - p_t(x \mid c) = (q_\theta - p) * K_{\sigma_t}$ and write $\widehat{\cdot}$ for Fourier transforms in x. Define

$$\widetilde{W}_t(\omega) := \|\omega\|^2 e^{-\sigma_t^2\|\omega\|^2}\, \Xi_t(\omega). \tag{41}$$

Theorem 3 (Conditional spectral sandwich under small perturbations)**.** *Assume 2, 4, and 9. Then there exist finite constants $0 < c_t \le C_t < \infty$ such that for all θ,*

$$\begin{aligned} &\int_0^1 w(t)\, c_t \int_{\mathbb{R}^d} \widetilde{W}_t(\omega) \\ &\mathbb{E}_c\big|\widehat{\delta}(\omega \mid c)\big|^2\, d\omega\, dt \;\lesssim\; \mathcal{L}_{\mathrm{FM}}(\theta) \\ &\mathcal{L}_{\mathrm{FM}}(\theta) \;\lesssim\; \int_0^1 w(t)\, C_t \\ &\int_{\mathbb{R}^d} \widetilde{W}_t(\omega)\, \mathbb{E}_c\big|\widehat{\delta}(\omega \mid c)\big|^2\, d\omega\, dt + R, \end{aligned} \tag{42}$$

with $0 \le R \le \int_0^1 w(t)\, \kappa_t\, \|\Delta_t\|_{L^\infty}^2\, dt$. Consequently:

1. *(*High-frequency down-weighting*) Since $\widetilde{W}_t(\omega) \propto \|\omega\|^2 e^{-\sigma_t^2\|\omega\|^2}$, large-$\|\omega\|$ discrepancies contribute exponentially little to $\mathcal{L}_{\mathrm{FM}}$.*
2. *(*Nullspace down-weighting*) If $|\widehat{H}(\omega)| = 0$ on Ω, then $\Xi_t(\omega) = 0$ and $\widetilde{W}_t(\omega) = 0$ on Ω, establishing exact blindness on $\ker(H)$ bands.*
3. *(*Loss–distortion gap*) By concentrating $\widehat{\delta}$ where $\widetilde{W}_t$ is tiny and controlling $\|\Delta_t\|_\infty$, one can keep $\mathcal{L}_{\mathrm{FM}}$ small while incurring large pixel/perceptual deviations (cf. Proposition 5).*

Takeaway 1 (Objective Bias): FM minimizes a *spectrally reweighted* discrepancy, where high frequencies and directions in the kernel of H are under-penalized. This leads to an *identifiability gap*, where visually distinct reconstructions can exhibit nearly identical FM loss values.

From objective bias to visible geometry: flow amplification. Let x follow the true conditional flow $\dot{x} = v_t^\star(x, c, t)$ (Assumption 3) and $\tilde{x}$ the learned flow $\dot{\tilde{x}} = v_\theta(\tilde{x}, c, t)$.

Lemma 3 (Conditional flow stability and geometric warp)**.** *With $e_t = v_\theta - v_t^\star$ and L the Lipschitz constant of $v_t^\star$,*

$$\|x(t) - \tilde{x}(t)\| \;\le\; e^{Lt}\left(\|x(0) - \tilde{x}(0)\| + \int_0^t \|e_s(\tilde{x}, c, s)\|\, ds\right). \tag{43}$$

Corollary 3 (Edge bending (heuristic))**.** *Let $\Gamma \subset \mathcal{X}$ be a high-curvature level set (edge/contour). Suppose e_t is predominantly supported where the learned dynamics deviates around edges (a region where small phase errors matter most). Then even when $\int_0^1 \mathbb{E}\|e_t\|^2 dt$ is small under the FM weighting, the spatial displacement of Γ under $\tilde{x}(\cdot)$ can be $O(e^L)$ relative to local curvature radii, producing visible* warps/kinks.

Intuition. Because FM underweights high-ω errors (due to heat-kernel smoothing and measurement passband), the learned drift can be slightly wrong where edges live. The ODE then *integrates* these local, directionally coherent errors into macroscopic geometric distortions (Lemma 3).

Why over-generation (hallucinated detail) appears. Two complementary mechanisms follow from Theorem 3:

Proposition 6 (Nullspace over-generation under down-weighting)**.** *Suppose there exists $\Omega \neq \emptyset$ with $|\widehat{H}(\omega)| = 0$ on Ω and the setting is linear–Gaussian so that $\Xi_t(\omega) = 0$ on Ω. If the generator class can realize perturbations supported on Ω (assumption on representational capacity within K_t), then for any $\epsilon > 0$ there is q_θ with $\mathcal{L}_{\mathrm{FM}}(\theta) \leq \epsilon$ whose samples $\tilde{x}$ contain arbitrarily strong components supported on Ω. Thus textures along* $\ker H$ *can be invented at negligible FM loss in this regime.*

Remark 3 (Scope vs. small-perturbation regime). The construction in Proposition 6 addresses the *attainable worst case* and is *not* restricted by the small-$\|\Delta_t\|_\infty$ assumption of Assumption 9. When $\|\Delta_t\|_\infty$ is explicitly constrained, the strength of over-generated $\ker H$ components is likewise bounded.

Proposition 7 (Early-time conditioning leakage)**.** *Suppose $w(t)$ gives nontrivial mass to large σ_t (early times). Then $\widetilde{W}_t(\omega) \to 0$ for all ω as $\sigma_t \to \infty$, so the early-time portion of Equation 13 is weakly informative about c. By Proposition 4, we have $\mathbb{E}_c \mathrm{TV}\big(p_t(\cdot \mid c), p_t(\cdot)\big) \to 0$ as $\sigma_t \to \infty$, hence the early-time loss is effectively unconditional in the strong-smoothing regime. The learned drift there tends toward the* unconditional *score, which injects prior textures that later stages cannot fully remove in underconstrained regions, contributing to* over-generation.

Summary of mechanisms for FM artifacts

- *Spectral bias* (Theorem 3(a)): high-ω penalties are tiny $\Rightarrow$ **edge softness, texture loss**.
- *Nullspace blindness* (Theorem 3(b), Prop. 6): H unpenalized $\Rightarrow$ **hallucinated detail** in underspecified regions.
- *Flow amplification* (Cor. 3): small drift errors near edges integrate to **geometric warps/kinks**.
- *Early-time leakage* (Prop. 7): weak conditioning at large σ_t injects prior textures $\Rightarrow$ **over-generation** that later steps cannot reliably remove.

Design implications for FM in restoration.

1. *Spectral reweighting.* Modify the loss with a deconvolution factor to counter the spectral weight (W_t globally, or $\widetilde{W}_t$ in the local view):

$$\tilde{\mathcal{L}} = \int_0^1 w(t)\, \mathbb{E}\left\| \Lambda_t^{-1}\big(s_\theta - \nabla \log p_t\big)\right\|^2, \qquad \widehat{\Lambda_t}(\omega) \propto \|\omega\|\, e^{-\sigma_t^2 \|\omega\|^2/2}\, \sqrt{\Xi_t(\omega)}, \tag{44}$$

   with *clipping* to avoid noise amplification.

2. *Data consistency.* Interleave FM steps with projections onto $\{x : \|Hx - c\| \leq \tau\}$ or add a strong penalty $\gamma \|HG_\theta(c) - c\|^2$ during training/sampling to break nullspace blindness [6]; see also PnP/RED frameworks [23, 65, 70, 82].
3. *Edge-aware guidance.* Upweight loss contributions near high $\|\nabla_x \Phi(x)\|$ (perceptual edges) or blend an IPM on high-pass features to prevent geometric warps.
4. *Time weighting.* Reduce $w(t)$ for very large σ_t or use schedules that spend more capacity near well-conditioned times where c carries information.

8. Adversarial Training: High-Fidelity Control Without Over-Generation

8.1. Objective and Assumptions

Under Equation 6, let $G_\theta(c, z)$ be a generator with latent z, and let $d_{\mathcal{F}}$ be an Integral Probability Metric (IPM) induced by a compact function class $\mathcal{F}$ (e.g., 1-Lipschitz critics for W_1) [3]. We consider the composite objective

$$\mathcal{J}(\theta) = \lambda\, d_{\mathcal{F}}\big(p, q_\theta\big) + \alpha\, \mathbb{E}\, \|G_\theta(c, z) - x\|_2^2 + \beta\, \mathbb{E}\, \|\Phi(G_\theta(c, z)) - \Phi(x)\|_2^2, \tag{45}$$

with nonnegative weights λ, α, β. Here q_θ is the model law from G_θ, and $\Phi : \mathcal{X} \to \mathbb{R}^m$ is a fixed perceptual map.

Assumption 10 (Regularity)**.** *$\mathcal{F}$ is compact and satisfies Danskin differentiability (Assumption 5); Φ is L_Φ-Lipschitz when relating to high-frequency seminorms; G_θ is differentiable with bounded Jacobian J_{G_θ} on the support of interest; expectations are finite.*

8.2. Gradient Representation and Stationarity

Lemma 4 (IPM gradient/subgradient for the critic)**.** *Assume $\mathcal{F} = \{f_\psi : \psi \in \Psi\}$ is a compact, parameterized class (e.g., spectrally-normalized networks) and the supremum is attained at some $\psi^\star$. Then*

$$\nabla_\theta d_{\mathcal{F}}(p, q_\theta) = -\,\mathbb{E}_{z,c}\big[J_{G_\theta}(z, c)^\top \nabla_x f_{\psi^\star}\big(G_\theta(c, z)\big)\big]\,. \tag{46}$$

If instead $\mathcal{F}$ is the full 1-Lipschitz ball (Wasserstein-1), the above expression defines a subgradient *(with the same leading minus sign) for any measurable selection $f^\star \in \arg\max_{f \in \mathcal{F}} (\mathbb{E}_p f - \mathbb{E}_{q_\theta} f)$.*

Proof. This is Assumption 5 for $G_\theta(c, z)$ with compact $\mathcal{F}$; see also standard IPM/Wasserstein results [2, 3, 72]. □

Theorem 4 (First-Order Stationarity)**.** *Under Assumption 10 and Lemma 4, any stationary point $\theta^\dagger$ (gradient or, for W_1, subgradient) of Equation 45 satisfies*

$$-\lambda\, \mathbb{E}\big[J_G^\top \nabla_x f^\star(G)\big] + 2\alpha\, \mathbb{E}\big[J_G^\top (G - x)\big] + 2\beta\, \mathbb{E}\big[J_G^\top J_\Phi^\top (\Phi(G) - \Phi(x))\big] = 0, \tag{47}$$

with $G = G_{\theta^\dagger}(c, z)$, $J_G = J_{G_{\theta^\dagger}}(c, z)$, J_Φ evaluated at G.

8.3. Off-Manifold Control and "Sharpness Ceiling"

We formalize two key effects of Equation 45: (i) off-manifold suppression via the IPM term and (ii) bounded over-sharpening relative to p ("sharpness ceiling").

Proposition 8 (Off-manifold penalty via W_1 tube bound)**.** *Let $\mathcal{M} = \mathrm{supp}(p)$ and $\mathcal{N}_\alpha(\mathcal{M}) = \{x : \mathrm{dist}(x, \mathcal{M}) \le \alpha\}$. If $q_\theta(\mathcal{N}_\alpha(\mathcal{M})) \le 1 - \mu_\alpha$, then*

$$W_1(p, q_\theta) \;\ge\; \alpha\,\mu_\alpha. \tag{48}$$

Thus any nontrivial mass placed outside an α-tube around the natural-image manifold induces at least $\alpha\mu_\alpha$ Wasserstein cost.

Proof. By the Kantorovich–Rubinstein duality, $W_1(p, q_\theta) = \sup_{\|f\|_{\mathrm{Lip}} \le 1}\{\mathbb{E}_p f - \mathbb{E}_{q_\theta} f\}$. Let $f(x) = \min\{\mathrm{dist}(x, \mathcal{M}), \alpha\}$, which is 1-Lipschitz. Then $f \equiv 0$ on $\mathcal{M}$ and $f(x) \ge \alpha$ for $x \notin \mathcal{N}_\alpha(\mathcal{M})$. Take $g = -f$, also 1-Lipschitz. Then:

$$\mathbb{E}_p g - \mathbb{E}_{q_\theta} g = -\mathbb{E}_p f + \mathbb{E}_{q_\theta} f \ge 0 + \alpha\,\mu_\alpha = \alpha\,\mu_\alpha. \tag{49}$$

Hence $W_1(p, q_\theta) \ge \alpha\,\mu_\alpha$. See [2]. □

Proposition 9 (Sharpness Ceiling via 1-Lipschitz Functionals)**.** *Let $d_{\mathcal{F}} = W_1$ (or more generally suppose there exists $c_{\mathcal{F}} > 0$ such that $d_{\mathcal{F}} \ge c_{\mathcal{F}} W_1$). Let $T : \mathcal{X} \to \mathcal{Y}$ be linear with $\|T\| \le 1$. Then, for $f(x) = \|Tx\|_2$ (1-Lipschitz),*

$$d_{\mathcal{F}}(p, q_\theta) \;\ge\; c_{\mathcal{F}}\,\big|\,\mathbb{E}_p\|Tx\|_2 - \mathbb{E}_{q_\theta}\|T\tilde{x}\|_2\,\big|. \tag{50}$$

Specifically, any deliberate amplification of band-/high-pass components—i.e., applying T as a band-pass filter or gradient—forces an increase in $d_{\mathcal{F}}$, providing a distributional upper bound for over-sharpening.

Proof. Since f is 1-Lipschitz, the Kantorovich–Rubinstein dual implies $W_1(p, q_\theta) \ge \mathbb{E}_p f - \mathbb{E}_{q_\theta} f$ [2]. Applying the same argument with $-f$ establishes the reverse inequality and yields the absolute difference. Using $d_{\mathcal{F}} \ge c_{\mathcal{F}} W_1$ gives Equation 50. □

8.4. Perceptual Anchoring: Deviation Bounds

We next show that, even without explicit data-consistency, the paired pixel/perceptual terms constrain nullspace-like deviations (including directions weakly reflected by H) because they are measured against the *ground-truth* x.

Proposition 10 (Component-wise control by the pixel term)**.** *Let $P : \mathcal{X} \to \mathcal{X}$ be any nonexpansive linear projector ($\|P\| \le 1$). With the α used in Equation 45,*

$$\mathbb{E}\,\|P(G_\theta(c, z) - x)\|_2^2 \;\le\; \frac{1}{\alpha}\,\mathcal{J}(\theta). \tag{51}$$

At any minimizer $\theta^\star$, $\mathbb{E}\,\|P(G_{\theta^\star} - x)\|_2^2 \le \inf_\theta \mathcal{J}(\theta)/\alpha$.

Proof. From Equation 45, one obtains $\mathcal{J}(\theta) \ge \alpha\,\mathbb{E}\|G_\theta - x\|^2 \ge \alpha\,\mathbb{E}\|P(G_\theta - x)\|^2$, where the second inequality follows from the contractive property $\|P\| \le 1$. □

Proposition 11 (Perceptual Alignment Controls Feature-Space Deviations)**.** *For any Φ,*

$$\mathbb{E}\,\|\Phi(G_\theta) - \Phi(x)\|_2^2 \;\le\; \frac{1}{\beta}\,\mathcal{J}(\theta). \tag{52}$$

Moreover, whenever Φ dominates a high-frequency seminorm Π (i.e., $\Pi(u) \le C\,\|\Phi(u)\|_2$ for u in the region of interest), we obtain

$$\begin{aligned}\mathbb{E}\,\Pi\big(G_\theta - x\big) \;&\le\; C\Big(\mathbb{E}\,\|\Phi(G_\theta) - \Phi(x)\|_2^2\Big)^{1/2} \\ &\le\; \frac{C}{\sqrt{\beta}}\,\sqrt{\mathcal{J}(\theta)}.\end{aligned} \tag{53}$$

Hence feature-space alignment upper-bounds perceptual/high-frequency deviations.

8.5. No Over-Generation at Optimum

We combine IPM and pixel/perceptual anchoring to show that stationary solutions cannot produce arbitrarily exaggerated details relative to p and x.

Theorem 5 (Combined Control of Over-Generation)**.** *Assume $d_{\mathcal{F}} = W_1$ (or, more generally, there exists $c_{\mathcal{F}} > 0$ such that $d_{\mathcal{F}} \ge c_{\mathcal{F}} W_1$). Let T be linear with $\|T\| \le 1$, and let P be any projector with $\|P\| \le 1$. Then, for any θ,*

$$\begin{aligned}\big|\,\mathbb{E}_{q_\theta}\|T\tilde{x}\|_2 - \mathbb{E}_p\|Tx\|_2\,\big| \;&\le\; \tfrac{1}{\lambda\, c_{\mathcal{F}}}\,\mathcal{J}(\theta), \\ \mathbb{E}\,\|P(G_\theta - x)\|_2^2 \;&\le\; \tfrac{1}{\alpha}\,\mathcal{J}(\theta),\end{aligned} \tag{54}$$

and

$$\mathbb{E}\,\|\Phi(G_\theta) - \Phi(x)\|_2 \;\le\; \frac{1}{\sqrt{\beta}}\,\sqrt{\mathcal{J}(\theta)}. \tag{55}$$

In particular, at any global minimizer $\theta^\star$, all three deviations are jointly minimized and cannot be simultaneously large. Thus, the adversarial training is biased toward faithful *reconstructions (natural-manifold and feature-aligned), lacking the tendency to over-generate ultra-sharp details.*

8.6. From AR/FM Biases to Adversarial Training Fidelity

Theorem 6 (Fidelity Guarantees for Low-Level Restoration)**.** *Consider $c = Hx + \eta$ under the spectral/regularity assumptions in Sections 7–8.5 of the Appendix, and let $p(x \mid c)$ denote the ground-truth posterior. Suppose generators are trained with: (1) conditional AR MLE (Equation 12); (2) FM Fisher minimization (Equation 13); (3) a composite adversarial training objective (Equation 45).*

1. ***Autoregression (AR).*** *Single-mode conditional factorization together with greedy or vanishing-temperature decoding enforces a deterministic resolution of posterior ambiguity along* $\ker H$*: genuinely distinct posterior modes contract toward midpoints, reducing variability and introducing structured artifacts. Under teacher forcing with test-time rollout, the total-variation discrepancy* can accumulate up to linear order *with sequence length in the worst case (Lemma 1), yielding distortions and over-generation [5, 44, 62, 66].*
2. ***Flow Matching (FM).*** *FM minimizes an integral of conditional Fisher divergences of Gaussian-smoothed posteriors with effective frequency weight:*

$$\widetilde{W}_t(\omega) \propto \|\omega\|^2 e^{-\sigma_t^2\|\omega\|^2}\, \Xi_t(\omega).$$

   Exponential high-frequency damping and null bands where $\Xi_t(\omega)$ *is tiny (including* $|\widehat{H}(\omega)| = 0$*) make FM insensitive to null-space errors, weak on textures/edges, and amplify local drifts into macroscopic warps or hallucinations [23, 45, 46, 70].*
3. ***Adversarial Training.*** *The composite objective enforces: (i) distributional alignment via an IPM critic, suppressing off-manifold mass; (ii) pixel-space anchoring to the ground truth, which bounds all measurable components of* x*; and (iii) perceptual alignment in feature space, which regulates high-frequency discrepancies. At any global minimizer these constraints hold jointly, yielding sharp, faithful reconstructions without uncontrolled over-generation [2, 3, 72].*

The preceding sections have established complementary theoretical limitations of autoregressive (AR) and flow matching (FM) objectives in the low-level restoration setting. For AR, the conditional factorization and single-mode decoding force posterior ambiguities in $\ker(H)$ to collapse deterministically, while exposure bias causes small local mismatches to accumulate into global artifacts and over-generation. For FM, the Fisher-based objective is spectrally reweighted by Gaussian smoothing and the forward operator (through Ξ_t), which underweights high-frequency errors, ignores $\ker(H)$ components, and allows unconditional textures to leak in at early times; small drift errors near edges are then amplified into visible warps. In contrast, the adversarial training formulation explicitly combines three complementary forces: (i) distributional alignment with the natural-image manifold via an IPM critic, (ii) pixel-level fidelity to the ground truth, and (iii) perceptual alignment in feature space. These terms jointly prevent both collapse and uncontrolled hallucination, while maintaining sharpness without exceeding the statistics of p.

Together, these complementary objectives provide a precise characterization of fidelity and stability (Theorem 6).

9. Frequency-Aware Low-Level Instructions

Universal Restoration Instruction r**.**

```
Analyze and return 1 line: Task: <task_token>,
Focus: <high | low>, Rationale: <brief reason>,
Pipeline: <step1 -> step2 -> ...>.
```

Interpretation. The role of r is to constrain the planner's output format and scope: it precludes free-form text, enforces a single categorical decision, ties that decision to an interpretable frequency regime, and requires both a concise rationale and a stepwise restoration plan.

Expert Rule P_{expert}**.**

```
You are an expert image restoration task classifier.
Available tasks (use EXACT lowercase tokens):
deraining | desnowing | dehazing | deblur | denoise
| light_enhancement | super_resolution.
Critical distinctions:
RAIN (deraining)
- Linear streaks: parallel/near-parallel elongated
lines
- Overlay effect: streaks cross sharp edges without
blurring them
- Directional: consistent orientation
(diagonal/vertical)
- Can coexist with low contrast, but streaks are
visible as distinct overlays
SNOW (desnowing)
- Particles: round/irregular white blobs, bokeh
discs
- Size variation: larger near, smaller far
- Random distribution, NOT linear/parallel
HAZE (dehazing)
- Depth-dependent: far objects more degraded than
near
- Milky appearance with desaturation
- Distance gradient clearly visible
BLUR (deblur)
- Edge smearing: boundaries themselves are
widened/soft
- Uniform softness or motion trails
- Object edges lose geometric precision
NOISE (denoise)
- Grain on crisp edges: edge structure intact but
covered by speckles
- Random texture in flat areas
- High-ISO appearance
UNDEREXPOSED (light_enhancement)
- Globally dark, no depth gradient
- Histogram left-biased, shadows crushed
- Color cast possible
LOW RESOLUTION (super_resolution)
- Insufficient spatial sampling: small native
H×W or strong aliasing
- Loss of fine textures; blockiness/jagged edges
when upscaled
- Distinct from blur: edges can appear
jagged/aliased rather than uniformly smeared
Frequency decision rules:
Choose high when degradation is dominated by
fine-scale artifacts or missing detail that lives in
high spatial frequencies:
- deblur (recover sharp edges and textures)
- denoise (suppress noisy high-frequency speckles
while preserving true details)
- deraining / desnowing (remove thin
streaks/particles and recover edge micro-structure)
Choose low when degradation is dominated by
global/slow-varying components:
```

- dehazing (restore low-frequency contrast/airlight; depth-dependent veiling)
- light_enhancement (global exposure/illumination and tone mapping)

If signals suggest mixed conditions, choose the dominant component according to the most visible impairment.

Pipeline rules:
- **DO NOT** use any task names in the Pipeline.

Interpretation. P_{expert} formalizes image restoration as a frequency-aware understanding problem, aligning perceptual degradations with their dominant spectral signatures. It specifies a principled decision boundary: fine-scale losses, rain, snow, noise, blur, map to high-frequency restoration, whereas global, slowly varying degradations, haze, illumination, map to low-frequency correction. By explicitly treating super-resolution as a sampling-limited case, P_{expert} extends the taxonomy beyond semantic labels toward signal-reconstruction logic. Building on this prior, the planner emits, for each instance, an auditable reasoning trace and an executable restoration plan that directly guide the downstream executor.

Label-free Low-level Feature Pool P_{hints}.

Use the following label-free features computed from image pixels to support your decision.
Image size: H, W (pixels). Include scale cues for potential super_resolution.

(1) Rain cues ϕ_{rain}: Grad. orientation: grayscale $g \in [0,1]$, (g_x, g_y), magnitude $m = \sqrt{g_x^2 + g_y^2}$; $\theta = \text{mod}(\arctan 2(g_y, g_x), \pi)$; weighted histogram $h(\theta)$ with weights m. **Scores:** line_score= $\max h / \sum h$, anisotropy= $(\max h - \text{mean}\, h)/(\text{mean}\, h)$. **Spectrum:** power P from FFT of g; annuli (mid/low) by radius; freq_ratio= $\text{mean}(P_{\text{mid}})/(\text{mean}(P_{\text{low}}) + \varepsilon)$.

(2) Snow cues ϕ_{snow}: Blobs: $b = \mathbb{1}[g > 0.78]$; connected components; small_blobs= $\#\{3 \le \text{area} \le 200\}$. **Isotropy:** reuse $h(\theta)$; lower anisotropy $\rightarrow$ more snow-like randomness.

(3) Noise cues ϕ_{noise}: Flat mask: $\Omega_{\text{flat}} = \{m < Q_{0.25}(m)\}$; local mean $\bar{g}$ (box 3×3); residual $r = g - \bar{g}$. **Stats:** noise_mad= $\text{median}(|r - \text{median}(r)|)$. **Chroma:** YCbCr std on Ω_{flat}, chroma_std= $\frac{1}{2}(\text{std(Cb)} + \text{std(Cr)})$. **Score:** noise_score= $0.6\,\sigma(50(\text{noise\_mad} - 0.0050)) + 0.4\,\sigma(50(\text{chroma\_std} - 0.0095))$.

(4) Blur cues ϕ_{blur}: Laplacian var: lapVar= $\text{Var}(L * g)$. **Spectrum ratio:** hf_energy= $\text{mean}(P_{\text{outer}})/(\text{mean}(P_{\text{inner}}) + \varepsilon)$. **Edge strength:** grad95= $Q_{0.95}(\sqrt{g_{x,s}^2 + g_{y,s}^2})$ with mild smoothing.

(5) Haze cues ϕ_{haze}: Dark channel: dark_mean= $\text{mean}(\min(R, G, B))$; **Saturation:** sat_mean= $\text{mean}(S)$ in HSV. **Depth proxy:** split top/bottom; depth_grad= $\text{mean}(Y_{\text{top}}) - \text{mean}(Y_{\text{bot}})$. **Composite:** haze_score= $0.4\,\sigma(7(\text{dark\_mean} - 0.33)) + 0.3\,\sigma(7(0.30 - \text{sat\_mean})) + 0.3\,\sigma(8(\text{depth\_grad} - 0.03))$.

(6) Exposure cues ϕ_{expo}: Luma stats: meanY= $\text{mean}(Y)$, p50= $Q_{0.50}(Y)$; underexposed if meanY < 0.32 or p50 < 0.26.

Recommended thresholds (for disambiguation): Rain: line_score> 0.16, anisotropy> 0.40, freq_ratio> 1.05; **Snow:** small_blobs> 25, anisotropy< 0.42. **Noise:** noise_score> 0.45; **Blur:** grad95< 0.17, lapVar< 0.27, hf_energy< 0.052. **Haze:** haze_score> 0.50, depth_grad> 0.03; **Underexp:** meanY< 0.32 or p50< 0.26.

Example (auto-filled at runtime): Rain: line=0.21, aniso=0.45, freq=1.08; Snow: blobs=31, aniso=0.38; Noise: mad=0.0061, chroma=0.0083, score=0.47; Blur: lapVar=0.256, hf=0.055, grad95=0.174; Haze: score=0.42, depth_grad=0.028, dark_mean=0.36, sat_mean=0.29; Exposure: meanY=0.31, p50=0.27; Size: H=480, W=320.

Interpretation. P_{hints} grounds each decision in measurable, pixel-level evidence, making the planner's reasoning traceable and reproducible. Gradient orientation, spectrum energy ratios, and brightness statistics are mapped to specific degradation cues, translating physical image phenomena into quantitative signals. High-frequency degradations (rain, snow, noise, blur) and low-frequency ones (haze, exposure) are thus separated by data-driven thresholds, while image size provides an explicit trigger for super-resolution. This turns frequency focus from a heuristic label into a verifiable diagnostic built on interpretable computational cues.

10. Detailed Metrics

To complement the main quantitative comparison in Table 1 and Table 2, we provide full per-benchmark results in Table 5 and Table 6 of the appendix. These tables report all five distortion–perception metrics (PSNR, SSIM, LPIPS, FID, DISTS) for each individual benchmark across seven restoration tasks (deraining, desnowing, dehazing, deblurring, denoising, low-light enhancement, and super-resolution), together with a unified comparison against recent AIO-IR methods [15, 34, 58, 78, 109] and SR baselines [41, 83, 92, 97, 98, 103]. Overall, FAPE-IR attains state-of-the-art or comparable performance on the majority of benchmarks, especially on weather-related degradations and challenging real-world benchmarks, where it substantially improves both PSNR/SSIM and perceptual metrics.

In the high-frequency–dominant restoration regimes (e.g., OutDoor, RainDrop, Snow100K-L/S, Rain100_H, GoPro-gamma, RealBlur-J/R, Urban100-15/25/50), our method consistently attains markedly lower LPIPS and DISTS together with competitive or clearly higher PSNR/-FID than prior AIO-IR approaches, indicating that the frequency-aware planner and band-specialized LoRA-MoE executor effectively suppress artifacts while preserving fine structures. In contrast, for low-frequency degradations such as URHI, ITS-val, and LOL-v1/v2, FAPE-IR achieves strong gains in FID and DISTS while maintaining high SSIM, reflecting better global contrast and color consistency under severe haze and illumination changes. A remaining challenge lies in Rain100_L and the BSD68-15/25/50 denoising benchmarks, where FAPE-IR is less competitive in PSNR. We attribute this mainly to a mismatch between our training distribution and these relatively simple degradations (light rain and synthetic Gaussian noise), together with the limited amount of pure denoising data: the planner tends to favor stronger high-frequency

Table 5. Unified quantitative comparison across all benchmarks. Our method (**FAPE-IR**) consistently achieves state-of-the-art or comparable performance across diverse restoration benchmarks.

Method	OutDoor					RainDrop					Rain100_L					Rain100_H				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
State-of-the-art AIO-IR methods																				
PromptIR [58]	18.40	0.67	0.41	154.96	0.30	23.48	0.80	0.18	61.42	0.12	37.41	0.98	0.02	8.84	0.03	15.93	0.52	0.46	175.42	0.27
FoundIR [34]	17.07	0.65	0.45	156.17	0.30	23.52	0.80	0.21	63.57	0.14	30.62	0.93	0.11	34.19	0.08	13.77	0.43	0.55	219.48	0.33
DFPIR [78]	14.06	0.60	0.50	176.41	0.38	22.52	0.79	0.20	64.91	0.14	37.99	0.98	0.02	8.25	0.03	16.07	0.54	0.44	166.89	0.26
MoCE-IR [109]	17.99	0.67	0.41	154.87	0.30	23.30	0.80	0.18	61.89	0.12	38.05	0.98	0.02	7.90	0.02	15.33	0.49	0.48	180.16	0.28
AdaIR [15]	18.23	0.67	0.41	155.67	0.30	23.37	0.80	0.18	61.89	0.12	38.00	0.98	0.02	7.58	0.02	15.93	0.52	0.46	175.01	0.27
Ours	28.16	0.83	0.09	25.16	0.07	25.83	0.80	0.11	20.86	0.11	33.18	0.93	0.04	10.21	0.03	27.01	0.82	0.13	33.58	0.09

Method	BSD68-15					BSD68-25					BSD68-50					Urban100-15				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
State-of-the-art AIO-IR methods																				
PromptIR [58]	28.60	0.85	0.20	49.74	0.14	27.15	0.72	0.32	86.65	0.21	23.52	0.48	0.52	164.09	0.33	38.98	0.98	0.01	5.28	0.05
FoundIR [34]	33.87	0.88	0.17	49.61	0.12	29.81	0.75	0.29	85.98	0.19	23.99	0.50	0.50	160.78	0.31	36.90	0.97	0.03	10.50	0.07
DFPIR [78]	30.78	0.88	0.15	45.11	0.11	28.41	0.75	0.28	81.08	0.19	23.42	0.49	0.50	161.16	0.31	37.80	0.97	0.03	11.79	0.08
MoCE-IR [109]	29.78	0.84	0.20	49.93	0.14	27.11	0.71	0.32	85.49	0.21	23.32	0.48	0.52	163.59	0.32	39.19	0.98	0.01	6.16	0.05
AdaIR [15]	29.58	0.84	0.20	49.85	0.13	27.51	0.71	0.32	85.31	0.21	23.34	0.48	0.52	164.68	0.33	38.94	0.98	0.02	6.11	0.05
Ours	34.21	0.91	0.09	26.94	0.07	31.79	0.85	0.15	44.19	0.11	27.87	0.72	0.27	76.01	0.17	31.09	0.93	0.02	7.45	0.05

Method	Urban100-25					Urban100-50					ITS-val					URHI				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
State-of-the-art AIO-IR methods																				
PromptIR [58]	35.36	0.96	0.02	9.37	0.06	29.40	0.89	0.08	31.77	0.12	21.51	0.89	0.15	37.38	0.11	26.18	0.95	0.05	11.30	0.05
FoundIR [34]	32.72	0.90	0.09	27.28	0.12	26.29	0.72	0.30	68.30	0.22	13.32	0.74	0.33	49.35	0.22	16.99	0.84	0.17	19.74	0.14
DFPIR [78]	34.35	0.94	0.05	21.67	0.10	29.52	0.84	0.13	49.97	0.17	12.42	0.71	0.32	47.00	0.23	24.64	0.94	0.06	11.97	0.05
MoCE-IR [109]	35.78	0.97	0.02	10.39	0.07	29.94	0.92	0.06	24.88	0.11	14.71	0.79	0.24	43.64	0.16	23.76	0.93	0.07	12.32	0.06
AdaIR [15]	35.53	0.96	0.02	9.14	0.06	29.36	0.89	0.09	33.24	0.13	19.44	0.87	0.17	39.66	0.13	24.46	0.94	0.06	11.83	0.05
Ours	30.17	0.91	0.03	9.93	0.06	27.82	0.86	0.06	17.44	0.09	34.07	0.97	0.04	6.11	0.04	31.62	0.96	0.04	8.31	0.04

Method	Snow100K-L					Snow100K-S					GoPro					GoPro-gamma				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
State-of-the-art AIO-IR methods																				
AdaIR [15]	21.03	0.70	0.32	38.69	0.19	27.40	0.84	0.21	14.87	0.12	29.03	0.87	0.18	24.94	0.12	27.96	0.86	0.19	28.49	0.12
FoundIR [34]	20.89	0.73	0.33	39.96	0.20	26.72	0.84	0.23	18.99	0.15	27.51	0.81	0.24	39.35	0.16	27.21	0.81	0.24	41.79	0.16
DFPIR [78]	18.41	0.67	0.35	48.50	0.21	23.18	0.81	0.25	21.14	0.15	30.09	0.89	0.16	21.59	0.10	28.84	0.88	0.17	25.59	0.11
MoCE-IR [109]	21.00	0.69	0.33	43.21	0.20	26.67	0.83	0.23	17.74	0.13	29.56	0.90	0.16	20.10	0.10	27.93	0.87	0.18	25.85	0.11
Ours	29.07	0.85	0.10	1.88	0.07	31.52	0.90	0.06	1.10	0.05	28.13	0.84	0.13	16.04	0.08	28.00	0.84	0.13	17.13	0.08

Method	RealBlur-J					RealBlur-R					LOL-v2					LOL-v1				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
State-of-the-art AIO-IR methods																				
AdaIR [15]	17.74	0.71	0.28	51.66	0.18	12.54	0.51	0.51	93.48	0.36	24.48	0.86	0.18	48.78	0.13	24.93	0.92	0.12	52.78	0.11
FoundIR [34]	28.27	0.85	0.19	40.88	0.15	30.66	0.93	0.15	46.12	0.18	14.44	0.66	0.33	77.47	0.23	14.54	0.75	0.28	95.49	0.21
DFPIR [78]	28.75	0.87	0.16	29.58	0.12	36.00	0.96	0.10	36.17	0.14	25.92	0.90	0.16	50.62	0.13	25.95	0.92	0.12	53.66	0.11
MoCE-IR [109]	15.76	0.68	0.30	52.80	0.19	11.55	0.48	0.53	96.66	0.37	23.25	0.89	0.16	44.31	0.12	24.93	0.92	0.11	44.90	0.09
Ours	30.56	0.87	0.10	18.21	0.07	37.77	0.97	0.05	14.74	0.07	25.07	0.90	0.11	32.94	0.09	28.95	0.92	0.11	47.67	0.09

Method	RealSR 2×					DrealSR 2×					RealSR 4×					DrealSR 4×				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
State-of-the-art SR methods																				
StableSR [83]	24.22	0.75	0.23	81.72	0.19	25.71	0.75	0.26	91.11	0.18	23.11	0.68	0.30	127.06	0.22	27.63	0.73	0.35	140.86	0.23
DiffBIR [41]	26.31	0.71	0.30	80.60	0.21	27.31	0.70	0.37	103.44	0.24	23.75	0.62	0.37	131.16	0.24	25.49	0.57	0.52	182.95	0.31
SeeSR [98]	25.90	0.75	0.28	100.75	0.22	28.05	0.77	0.30	107.08	0.23	24.05	0.69	0.32	128.56	0.24	28.78	0.77	0.34	152.01	0.25
PASD [103]	26.19	0.76	0.24	92.31	0.19	27.96	0.77	0.27	107.73	0.20	24.69	0.71	0.31	131.09	0.21	28.97	0.78	0.33	150.68	0.22
OSEDiff [97]	25.56	0.76	0.25	100.26	0.20	27.37	0.79	0.27	114.31	0.20	24.51	0.72	0.30	124.37	0.21	28.84	0.79	0.31	131.27	0.22
PURE [92]	23.59	0.65	0.32	83.81	0.22	25.84	0.68	0.37	108.30	0.23	22.20	0.59	0.39	127.72	0.25	26.53	0.64	0.44	158.88	0.26
Ours	29.99	0.88	0.13	51.05	0.12	30.52	0.89	0.14	60.30	0.12	25.55	0.78	0.24	105.75	0.19	28.42	0.84	0.25	126.45	0.19

suppression, which can lead to slight over-smoothing and thus lower PSNR on texture-rich scenes, while still preserving favorable perceptual metrics.

Table 6 further reports no-reference image quality metrics on the same set of benchmarks. Across most restoration benchmarks (excluding SR), FAPE-IR improves or matches the best scores on at least three of the five IQA indicators, confirming that our adversarial training and frequency regularization translate into reconstructions that are also preferred by hand-crafted NR-IQA measures. For real-world super-resolution, however, FAPE-IR exhibits a different pattern: while Table 5 shows clear advantages in full-reference distortion and perceptual metrics (PSNR/SSIM, LPIPS, FID, DISTS), the no-reference scores in Table 6 are sometimes worse than those of competing SR methods. We speculate that this discrepancy stems from the SR ground-truth images themselves, whose statistics yield relatively poor scores under standard NR-IQA metrics. By recovering textures and statistics closer to these ground truths, FAPE-IR improves full-reference and learned perceptual metrics but can be penalized by hand-crafted no-reference indicators. Overall, the tables demonstrate that FAPE-IR maintains robust distortion–perception trade-offs across diverse degradations and benchmarks, while also highlighting the importance of higher-quality SR training data and more reliable NR-IQA models for future work, especially on challenging benchmarks such as Urban100 and SR benchmarks.

11. Other Visualizations and FM Analysis

In this section, we provide additional qualitative results that are not included in the main paper due to space limitations.

Table 6. Unified quantitative comparison across all benchmarks. Our method (**FAPE-IR**) consistently achieves state-of-the-art or comparable performance across diverse restoration benchmarks.

Method	OutDoor					RainDrop					Rain100_L					Rain100_H				
	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑
State-of-the-art AIO-IR methods																				
PromptIR [58]	4.59	69.40	0.68	0.69	0.50	6.68	61.27	0.58	0.45	0.39	7.18	59.42	0.58	0.31	0.46	3.16	65.43	0.66	0.46	0.51
FoundIR [34]	4.76	68.41	0.67	0.63	0.48	8.48	60.45	0.57	0.38	0.41	6.32	61.27	0.57	0.32	0.40	3.65	67.02	0.66	0.42	0.50
DFPIR [78]	4.60	69.64	0.68	0.70	0.50	6.38	61.81	0.59	0.46	0.40	6.87	57.64	0.52	0.28	0.40	3.17	64.49	0.65	0.42	0.50
MoCE-IR [109]	4.59	69.74	0.68	0.70	0.50	7.20	62.47	0.59	0.46	0.41	7.11	58.90	0.58	0.29	0.47	3.21	65.56	0.66	0.45	0.52
AdaIR [15]	4.60	69.67	0.68	0.70	0.50	6.89	61.96	0.59	0.45	0.40	7.19	58.95	0.58	0.29	0.48	3.15	65.35	0.66	0.49	0.51
Ours	5.04	69.34	0.68	0.67	0.48	5.04	66.57	0.64	0.61	0.45	3.42	68.26	0.68	0.39	0.58	2.99	69.12	0.70	0.46	0.55

Method	BSD68-15					BSD68-25					BSD68-50					Urban100-15				
	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑
State-of-the-art AIO-IR methods																				
PromptIR [58]	5.25	45.68	0.57	0.55	0.33	5.71	39.01	0.54	0.47	0.30	5.33	72.86	0.52	0.38	0.28	5.33	72.86	0.75	0.62	0.73
FoundIR [34]	5.34	46.89	0.57	0.58	0.34	5.80	40.08	0.55	0.50	0.31	7.32	37.11	0.52	0.43	0.30	4.25	71.22	0.72	0.61	0.67
DFPIR [78]	5.27	48.63	0.57	0.60	0.35	5.66	42.49	0.55	0.54	0.33	7.05	39.04	0.52	0.45	0.32	5.57	73.52	0.76	0.67	0.73
MoCE-IR [109]	5.38	46.03	0.57	0.55	0.33	5.80	39.04	0.55	0.47	0.30	7.12	35.44	0.52	0.39	0.28	5.46	73.08	0.75	0.63	0.74
AdaIR [15]	5.30	45.57	0.57	0.56	0.33	5.76	38.84	0.55	0.46	0.30	7.11	35.16	0.52	0.39	0.28	5.20	72.78	0.75	0.63	0.73
Ours	5.13	51.60	0.58	0.64	0.36	4.87	49.24	0.56	0.50	0.34	4.64	44.90	0.74	0.64	0.70	4.90	72.17	0.74	0.64	0.70

Method	Urban100-25					Urban100-50					ITS-val					URHI				
	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑
State-of-the-art AIO-IR methods																				
PromptIR [58]	4.79	72.39	0.74	0.61	0.72	4.11	70.23	0.68	0.56	0.68	4.53	51.32	0.50	0.10	0.35	4.21	54.28	0.64	0.27	0.35
FoundIR [34]	3.86	66.26	0.69	0.58	0.61	4.34	55.79	0.63	0.54	0.50	5.65	49.82	0.50	0.14	0.36	5.16	54.44	0.63	0.30	0.35
DFPIR [78]	5.86	73.15	0.76	0.64	0.69	7.13	69.97	0.71	0.58	0.56	5.10	46.81	0.48	0.11	0.34	4.16	54.11	0.63	0.24	0.35
MoCE-IR [109]	5.54	73.36	0.75	0.60	0.73	4.80	72.46	0.73	0.56	0.69	5.01	49.54	0.49	0.11	0.36	4.24	54.34	0.63	0.27	0.36
AdaIR [15]	5.01	72.75	0.75	0.62	0.72	4.13	69.45	0.68	0.58	0.67	4.65	50.44	0.49	0.10	0.34	4.22	54.20	0.63	0.28	0.35
Ours	4.92	71.90	0.73	0.62	0.69	4.96	70.73	0.70	0.57	0.66	5.66	47.88	0.55	0.13	0.30	4.47	53.76	0.64	0.23	0.35

Method	Snow100K-L					Snow100K-S					GoPro					GoPro-gamma				
	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑
State-of-the-art AIO-IR methods																				
AdaIR [15]	3.61	55.64	0.63	0.42	0.42	3.56	60.12	0.66	0.47	0.46	4.86	46.63	0.57	0.22	0.31	4.90	45.67	0.56	0.21	0.30
FoundIR [34]	4.53	57.99	0.64	0.39	0.42	4.56	62.57	0.67	0.43	0.46	5.24	40.86	0.49	0.22	0.26	5.25	41.06	0.50	0.22	0.26
DFPIR [78]	4.09	54.86	0.62	0.39	0.41	3.99	59.47	0.65	0.43	0.44	4.73	50.62	0.60	0.23	0.34	4.79	49.54	0.59	0.22	0.33
MoCE-IR [109]	3.85	55.91	0.63	0.38	0.42	3.76	60.54	0.66	0.43	0.46	4.61	51.14	0.60	0.25	0.35	4.70	48.42	0.58	0.23	0.33
Ours	3.66	62.51	0.66	0.41	0.45	3.55	63.57	0.67	0.43	0.47	4.33	53.83	0.61	0.24	0.39	4.32	53.91	0.60	0.24	0.40

Method	RealBlur-J					RealBlur-R					LOL-v2					LOL-v1				
	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑
State-of-the-art AIO-IR methods																				
AdaIR [15]	5.17	42.85	0.53	0.22	0.29	5.72	41.24	0.50	0.20	0.27	4.27	63.12	0.64	0.41	0.51	4.35	70.06	0.63	0.38	0.58
FoundIR [34]	5.56	41.41	0.53	0.21	0.28	8.28	29.52	0.50	0.24	0.22	5.15	56.67	0.65	0.44	0.44	5.72	65.41	0.66	0.38	0.51
DFPIR [78]	5.45	48.50	0.58	0.24	0.33	8.33	28.57	0.52	0.19	0.22	4.33	64.22	0.64	0.39	0.52	4.55	69.37	0.62	0.36	0.57
MoCE-IR [109]	5.23	44.07	0.53	0.23	0.30	5.99	43.54	0.51	0.23	0.29	4.27	64.95	0.65	0.43	0.52	4.47	71.48	0.64	0.42	0.60
Ours	4.92	52.27	0.61	0.24	0.39	6.71	32.52	0.57	0.22	0.26	4.74	65.31	0.66	0.40	0.50	4.92	69.84	0.64	0.39	0.59

Method	RealSR 2×					DrealSR 2×					RealSR 4×					DrealSR 4×				
	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	TOPIQ↑
State-of-the-art SR methods																				
StableSR [83]	6.62	63.20	0.63	0.63	0.51	6.58	60.31	0.60	0.63	0.51	5.86	58.56	0.57	0.58	0.47	6.91	51.74	0.52	0.58	0.45
DiffBIR [41]	5.97	69.44	0.66	0.70	0.68	5.81	66.79	0.63	0.70	0.67	5.60	69.55	0.65	0.71	0.68	7.00	65.88	0.61	0.71	0.66
SeeSR [98]	5.71	71.46	0.67	0.72	0.72	6.05	67.93	0.65	0.70	0.70	5.47	70.43	0.65	0.70	0.71	6.36	64.96	0.60	0.69	0.67
PASD [103]	5.12	67.67	0.63	0.62	0.61	5.66	65.53	0.61	0.65	0.63	5.21	64.63	0.58	0.58	0.58	6.94	57.85	0.52	0.57	0.55
OSEDiff [97]	5.81	70.55	0.66	0.69	0.64	6.02	67.61	0.63	0.69	0.63	5.74	69.14	0.63	0.67	0.63	6.78	62.68	0.58	0.69	0.59
PURE [92]	5.34	69.47	0.66	0.72	0.65	6.18	65.95	0.62	0.70	0.63	5.77	66.84	0.62	0.69	0.61	7.11	60.14	0.57	0.67	0.58
Ours	7.05	51.94	0.53	0.35	0.37	7.67	47.74	0.50	0.34	0.37	7.63	52.65	0.48	0.39	0.40	8.79	46.53	0.44	0.45	0.41

These visualizations cover a broader range of degradations and benchmarks, further illustrating the effectiveness and generalization ability of our FAPE-IR. For clarity, we group the results into two parts: (1) supplementary qualitative comparisons, and (2) visualizations derived from our flow-matching training process.

11.1. More Qualitative Comparisons

Figures 10–11 present additional qualitative comparisons on deraining, desnowing, deblurring, dehazing, denoising, low-light enhancement, and real-world super-resolution. Figure 10 compares FAPE-IR with recent unified models, while Figure 11 focuses on strong AIO-IR baselines. Across all tasks, FAPE-IR introduces fewer artifacts, better respects the input content, and more faithfully preserves high-frequency details than both categories of baselines.

- **Rainy scenes (deraining):** Unified models often hallucinate textures or alter scene layouts, and AIO-IR methods tend to leave residual streaks or over-smooth details. In contrast, FAPE-IR suppresses rain streaks more thoroughly while retaining local texture contrast, consistent with the improvements in LPIPS and DISTS.
- **Snow scenes (desnowing):** Competing methods either fail to fully remove veiling snow or over-smooth the background. FAPE-IR removes both small particles and large translucent flakes while maintaining the underlying structures of objects such as buildings and vegetation.
- **Denoising:** Under heavy Gaussian noise, unified models may introduce unnatural textures, whereas AIO-IR baselines can blur fine patterns. FAPE-IR better preserves sharp edges and regular patterns (e.g., window grids) with fewer color shifts and blotchy artifacts.
- **Real-world blur (deblurring):** The proposed planner more effectively separates structural edges from noise-like blur, leading to sharper reconstructions with substantially fewer ringing and overshoot artifacts than both unified models and AIO-IR methods.
- **Dehazing and low-light images:** For hazy and underexposed scenes, baseline methods often exhibit residual haze, elevated noise, or strong color bias. FAPE-IR produces smoother illumination transitions, more balanced global contrast, and more natural color tones.

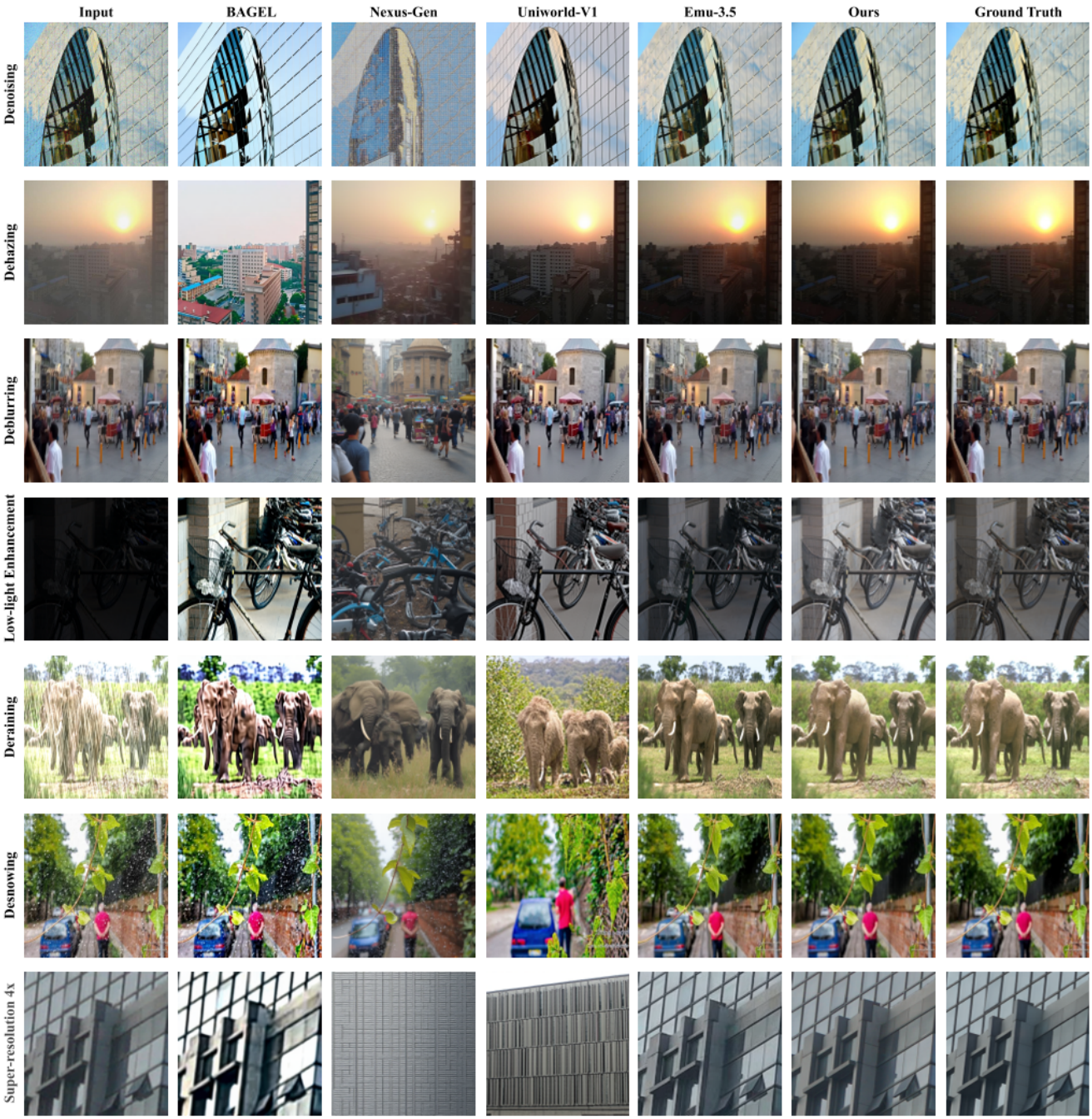

Figure 10. Comparison among unified models, including BAGEL [17], Nexus-Gen [111], Uniworld-V1 [40], and Emu3.5 [16].

- **Super-resolution:** On real-world $\times 4$ super-resolution, unified models sometimes hallucinate high-frequency patterns that deviate from the ground truth, while AIO-IR methods tend to over-smooth repetitive structures. FAPE-IR reconstructs sharper, more regular textures (e.g., building facades) that better match the ground-truth distribution, despite the NR-IQA metrics being challenging and sometimes misaligned with human perception.

These additional qualitative comparisons further corroborate the distortion–perception trade-offs discussed in the main paper and illustrate that FAPE-IR scales more reliably than both unified models and existing AIO-IR approaches across diverse degradation types.

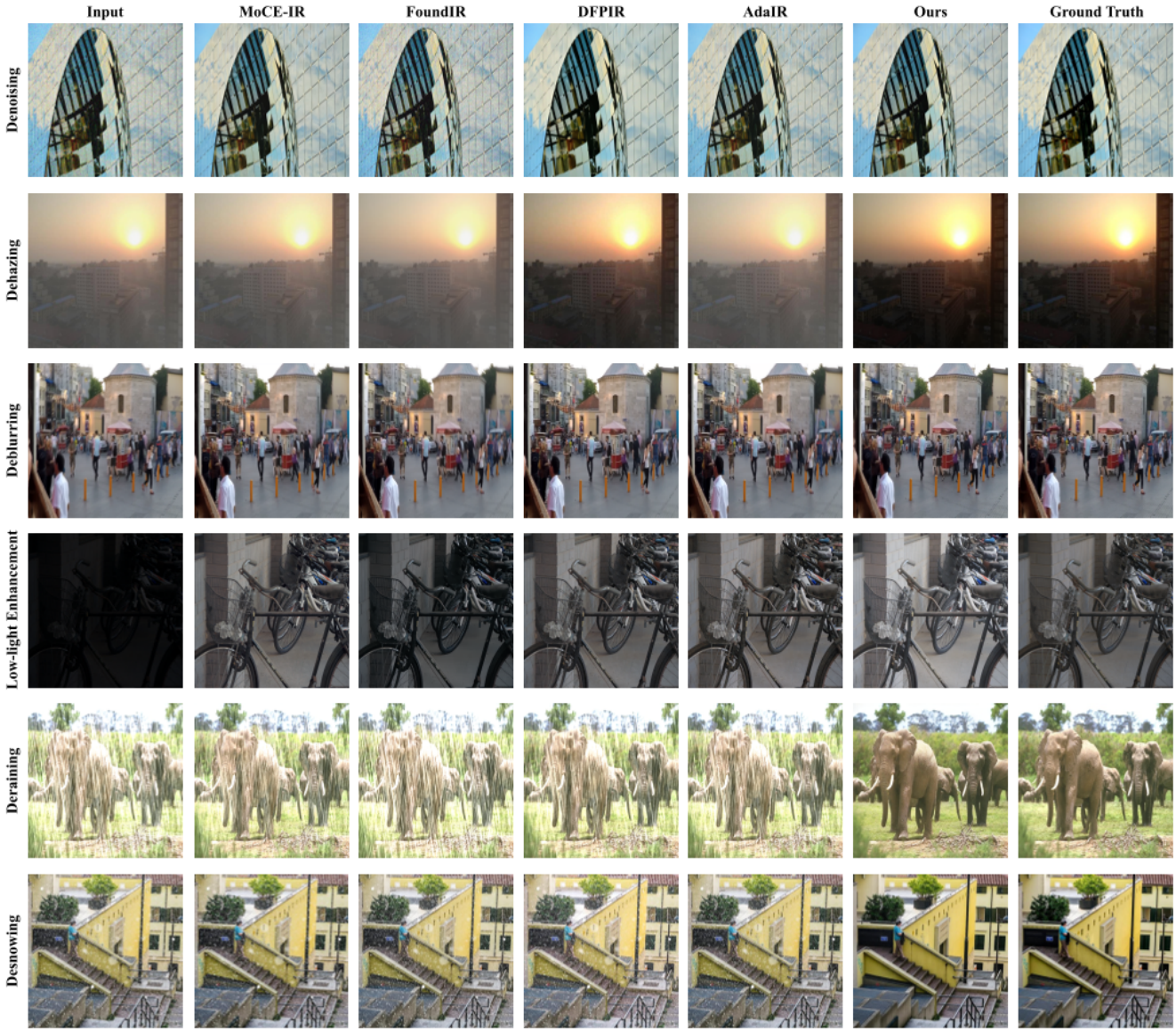

Figure 11. Qualitative comparison of restoration results produced by FAPE-IR and state-of-the-art AIO-IR models.

11.2. Flow-Matching Visualizations

In our early exploratory stage, we attempted to train the entire FAPE-IR framework using a standard flow-matching (FM) objective [45]. Figure 12 shows representative qualitative results on real-world super-resolution benchmarks. While the FM-trained variant is able to remove part of the degradation and produce sharper outputs than the input LR images, it also generates a large number of artifacts and unrealistic details: building facades exhibit irregular, "painted" textures, edges become locally distorted, and fine patterns are often hallucinated rather than faithfully reconstructed from the input. These failure cases provide an empirical counterexample to the naive expectation that FM alone is sufficient for all-in-one restoration in pixel space. In ill-posed tasks such as real-world SR, the learned flow tends to overfit the training distribution and prioritize distribution matching over content preservation, leading to spurious high-frequency components that hurt perceptual realism. This observation is consistent with our theoretical motivation, and it prompted us to explore additional adversarial training and frequency-aware regularization to better constrain the planner–executor pipeline.

Although this FM-based variant is discarded in our final system, we include it here for completeness, given the widespread use of flow-matching in recent unified generative models. In future work, we plan to investigate ways to mitigate these artifacts, e.g., by incorporating stronger structural priors or geometry-aware constraints into the flow

Figure 12. Qualitative results of training our framework with a standard flow-matching (FM) objective on real-world super-resolution. Although the FM-trained variant can sharpen some structures, it also introduces severe artifacts and unrealistic high-frequency details (e.g., distorted edges and hallucinated textures), which motivates our final design choices for FAPE-IR.

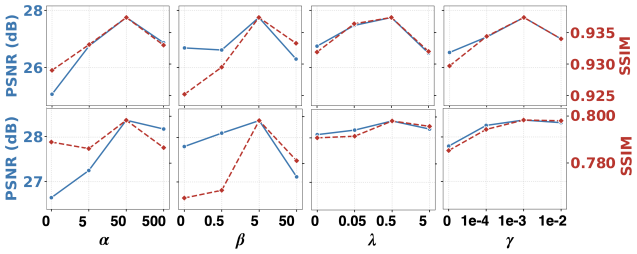

Figure 13. Hyperparameter sensitivity analysis of the four loss weights α, β, λ, and γ on URHI and BSD68-15.

field, so that FM can be more safely integrated into general-purpose restoration frameworks like FAPE-IR.

11.3. Hyperparameter Sensitivity Analysis

To further validate the robustness of our objective design, we provide a more detailed hyperparameter sensitivity analysis for the four loss weights, namely α, β, λ, and γ. Specifically, for each loss term, we vary its weight to $\{0\times, 0.1\times, 1\times, 10\times\}$ relative to the default setting, while keeping all remaining loss weights fixed. For efficiency, all variants are trained for 5k iterations under the same training protocol. Figure 13 reports the corresponding results on two representative benchmarks, i.e., URHI and BSD68-15, in terms of both PSNR and SSIM. Overall, the results show a clear and consistent trend: each loss term contributes positively to the final performance, and removing any of them (i.e., setting the corresponding weight to 0) leads to noticeable degradation. This confirms that the four components are complementary rather than redundant.

11.4. Ablation on High-Frequency Restoration

In the main text, we present the structure ablation results on the low-frequency restoration setting. To provide a more complete picture, we further report the corresponding ablation results on a high-frequency benchmark, i.e., BSD68-15, in Table 7. The overall trend is consistent with the observations in the main text. First, introducing Qwen2.5-VL already brings a clear improvement over the base model, indicating that semantic guidance from the vision-language planner is beneficial not only for low-frequency recovery but also for more challenging high-frequency restoration. Second, enabling the frequency-aware text router (Freq-U) further improves performance, showing that frequency-conditioned text routing helps better align semantic priors with the restoration process. Third, adding the FIR spectral router (Freq-G) leads to the largest gain, which confirms the importance of explicit frequency-aware expert selection in handling complex high-frequency degradations.

Table 7. Ablation on BSD68-15 benchmark. Qwen: remove Qwen2.5-VL; Freq-U: remove frequency-aware text router; Freq-G: remove FIR spectral router; r: LoRA rank in expert combos.

Qwen	Freq-U	Freq-G	r=4	r=8	r=16	PSNR↑	SSIM↑
✗	✗	✗		✓		30.81	0.90
✓	✗	✗		✓		31.19	0.90
+ Freq-U enabled							
✓	✓	✗		✓		31.27	0.90
✓	✓	✗		✓✓		31.69	0.90
✓	✓	✗	✓		✓	31.55	0.91
✓	✓	✗		✓	✓	32.41	0.91
+ Freq-G enabled							
✓	✓	✓		✓✓		33.57	0.91
✓	✓	✓	✓		✓	33.12	0.91
✓	✓	✓		✓	✓	33.20	0.91

We also compare different LoRA rank combinations. Among all tested settings, the combination centered on $r = 8$ achieves the best overall performance, while configurations involving $r = 4$ or $r = 16$ are generally less effective. This suggests that a moderate LoRA rank provides a better trade-off between capacity and specialization for expert routing. Overall, these supplementary results on BSD68-15 further validate that each proposed component contributes positively and that the conclusions drawn from the low-frequency setting generalize well to the high-frequency restoration scenario.