PRIMAL-DUAL CONSTRAINED DIFFUSION SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many applications call for generative models to provide samples that satisfy requirements, including target semantic or visual features in image generation or observation consistency for inverse problems. Although diffusion models excel at capturing complex distributions, their stochastic, iterative nature makes it difficult to impose such requirements. Typical approaches, known as *guidance*, tend to *encourage rather than enforce* desired properties and are often sensitive to hyperparameters such as guidance weights and step size schedules. Moreover, they typically incur substantial memory and computation costs due to retraining or backpropagating through the diffusion model, which limits the type and number of requirements they can impose. This work addresses these challenges by casting constrained diffusion sampling as an optimization problem in which each requirement is described as a generalized moment constraint. Using duality, it shows that the solution to this problem is achieved by a reverse dynamic akin to a parametrized Doob's h-transform of the original diffusion process, whose parameters are the problem's Lagrange multipliers. It then introduces Primal-Dual Constrained Diffusion Sampling (PD-CDS), a training-free algorithm that automatically adjusts these parameters throughout the diffusion, and derives convergence guarantees. The generality of this approach is illustrated by inverse problems, semantic control tasks, and visual property constraints, separately and combined. By precluding backpropagation through the diffusion model, PD-CDS is able to match and even outperform existing methods while reducing computation and memory costs.

1 INTRODUCTION

Denoising diffusion models (DDMs) have emerged as powerful tools for sampling complex, high-dimensional distributions using learned diffusion processes (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021a; Karras et al., 2022). Many applications, however, call for samples to satisfy additional requirements, such as semantic/statistical properties (e.g., (Dhariwal & Nichol, 2021; Ho & Salimans, 2022)) or observation consistency when DDMs are used as priors for inverse problems (e.g., (Chung et al., 2023; Oliviero-Durmus et al., 2025)). A variety of *guidance* methods have been proposed to achieve this purpose, based on conditioning (e.g., (Chung et al., 2023; Song et al., 2023a; Moufad et al., 2025)), optimization techniques (e.g., (Lugmayr et al., 2022; Zampini et al., 2025)), and loss- or energy-based mechanisms (e.g., (Song et al., 2023b; Ye et al., 2024)).

Though successful in many applications, current guidance approaches exacerbate a key challenge with DDMs: inference time. Indeed, they can significantly increase computational and memory costs due to their need for solving inner optimization problems (e.g., (Moufad et al., 2025; Bansal et al., 2024a)) or backpropagating through the DDM (e.g., (Chung et al., 2023; Song et al., 2023b; Bansal et al., 2024a)). Approaches that involve retraining DDMs, such as classifier-free guidance (Ho & Salimans, 2022), also make it cumbersome to add or modify generation requirements. Despite this increased complexity, most guidance methods *encourage rather than enforce* the desired conditions, providing little quantitative requirement satisfaction guarantee. This is particularly critical since they often rely on hyperparameters such as guidance weights and step sizes whose values are task-dependent and greatly affect performance (Ye et al., 2024). This issue is exacerbated when composing heterogeneous requirements, such as trying to generate a bright, high-contrast image of a smiling older person. Hyperparameters must then be jointly tuned, leading to an exponential increase in possible configurations. Lower complexity approaches, such as those based on projections, are also restricted in the type and number of constraints they can impose (Lugmayr et al., 2022; Christopher et al., 2024). These challenges effectively limit the applicability of these methods.

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.

To overcome these issues, this work casts guided diffusion as a constrained optimization problem, setting out to sample from the closest distribution to the one implied by a pretrained DDM that satisfies a set of generalized moment constraints (Sec. 3.1). It then uses duality to show that the solution of this problem can be achieved by a *parametrized* Doob's h-transform of the original diffusion process, whose optimal parameters are the problem's Lagrange multipliers (Sec. 3.2). In particular, this provides a principled justification for the use of "guidance weights" and provides an automatic way of determining their values. Building on these results, it introduces Primal-Dual Constrained Diffusion Sampling (PD-CDS), a training-free sampling algorithm that automatically determines the guidance parameters without backpropagating through the DDM (Sec. 4). Finally, it establishes convergence guarantees for PD-CDS (Thm. 4.4) and showcases its performance in a myriad of applications involving solving inverse problems, controlling semantic features, and imposing visual properties as well as their combinations (Sec. 5). With PD-CDS, it becomes straightforward to generate bright, high-contrast images of smiling older individuals (Fig. 1).

2 PRELIMINARIES

2.1 DIFFUSION MODEL

DDMs generate samples $X_0 \sim \pi_0$, a distribution over $\mathbb{R}^d$, by simulating a Markov chain $(X_t)_{t=T}^0$ backward in time starting from $X_T \sim \mathcal{N}(0, I)$, a fixed standard Gaussian prior, and iteratively applying the transition kernel

$$K_{\theta,t}(\cdot \mid x_t) \triangleq \mathbb{P}[X_{t-1} \in \cdot \mid X_t = x_t] = \mathcal{N}\big(m_{\theta,t}(x_t), \beta_t I\big), \quad t = T, \dots, 1, \tag{1}$$

where $\beta_t \in (0,1)$ is a decreasing sequence defining the *noise schedule* and $m_{\theta,t}(x) = \alpha_t^{-1/2}\big(x + \beta_t s_\theta(x,t)\big)$, for $\alpha_t = 1 - \beta_t$ and a pretrained model s_θ. The model s_θ is known as a *denoiser* because it is typically trained to estimate samples from the target distribution $\tilde{X}_0 \sim \tilde{\pi}_0$ based on

$$\tilde{X}_t = \rho_t \tilde{X}_0 + \sigma_t \varepsilon, \quad \text{for } \varepsilon \sim \mathcal{N}(0, I), \tag{2}$$

where $\rho_t = \sqrt{\bar{\alpha}_t}$, $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$, and $\bar{\alpha}_t = \prod_{j=1}^t \alpha_j$ (Song & Ermon, 2019; Ho et al., 2020). In doing so, $s_\theta(x,t)$ actually approximates the score $\nabla \log \tilde{p}_t$, where $\tilde{p}_t$ denotes the density of the distribution $\tilde{\pi}_t$ of $\tilde{X}_t$. What is more, the marginal distributions π_t of X_t generated by the DDM (1) can be shown to approximate $\tilde{\pi}_t$. Hence, for suitable noise schedules, the DDM output X_0 turns out to be an approximate sample of the target $\tilde{\pi}_0$ (see (Chen et al., 2023; Silveri & Ocello, 2025; Conforti et al., 2025) and App. B). The exact connection between forward and reverse dynamics in continuous-time is discussed in (Anderson, 1982; Song et al., 2021b). Note that there exist multiple variations of this scheme (referred to as *denoising diffusion probabilistic model*, *DDPM*), including alternative definitions of $\alpha_t, \beta_t, \rho_t, \sigma_t$ and reverse kernels $K_{\theta,t}$ (see, e.g., (Song et al., 2021a; Karras et al., 2022)). This work is largely agnostic to these differences and can be combined with many of these alternatives, though for clarity, the remainder of the manuscript deals only with DDPMs.

2.2 GUIDED DIFFUSION

A broad class of applications requires samples not from $\tilde{\pi}_0$, but from a distribution whose density is

$$\tilde{p}_0^w(x) \propto w(x)\tilde{p}_0(x), \tag{3}$$

where $\tilde{p}_0$ denotes the density of $\tilde{\pi}_0$ and $w : \mathcal{X} \to \mathbb{R}_+$ is a fixed weighting function. This includes inverse problems, where w denotes a likelihood model, or conditional sampling, where w is the indicator function of an event. Rather than training a new DDM targeting (3), it can be shown that the score of the distribution of $\tilde{X}_t$ obtained by drawing $\tilde{X}_0 \sim \tilde{p}_0^w$ in (2) is given by

$$\nabla_{x_t} \log \tilde{p}_t^w(x_t) = \nabla_{x_t} \log \tilde{p}_t(x_t) + \nabla_{x_t} \log h_t^w(x_t), \tag{4}$$

where $h_t^w(x_t) = \mathbb{E}\big[w(\tilde{X}_0) | \tilde{X}_t = x_t\big]$ is sometimes referred to as *Doob's h-transform* (Rogers & Williams, 2000). Since the denoiser s_θ from (1) is trained to approximate $\nabla_{x_t} \log \tilde{p}_t(x_t)$, (4) provides a correction that, when applied to (1), yields (approximate) samples from $\tilde{p}_0^w$ rather than $\tilde{\pi}_0$. Since $\nabla_{x_t} \log h_t^w(x_t)$ *guides* the diffusion towards $\tilde{\pi}_0^w$, it is often referred to as a *guidance*.

In general, h_t^w is intractable and guidance methods use approximations based on atomic distributions or (pretrained) parametrizations (e.g., (Didi et al., 2023; Chung et al., 2023; Denker et al., 2024)), Gaussian distributions (e.g., (Ho et al., 2022; Song et al., 2023a)), or Monte Carlo and variational methods (e.g., (Song et al., 2023b; Moufad et al., 2025)). Additionally, ensuring that the final samples satisfy a desired property often requires the guidance term in (4) to be reweighted. Consider, for instance, Diffusion Posterior Sampling (DPS), a guidance method conceived to solve inverse problems involving measurements $Y \sim \mathcal{N}\big(\mathcal{A}(X_0), \nu^2 I\big)$, where $\mathcal{A} : \mathcal{X} \to \mathbb{R}^n$ denotes a forward operator and $\nu > 0$, the noise variance. It replaces the transition kernel mean in (1) by

$$m_{\theta,t}^w(x) = \alpha_t^{-1/2}\Big(x + \beta_t s_\theta(x,t) + \gamma_t \alpha_t \nabla_{x_t} \log \hat{h}_t(x_t)\Big), \text{ for } \hat{x}_{0|t}(x) = \frac{x + (1-\bar{\alpha}_t) s_\theta(x,t)}{\sqrt{\bar{\alpha}_t}}, \quad (5)$$

where $\hat{h}_t(x) = \exp\big(-\|y - \mathcal{A}(\hat{x}_{0|t}(x))\|^2\big)$ and $\gamma_t > 0$ is a hyperparameter (Chung et al., 2023). Here, $\hat{x}_{0|t}(x)$ estimates $\mathbb{E}[\tilde{X}_0 \mid \tilde{X}_t = x]$ by replacing the exact score in Tweedie's identity from (Efron, 2011) with s_θ (see Sec. B). DPS and other guidance methods tend to be sensitive to the value of the guidance scale γ_t that must place sufficient weight on the correction to achieve the desired behavior (e.g., consistency with $\mathcal{A}$) without significantly deviating from the original kernel so as not to introduce artifacts. Different heuristics have been proposed to adjust γ_t (e.g., (Yang et al., 2024)).

Alternative, optimization-inspired guidance methods have also been proposed, based on projections (Lugmayr et al., 2022; Christopher et al., 2024), trust regions (Huang et al., 2024), proximal maps (Jayaram et al., 2025; Zampini et al., 2025), and variable splitting methods (Blanke et al., 2026). Nevertheless, they also rely on sensitive hyperparameters (e.g., step sizes and trust region radii). They are also restricted to requirements for which projections or proximal maps exist in closed form. Otherwise, they need to solve an optimization problem at each inference step.

3 Constrained Diffusion Sampling

3.1 The constrained diffusion problem

To address the challenges of sensitivity to hyperparameters, high computational cost, and limited requirement types, we take a hybrid approach that enables a kind of Doob's h-transform (4) to explicitly solve a constrained optimization problem. Consider the problem of sampling not from π_0 (or $\tilde{\pi}_0$ or $\tilde{p}_0^w$), as do the DDMs from Sec. 2, but from a nearby distribution that satisfies a set of generalized moment constraints. Explicitly, for a collection of (sub)differentiable constraint functions $g_i : \mathbb{R}^d \to \mathbb{R}$, we wish to sample from the distribution $\mu^\star$ that solves

$$\begin{aligned} P^\star = \min_{\mu \in \mathcal{P}} \quad & \mathrm{KL}(\mu \| \pi_0) \\ \text{subject to} \quad & \mathbb{E}_{X\sim\mu}\big[g_i(X)\big] \le 0, \quad i = 1, \dots, m, \end{aligned} \quad \text{(PI)}$$

where $\mathcal{P}$ denotes the space of probability measures on $(\mathbb{R}^d, \mathcal{B})$, with $\mathcal{B}$ the Borel σ-algebra, and whose objective is the Kullback-Leibler (KL) divergence defined as

$$\mathrm{KL}(\mu \| \pi_0) = \int \log\left(\frac{d\mu}{d\pi_0}\right) d\mu,$$

for the Radon-Nikodym derivative $\frac{d\mu}{d\pi_0}$, when μ is absolutely continuous with respect to π_0 ($\mu \ll \pi_0$) and $+\infty$ otherwise (Durrett, 2010). We let the expectation in (PI) be $+\infty$ whenever g is not μ-integrable, rendering those measures infeasible. Note that requirements of the form $\mathbb{E}_\mu[\tilde{g}_i] \le c_i$ can be imposed by taking $g_i(x) = \tilde{g}_i(x) - c_i$. Problem (PI) is closely related to the constrained sampling problem tackled in (Liu et al., 2021; Chamon et al., 2024), though these works consider the Markov Chain Monte Carlo (MCMC) setting where $p_0(x) \propto e^{-f(x)}$ for a known f.

In contrast to previous guidance approaches, (PI) imposes constraints on the distribution implied by a pretrained DDM rather than specifying task-dependent changes to its reverse dynamics (1). It is more general than restricting the samples themselves, as do projection methods (Lugmayr et al., 2022; Christopher et al., 2024), since (PI) can ensure $\tilde{g}_i(x) \le 0$ almost surely by letting $g_i(x) = [\tilde{g}_i(x)]_+$ (see, e.g., (Liu et al., 2021; Chamon et al., 2024)). It also provides a more interpretable framework than the weights (3), accommodating multiple constraints without adding new guidance scales γ_t. Next, we illustrate the wide range of behaviors that can be cast in (PI) (see also Sec. 5).

Inverse problems. Given the forward operator $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^n$, an inverse problem seeks $x \in \mathbb{R}^d$ such that $\mathcal{A}(x)$ is close to a measurement $y \in \mathbb{R}^n$. Since this problem is typically under-determined ($n < d$), we instead seek a distribution of x close to a prior π_0 implied by a pretrained DDM. This task can be cast using (PI) with, for a tolerance $c_{\text{inv}} > 0$,

$$g_{\text{inv}}(x) = \|A(x) - y\|^2 - c_{\text{inv}} \quad \text{or} \quad g_{\text{inv}}(x) = \|A(x) - y\|_1 - c_{\text{inv}},$$

depending on whether we consider Gaussian- (as in DPS (5)) or Laplace-distributed errors (Robert & Casella, 2004). Inverse problems are found in applications such as inpainting and deblurring.

Loss/reward optimization. Let $\ell : \mathbb{R}^d \to \mathbb{R}$ be a (sub)differentiable loss denoting the departure from a desired condition. The task of generating samples from π_0 that yield small losses (large rewards) can be formulated in (PI) using $g_{\text{loss}}(x) = \ell(x) - c_{\text{loss}}$ for an upper bound c_{loss}. This is useful, for instance, to perform semantic control by taking $\ell(x) = -\log\big(\varphi(x)\big)$ for a classifier $\varphi : \mathcal{X} \to (0, 1]$ pretrained to detect, e.g., the presence of a truck or a smile. It can also be used to adjust visual features of an image by taking $\ell(x) = |\phi(x) - b|$, where ϕ is a measure of, e.g., brightness or contrast, and b is the target level.

Multiple requirements. A key advantage of (PI) is its ability to describe complex tasks involving multiple, heterogeneous requirements by simply stacking g_i's. It is straightforward, for instance, to generate images of a smiling individual with brightness levels that match a specific aesthetic or incorporate smoothness priors in imaging inverse problems. Note that each requirement is imposed individually rather than by means of a fine-tuned weighted combination of objective functions.

Still, a key challenge with (PI) is that π_0 is only known implicitly from the reverse process (1) of a pretrained DDM. Next, we address this issue by deriving a result akin to Doob's h-transform from Sec. 2.2, but for sampling from $\mu^\star$ instead of p_0^w.

3.2 A Doob-like transform

Let $g : \mathbb{R}^d \to \mathbb{R}^m$ collect the constraint functions g_i and define the Lagrangian of (PI) as

$$L(\mu, \lambda) \triangleq \mathrm{KL}(\mu \| \pi_0) + \lambda^\top \mathbb{E}_\mu[g], \quad \text{for } \lambda \in \mathbb{R}^m_+. \tag{6}$$

Its *dual problem* is then defined in terms of the dual function $d(\lambda) \triangleq \min_{\mu \in \mathcal{P}} L(\mu, \lambda)$ as

$$D^\star = \sup_{\lambda \in \mathbb{R}^m_+} d(\lambda) \tag{DI}$$

A solution $\lambda^\star$ of (DI) is called a *Lagrange multiplier*, the set of which is denoted by $\Lambda^\star$. Note that (DI) is a convex optimization program, given that d is concave (it is the minimum of a set of affine functions) (Bertsekas, 2009). It is also a relaxation of (PI), which yields $D^\star \leq P^\star$. Next, we show conditions under which this inequality is tight (strong duality), implying we can solve the infinite dimensional (PI) by solving the finite dimensional, differentiable, convex problem (DI).

Proposition 3.1. *Suppose the g_i are π_0-integrable and bounded below π_0-almost surely and that there exist $\mu^\dagger \in \mathcal{P}$ such that $\mathrm{KL}(\mu^\dagger \| \pi_0) < +\infty$ and $\mathbb{E}_{\mu^\dagger}[g_i] \leq -\delta < 0$ for all $i = 1, \ldots, m$. For $\lambda \in \mathbb{R}^m_+$, define μ_0^λ through its density*

$$p_0^\lambda(x) = Z_\lambda^{-1} \exp\big[-\lambda^\top g(x)\big] p_0(x), \quad \textit{for } Z_\lambda \triangleq \mathbb{E}_{\pi_0}[\exp\big(-\lambda^\top g\big)].$$

Then, $d(\lambda) = L(\mu_0^\lambda, \lambda)$ is differentiable with $\nabla d(\lambda) = \mathbb{E}_{\mu_0^\lambda}[g]$. Moreover, $P^\star = D^\star$, there exists a finite R such that $\|\lambda^\star\| \leq R$ for all $\lambda^\star \in \Lambda^\star$, and $\mu^\star = \mu_0^{\lambda^\star}$ for all $\lambda^\star \in \Lambda^\star$.

See App. C.1 for a proof. The strict feasibility assumption in Prop. 3.1, known as *Slater's conditions* (Bertsekas, 2009), makes it possible to obtain an explicit value for R as a function of δ (see, e.g., (Chamon et al., 2023, Lemma III.2)). Hence, (DI) need only be solved over $\Lambda = \{\lambda \in \mathbb{R}^m_+ : \|\lambda\| \leq R\}$. Still, doing so depends on the implicitly-defined π_0. The next result shows how to sample from μ_0^λ using only a pretrained DDM.

Proposition 3.2. *Under the assumptions of Prop. 3.1, a sample $X_0^\lambda \sim \mu_0^\lambda$ can be obtained from the backward Markov chain $(X_t^\lambda)_{t=T}^0$ with transition kernel $K_{\theta,t}^\lambda$ and initial state $X_T^\lambda \sim \mu_T^\lambda$, where*

$$\frac{d\mu_T^\lambda}{d\pi_T} = \frac{h_T^\lambda}{Z_\lambda}, \quad K_{\theta,t}^\lambda(\cdot|x_t) = K_{\theta,t}(\cdot|x_t)\frac{h_{t-1}^\lambda(\cdot)}{h_t^\lambda(x_t)}, \quad \textit{and} \quad h_t^\lambda(x) = \mathbb{E}\big[e^{-\lambda^\top g(X_0)} \;\big|\; X_t = x\big]. \tag{7}$$

The distributions μ_t^λ of the marginals X_t^λ are defined by the densities $p_t^\lambda(x) = Z_\lambda^{-1} h_t^\lambda(x) p_t(x)$.

A proof is provided in App. C.2. Together, Prop. 3.1 and 3.2 yield a Doob-like characterization of constrained diffusion sampling.

Corollary 3.3 ($h^\star$-transform)**.** *Let $\lambda^\star$ be a solution of* (DI)*. Then, starting from $\mu_T^{\lambda^\star}$ and applying the modified reverse kernel $K_{\theta,t}^{\lambda^\star}$ in* (7) *for $t = T, \dots, 1$ yields the solution of* (PI)*.*

If Doob's h-transform modifies diffusion processes to reach the conditional or reweighted target $\tilde{p}_0^w$ in (3), the $h^\star$-transform of Cor. 3.3 does the same for constrained targets. Surely, $h_t^{\lambda^\star}$ determines the corrections on the DDPM kernel $K_{\theta,t}$ engendered by the constraints. In fact, the score of the reverse process marginals becomes $\nabla_x \log p_t^\star(x) = \nabla_x \log p_t(x) + \nabla_x \log h_t^{\lambda^\star}(x)$, an expression very similar to the h-transform in (4).

These results also suggest a method to approximate $\mu^\star$ without direct access to π_0. Indeed, it can be obtained by taking $\lambda_0 = 0$ and iterating over the idealized algorithm

$$\lambda_{k+1} = \Pi_\Lambda\Big[\lambda_k + \eta_k \, \mathbb{E}_{\mu_0^{\lambda_k}}[g]\Big], \quad \text{for } \mu_t^{\lambda_k} = \begin{cases} \frac{h_T^{\lambda_k}}{Z_\lambda}\pi_T, & t = T \\ \int K_{\theta,t+1}^{\lambda_k}(\cdot \mid x)\mu_{t+1}^{\lambda_k}(dx), & t = T-1, \dots, 0 \end{cases}, \tag{8}$$

where $\eta_k > 0$ is a step size and Π_Λ denotes the projection onto Λ. Recall from Prop. 3.1 that $\mathbb{E}_{\mu_0^\lambda}[g] = \nabla d(\lambda)$ and that $\lambda^\star \in \Lambda$, so that (8) is effectively a projected gradient ascent algorithm for (DI). Since it is a concave optimization problem, its convergence can be established under very general conditions (see, e.g., (Bertsekas, 2015)). Yet, though (8) relies only on the $K_{\theta,t}$ from (1), it remains intractable. In what follows, we develop a practical constrained diffusion algorithm based on (8) with convergence guarantees that we show to match or outperform existing baselines at a fraction of their computation and memory costs (Sec. 5).

4 Primal-Dual Constrained Diffusion Sampling

The idealized algorithm (8) manipulates probability distributions μ_t^λ and requires computing the full reverse diffusion process for each dual update. In this section, we derive PD-CDS (Alg. 1), an algorithm that updates the dual variables along a single trajectory (*dual step*) based on samples from $\mu_t^{\lambda_t}$ generated using a DDPM-like algorithm (*primal step*) [1].

Dual update. By the tower property, for any $\lambda \in \mathbb{R}_+^m$,

$$\mathbb{E}_{\mu_0^\lambda}[g] = \mathbb{E}_{X_t^\lambda \sim \mu_t^\lambda}\Big[\mathbb{E}_{\mu_{0|t}^\lambda}\big[g(X_0^\lambda) \mid X_t^\lambda\big]\Big], \quad \text{where } \mu_{0|t}^\lambda = \int \prod_{\tau=1}^{t} K_{\theta,\tau}^\lambda(x_{\tau-1}|x_\tau)dx_{1:t-1}, \tag{9}$$

i.e., $\mu_{0|t}^\lambda$ is the conditional law of X_0^λ given X_t^λ. From (9), we can construct a stochastic approximation of the dual update in (8), namely $\hat{r}_t(\lambda, x) = \mathbb{E}_{\mu_{0|t}^\lambda}\big[g(X_0^\lambda)|X_t^\lambda = x\big]$. Note that $\hat{r}_t$ is an unbiased estimate of $\nabla d(\lambda) = \mathbb{E}_{\mu_0^\lambda}[g]$ for $x \sim \mu_t^\lambda$. To obtain a tractable approximation of $\hat{r}_t$, we use a first-order approximation of g to write $\hat{r}_t(\lambda, x) \approx g\big(\mathbb{E}_{\mu_{0|t}^\lambda}[X_0^\lambda|X_t^\lambda = x]\big)$. Following the DDPM approximation (Song & Ermon, 2019; Ho et al., 2020), we regard μ_t^λ as being close to the corresponding marginal of the forward process (2) started from $\tilde{X}_0^\lambda \sim \mu_0^\lambda$. This motivates using Tweedie's identity to approximate the conditional expectation and write

$$\hat{r}_t(\lambda, x) \approx g\big(\hat{x}_{0|t}^\lambda(x)\big), \quad \text{for } \hat{x}_{0|t}^\lambda(x) = \frac{x + (1-\bar{\alpha}_t)\nabla \log p_t^\lambda(x)}{\sqrt{\bar{\alpha}_t}}. \tag{10}$$

To proceed, observe that $\nabla \log p_t^\lambda(x) = \nabla \log p_t(x) + \nabla \log h_t^\lambda(x) \approx s_\theta(x,t) + \nabla \log h_t^\lambda(x)$. To approximate the remaining "h^λ-transform," use a first-order approximation of h_t^λ together with the same Tweedie's estimate as above to get

$$\nabla \log h_t^\lambda(x_t) \approx -\sum_{i=1}^m \lambda_i J_{\hat{x}_{0|t}}(x_t)^\top \nabla g_i(x)\big|_{x=\hat{x}_{0|t}(x_t)}, \tag{11}$$

[1] Note that *sampling from* μ_t^λ is not technically a primal update of the saddle point problem (8), though it is related to it by the Fokker-Planck equation (see (Chamon et al., 2024) for more details).

Algorithm 1 PD-CDS (Primal-Dual Constrained Diffusion Sampling)

Input: Pretrained DDM s_θ, constraint functions g_i, noise schedule β_t, and step sizes η_t

1: Set $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{j=1}^{t} \alpha_j$, $\lambda_T = 0$, and sample $X_T \sim \mathcal{N}(0, I)$

2: **for** $t = T, \dots, 1$ **do**

3: $\hat{x}_{0|t} = \dfrac{X_t + (1-\bar{\alpha}_t) s_\theta(X_t, t)}{\sqrt{\bar{\alpha}_t}}$, $s_t^\lambda = -\displaystyle\sum_{i=1}^{m} \lambda_{t,i} \nabla g_i(\hat{x}_{0|t})$, and $\hat{x}_{0|t}^{\lambda_t} = \hat{x}_{0|t} + \dfrac{(1-\bar{\alpha}_t) s_t^\lambda}{\sqrt{\bar{\alpha}_t}}$

4: $X_{t-1} \sim \mathcal{N}\left(\dfrac{X_t + \beta_t \big(s_\theta(X_t, t) + s_t^\lambda\big)}{\sqrt{\alpha_t}}, \beta_t I \right)$ ▷ Primal step: reverse diffusion

5: $\lambda_{t-1,i} = \Pi_\Lambda \Big[\lambda_{t,i} + \eta_t g_i\big(\hat{x}_{0|t}^{\lambda_t}\big) \Big]$ ▷ Dual step: (stochastic) gradient ascent

6: **end for**

7: **return** X_0

where $\hat{x}_{0|t}$ is defined in (5) and $J_{\hat{x}_{0|t}}$ its Jacobian. We explicitly write the chain rule for $\nabla_x g(\hat{x}_{0|t}(x))$ in (11) to highlight a key source of complexity for guidance methods: the fact that h-transforms depend on $J_{\hat{x}_{0|t}}$ whose computation requires backpropagating through the DDM $s_\theta(x, t)$. Nevertheless, note that for $s_\theta(\cdot, t) = \nabla \log \tilde{p}_t$, it holds that $J_{\hat{x}_{0|t}}(x) \propto \Sigma_{0|t}(x)$, the conditional covariance of $\tilde{X}_0 | \tilde{X}_t$ (see, e.g., (Dytso et al., 2023)). Whenever $\Sigma_{0|t}$ is approximately proportional to I, the Jacobian acts as a rescaling of the dual variable λ_i that can be automatically compensated by the dual updates in step 5 of Alg. 1. This occurs, for instance, when the reconstruction errors are sufficiently decorrelated. This motivates omitting the Jacobian in the approximation s_t^λ used in Alg. 1, which we found to have little performance impact in our experiments (see Sec. 5).

Primal update. Start by noticing that for $\lambda_T = 0$, (7) gives $\mu_T^0 = \pi_T = \mathcal{N}(0, I)$, justifying the initialization in Alg. 1. We then approximate the reverse transitions using the DDPM Gaussian parametrization (Song & Ermon, 2019; Ho et al., 2020), applied to the score of μ_t^λ, which combined with Prop. 3.2 and $s_\theta(x, t) \approx \nabla \log p_t(x)$ yields

$$\hat{K}_{\theta,t}^\lambda(\cdot \mid x_t) = \mathcal{N}\Big(m_{\theta,t}(x_t) + \alpha_t^{-1/2} \beta_t \nabla \log h_t^\lambda(x_t),\ \beta_t I \Big), \tag{12}$$

for h_t^λ as in (7) and $m_{\theta,t}(x_t)$ as in the original DDPM (1). If not for the adaptive dual variable λ and various scaling factors, (12) essentially describes the DPS algorithm (Chung et al., 2023) [see (5)]. Yet, these distinctions are crucial: the solution of (PI) is $\mu_0^{\lambda^\star}$ for an (*a priori*) unknown $\lambda^\star$ that the dual updates in step 5 seek to find. Replacing $\nabla \log h_t^\lambda(x_t) \approx s_t^\lambda$ then yields step 4 of Alg. 1.

Finally, note that sampling from $\mu_t^{\lambda_t}$ would require restarting the reverse process from $X_T^{\lambda_t} \sim \mu_T^{\lambda_t}$, which would substantially increase inference cost. Instead, Alg. 1 continues directly from time $t+1$ using $X_{t+1}^{\lambda_t}$ as a warm start. If the dual variables change slowly relative to the sampling dynamics, this approximation is similar to that used in two-timescale algorithms (see, e.g., (Borkar, 1997)). In the sequel, we show that this mismatch and the preceding approximations do not hinder the ability of Alg. 1 to approximate $\lambda^\star$.

Theoretical guarantee. To isolate the effect of our approximations from those already used in the development of DDPM (Song & Ermon, 2019; Ho et al., 2020), we study the convergence of Alg. 1 towards the solution of a primal-dual pair that replaces the target of (PI)–(DI) by $\tilde{\pi}_0$ from (2). Namely,

$$\tilde{\mu}^\star = \operatorname*{argmin}_{\mu \in \mathcal{P}} \ \mathrm{KL}(\mu \| \tilde{\pi}_0) \ \text{ subject to } \ \mathbb{E}_{X \sim \mu}\big[g(X)\big] \leq 0 \tag{$\widetilde{\text{PI}}$}$$

$$\tilde{\lambda}^\star \in \operatorname*{argmax}_{\lambda \in \mathbb{R}_+^m} \ \tilde{d}(\lambda) \triangleq \min_{\mu \in \mathcal{P}} \ \tilde{L}(\mu, \lambda) \tag{$\widetilde{\text{DI}}$}$$

with $\tilde{L}(\mu, \lambda) = \mathrm{KL}(\mu \| \tilde{\pi}_0) + \lambda^\top \mathbb{E}_\mu[g]$. In particular, this avoids having to deal with the approximation errors between the marginals π_t and $\tilde{\pi}_t$ of the reverse and forward processes (1)–(2), already considered in previous work, e.g., (Chen et al., 2023; Silveri & Ocello, 2025; Conforti et al., 2025). By showing that $\tilde{\pi}_0 \approx \pi_0$, they also show that ($\widetilde{\text{PI}}$)–($\widetilde{\text{DI}}$) are close to (PI)–(DI). Before introducing our assumptions, we state a key remark.

Remark 4.1. Under the conditions of Prop. 3.1 with π_0 replaced by $\tilde{\pi}_0$, the results of Sec. 3.2 apply to ($\widetilde{\text{PI}}$)–($\widetilde{\text{DI}}$) by replacing p_t with $\tilde{p}_t$. Indeed, $\tilde{d}(\lambda) = \tilde{L}(\tilde{\mu}_0^\lambda, \lambda)$ for $\tilde{\mu}_0^\lambda$ with density $\tilde{p}_0^\lambda(x) = \tilde{Z}_\lambda^{-1} e^{-\lambda^\top g(x)} \tilde{p}_0(x)$ with $\tilde{Z}_\lambda = \mathbb{E}_{\tilde{\pi}_0}[e^{-\lambda^\top g}]$. Also, $\tilde{\mu}^\star = \tilde{\mu}_0^{\tilde{\lambda}^\star}$ (Prop. 3.1). Additionally, for $\tilde{K}_t(\cdot|x) = \mathbb{P}(\tilde{X}_{t-1} \in \cdot | \tilde{X}_t = x)$, we can sample from $\tilde{\mu}_0^\lambda$ by transporting $\tilde{X}_T^\lambda \sim \tilde{\mu}_T^\lambda$ using the reverse kernel

$$\tilde{K}_t^\lambda(\cdot|x) = \tilde{K}_t(\cdot|x) \frac{\tilde{h}_{t-1}^\lambda(\cdot)}{\tilde{h}_t^\lambda(x)}, \quad \text{where } \tilde{h}_t^\lambda(x) = \mathbb{E}\big[e^{-\lambda^\top g(\tilde{X}_0)} \,\big|\, \tilde{X}_t = x\big] \text{ and } \frac{d\tilde{\mu}_T^\lambda}{d\tilde{\pi}_T} = \frac{\tilde{h}_T^\lambda}{\tilde{Z}_\lambda}, \quad (13)$$

obtaining marginals $\tilde{X}_t^\lambda$ with density $\tilde{p}_t^\lambda(x) = \tilde{Z}_\lambda^{-1} \tilde{h}_t^\lambda(x) \tilde{p}_t(x)$ (Prop. 3.2).

Assumption 4.2. The constraint functions g are continuously differentiable,

$$\sup_x \|g(x)\| \le G, \;\; \sup_x \|J_g(x)\|_{\text{op}} \le L_g, \;\; \text{and} \;\; \|\tilde{X}_0\| \le R_X \text{ almost surely},$$

where $J_g(x)$ denotes the Jacobian of g at x and $\|\cdot\|_{\text{op}}$ denotes the operator (spectral) norm.

Assumption 4.3. Suppose the pretrained score is exact, i.e., $s_\theta(\cdot, t) = \nabla \log \tilde{p}_t$ for all t, and let

$$\big\|s_t^\lambda(x) - \nabla \log \tilde{h}_t^\lambda(x)\big\| \le \varepsilon_h, \text{ for all } t, \quad \text{and} \quad \mathrm{KL}\Big(\tilde{K}_t^\lambda(\cdot|x) \,\Big\|\, \tilde{K}_t^{\lambda,\mathrm{G}}(\cdot|x)\Big) \le \frac{\varepsilon_{\mathrm{GD}}^2}{2}, \text{ for } t \ge 2,$$

uniformly over $x \in \mathbb{R}^d$ and $\lambda \in \Lambda$ for

$$s_t^\lambda(x) = -\sum_{i=1}^m \lambda_i \nabla g_i(\xi)\big|_{\xi = \hat{x}_{0|t}(x)} \quad \text{and} \quad \tilde{K}_t^{\lambda,\mathrm{G}}(\cdot|x) \triangleq \mathcal{N}\bigg(m_{\theta,t}(x) + \frac{\beta_t}{\sqrt{\alpha_t}} \nabla \log \tilde{h}_t^\lambda(x), \beta_t I_d\bigg).$$

Ass. 4.2 is typical in the analysis of optimization algorithms (Nedić & Ozdaglar, 2009; Bertsekas, 2015). Note that only $\tilde{X}_0$ must be almost surely bounded; the states $(\tilde{X}_t)_{t>0}$ can have unbounded support. Ass. 4.3 describes our approximations, where we take the pretrained DDM to match the score of the forward process and isolate the guidance (ε_h) and reverse kernel approximation ($\varepsilon_{\mathrm{GD}}$) errors. Note that the latter starts from $t = 2$, which is important to stay consistent with the boundedness hypothesis on $\tilde{X}_0$. The effects of approximating the score and assuming that the reverse kernels are Gaussian have been studied in the context of DDPM in, e.g., (Chen et al., 2023; Silveri & Ocello, 2025; Conforti et al., 2025). Incorporating these error into our analysis is left for future work.

Theorem 4.4. *Consider a fixed β schedule and let the dual step size in Alg. 1 be $\eta = 2R(G\sqrt{T})^{-1}$. Then, under the conditions of Prop. 3.1 and Ass. 4.2 and 4.3, the sequence λ_t it produces satisfies*

$$\tilde{d}(\tilde{\lambda}^\star) - \frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\tilde{d}(\lambda_t)\big] \le C + O\left(\frac{1}{\sqrt{T}}\right), \quad \textit{where } C = O\big(R(1 + \varepsilon_h + \varepsilon_{GD})\big). \quad (14)$$

A precise statement along with a proof can be found in App. D. Thm. 4.4 shows that, despite its differences with the idealized (8), Alg. 1 still approximates the value of ($\widetilde{\text{DI}}$) at a typical rate under our assumptions (see, e.g., (Nedić & Ozdaglar, 2009; Bertsekas, 2015)). The bias C of this approximation is directly impacted by the errors from Ass. 4.3. It can therefore be reduced by, e.g., improving the approximation of $\nabla \log \tilde{h}_t^\lambda$ by computing the Jacobian in (11) or using the refinements from (Moufad et al., 2025; Rout et al., 2024). Interestingly, C is also affected by the magnitude of the Lagrange multipliers R, a measure of requirement sensitivity. Indeed, for small enough ϵ, the objective of ($\widetilde{\text{PI}}$) stands to decrease by $\tilde{\lambda}_i^\star \epsilon$ if the constraint $\mathbb{E}_\mu[g_i] \le 0$ is relaxed to $\mathbb{E}_\mu[g_i] \le \epsilon$ (Bertsekas, 2009; Bonnans & Shapiro, 2000). It is therefore harder for Alg. 1 to enforce constraints that require large deviations from the pretrained DDM $\tilde{\pi}_0$. Doing so accurately calls for smaller errors $\varepsilon_h, \varepsilon_{\mathrm{GD}}$.

5 Experiments

We evaluate PD-CDS on image reconstruction (inverse problems) and semantic-photometric control (loss optimization) applications, particularly in settings that combine multiple requirements. We compare its performance and computational cost to different baselines, including DPS (Chung et al., 2023), DAPS (Zhang et al., 2025a), DiffPIR (Zhu et al., 2023), ΠGDM (Song et al., 2023a), and RED-diff (Mardani et al., 2024) for inverse problems; and TFG (Ye et al., 2024), UGD (Bansal et al.,

Table 1: Inverse problems on images from FFHQ: FID and mean LPIPS (lower is better; see App. E.3 for additional metrics). Best results are shown in bold-face and second-best are underlined.

	Gaussian deblur		Motion deblur		SR $\times 4$		Nonlinear deblur		Desolarization	
Method	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS
PD-CDS	69.38	0.145	<u>56.41</u>	<u>0.075</u>	62.54	<u>0.122</u>	<u>35.38</u>	0.150	<u>114.02</u>	<u>0.295</u>
PD-CDS + TV	<u>61.10</u>	0.160	**46.12**	**0.061**	<u>53.23</u>	0.123	**29.65**	<u>0.107</u>	179.56	0.438
DAPS	77.73	0.146	75.54	0.128	75.98	0.131	42.60	**0.064**	**79.47**	**0.174**
DiffPIR	**60.00**	<u>0.141</u>	80.29	0.168	61.81	0.146	96.61	0.386	139.13	0.348
DPS	78.41	**0.130**	85.67	0.169	93.76	0.190	284.71	0.626	221.31	0.555
ΠGDM	79.03	0.159	91.53	0.196	**43.65**	**0.077**	–	–	–	–

2024b), LGD (Song et al., 2023b), and FreeDoM (Yu et al., 2023) for loss-based guidance. When comparing methods, we use the same operator, DDM, observations, classifiers, and seeds. Though step 5 in Alg. 1 imposes an upper bound R on the norm of the dual variables, the results in this section are obtained by projecting only onto the non-negative orthant, precluding the need to determine R. Full implementation details can be found in App. E.1.

Inverse problems. We start by considering linear and nonlinear imaging problems, namely (1) $\times 4$ super-resolution (from bicubically downsampled targets); (2) Gaussian and motion deblurring (same setup as (Chung et al., 2023)); (3) nonlinear deblurring, using the model from (Tran et al., 2021); and (4) desolarizing, for images degraded by inverting the value of pixels above a threshold of 0.5 (see App. E for additional results on inpainting). We use the first 100 images of the FFHQ validation set and the pretrained pixel-space diffusion model from Chung et al. (2023) as a prior.

We consider $g_{\text{inv}}(x) = (c_{\text{inv}}N)^{-1}\|\mathcal{A}(x) - y\|_1 - 1$ where $c_{\text{inv}} = 0.01$ and N is the number of observed pixels. This normalized version of the constraint from Sec. 3.1 is used across problems with the same values to showcase the robustness of PD-CDS. We also consider a smoothness prior commonly used in image reconstruction (Chambolle & Pock, 2011) by taking $g_{\text{TV}}(x) = c_{\text{TV}}^{-1}\text{TV}(x) - 1$, where $c_{\text{TV}} = 0.1$ and TV denotes the total variation (TV) norm implemented as in (Rudin et al., 1992) (more details in App. E.1).

Results. Table 1 reports FID and mean LPIPS (values for PSNR and SSIM; results for RED-diff; and standard deviations can be found in App. E). Observe that PD-CDS is competitive overall and excels, together with DAPS, on nonlinear problems. Note, however, that DAPS takes on average twice as long to compute each sample (being up to 7 times slower for nonlinear blur, see Table 9 in the appendices). Despite their similarities, PD-CDS achieves lower FID and LPIPS than DPS on all but one task even without backpropagating through the DDM (reducing by 2 the memory costs). Only DiffPIR and ΠGDM are an order of magnitude faster than PD-CDS, though they can only properly handle linear measurement operators $\mathcal{A}$. ΠGDM in particular only applies to linear $\mathcal{A}$ (and is substantially more memory intensive). The performance of PD-CDS can be further improved in all but the desolarization setting at virtually no additional computational cost by adding a TV prior.

Semantic-photometric control. We now consider the problem of generating images with specific semantic (`Smiling`, `Not-Young`) and photometric attributes (brightness, contrast) using the pretrained DDPM from Ho et al. (2020) (trained on CelebA-HQ). We start from a single semantic requirement (e.g., `Smiling`), then combine semantic and photometric requirements (e.g., `Smiling` and brightness), and conclude by generating images that satisfy four constraints, i.e., `Smiling`, `Not-Young`, brightness, and contrast. Other combinations as well as class-conditional results on CIFAR-10 can be found in App. E. Every method is evaluated based on 200 generated images.

The semantic constraints are imposed using a pretrained classifier φ (see details in App. E.1) and the constraint functions described in Sec. 3.1 with $c_{\text{sem}} = -\log(0.97)$ for `Smiling` and $c_{\text{sem}} = -\log(0.99)$ for `Not-Young`. Similarly for the photometric features (with ϕ evaluating the brightness or contrast of the image) with $c_{\text{phot}} = 0.01$. For the baselines, we sum all guidance losses with unit weights as in (Ye et al., 2024).

Results. Table 2 reports accuracy (Acc.), measured by an independent classifier (not φ), the mean absolute error (MAE) to the photometric target b, and KID (Bińkowski et al., 2018) with respect

Table 2: Semantic (Acc.) and photometric (MAE) control. Best results are shown in bold-face and second-best are underlined.

	Smiling		*Smiling + brightness* ($b = 0.65$)			
Method	Acc.↑	KID↓	Acc.↑	MAE↓	Joint acc.↑	KID↓
PD-CDS	95.0%	**0.0315**	98.0%	**0.0001**	**98.0**	**0.0312**
TFG	**100.0%**	0.0408	99.0%	0.0162	60.0	0.0601
UGD	**100.0%**	0.0615	94.0%	0.0126	62.5	0.0549
LGD	99.5%	0.0478	**100.0%**	0.1139	7.0	0.0744
FreeDoM	**100.0%**	0.0639	93.5%	0.0105	75.0	0.0529

Figure 1: Samples under multiple constraints.



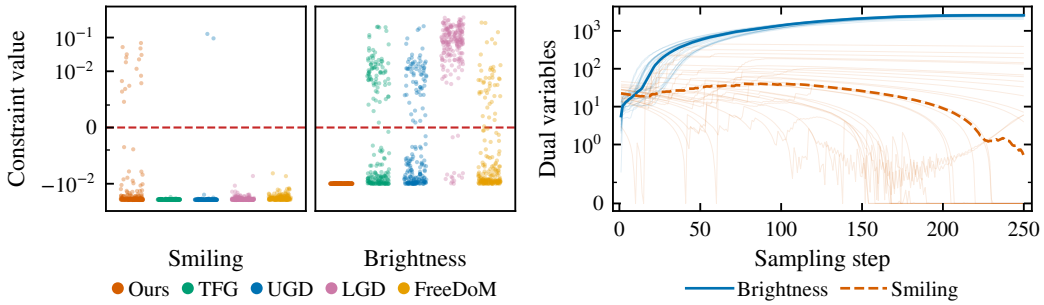


Figure 2: Constraint violation $g_{\text{sem}}, g_{\text{phot}}$ (negative means feasible, left) and dual variables trajectories (right) for `Smiling` + brightness.

to 1,000 CelebA-HQ validation images. Joint accuracy (Joint acc.) denotes the proportion of generated samples that are simultaneously classified as belonging to the semantic target and satisfy the photometric constraint. While all approaches are effective at imposing the `Smiling` requirement, PD-CDS achieves the lowest KID among evaluated methods. This is due to the tolerance c_{sem} that prevents overfitting to the loss, which allows the DDM to retain sample diversity. This is achieved with half the memory usage of the baselines (see Table 10 in the appendix).

This fact is all the more important when integrating the high brightness requirement. Other approaches are unable to target both goals (Fig. 2(left)), leading to considerably lower joint accuracies. PD-CDS, on the other hand, successfully applies both constraints without sacrificing its KID score. This occurs because the dual variables λ in Alg. 1 dynamically adapt the strength of each guidance loss during the inference process. Indeed, observe from Fig. 2(right) that while the brightness constraint must be imposed throughout the diffusion, `Smiling` is mostly relevant early on. Nevertheless, Fig. 2(right) also shows 25 individual curves to showcase the fact that the dual variable trajectories are highly dependent on the specific image. Hence, manually fine-tuning these parameters is entirely impractical. This issue is quickly exacerbated as more requirements are introduced. Yet, this is not an issue for PD-CDS since λ is a parameter automatically adjusted by the algorithm. It can therefore effectively handle all four semantic and photometric constraints at the same time (Fig. 1).

6 Conclusion

PD-CDS addresses key challenges of existing diffusion guidance approaches, some of which depend on sensitive hyperparameters, limit the type and number of constraints, display high computational costs, and/or do not guarantee constraint satisfaction. Based on constrained optimization and duality, it provides a general approach to simultaneously impose a variety of constraints, enabling it to match and even outperform existing methods while reducing computation and memory costs. We expect that certain coarse approximations used in PD-CDS, together with the analysis of their impact on its performance, can still be refined and combined with existing DDPM convergence results to provide a complete description of its dynamics.

AI USE STATEMENT

We have not used generative AI tools for developing theoretical models or conceptualizing the theoretical framework. For the mathematical proofs, generative AI tools were used to assist in verifying algebraic derivations, writing the proof by establishing clear logical transitions between steps, refining intermediate claims, and providing technical ingredients for the mathematical arguments. Additionally, we used generative AI tools to edit software code, identify relevant literature, and assist with copy-editing of the manuscript (not writing). We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI. Other required disclosure tasks are not applicable to this work.

REFERENCES

ahsanjavid. convnext-tiny-finetuned-cifar10. Hugging Face model repository, https://huggingface.co/ahsanjavid/convnext-tiny-finetuned-cifar10, 2022. Apache-2.0 license. Accessed September 2026.

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

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, 2024a.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, 2024b.

D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

D. P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Matthieu Blanke, Yongquan Qu, Sara Shamekh, and Pierre Gentine. Strictly constrained generative modeling via split augmented langevin sampling. In *International Conference on Learning Representations*, 2026.

J. F. Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, 2000.

Vivek S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5): 291–294, 1997.

Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. Ensembling with deep generative views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14997–15007, 2021.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

L. F. O. Chamon, S. Paternain, M. Calvo-Fullana, and A. Ribeiro. Constrained learning with non-convex losses. *IEEE Trans. on Inf. Theory*, 69[3]:1739–1760, 2023.

Luiz F. O. Chamon, Mohammad Reza Karimi, and Anna Korba. Constrained sampling with primal-dual langevin monte carlo. In *Advances in Neural Information Processing*, 2024.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Jacob K. Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing*, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109, 2025.

Alexander Denker, Francisco Vargas, Shreyas Padhy, Kieran Didi, Simon Mathis, Vincent Dutordoir, Riccardo Barbano, Emile Mathieu, Urszula Julia Komorowska, and Pietro Lio. DEFT: Efficient fine-tuning of diffusion models by learning the generalised h-transform. In *Advances in Neural Information Processing*, 2024.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing*, 2021.

Kieran Didi, Francisco Vargas, Simon V. Mathis, Vincent Dutordoir, Emile Mathieu, Urszula J. Komorowska, and Pietro Lio. A framework for conditional diffusion modelling with applications in motif scaffolding for protein design. *arXiv preprint arXiv:2312.09236*, 2023.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2010.

Alex Dytso, H. Vincent Poor, and Shlomo Shamai (Shitz). Conditional mean estimation in Gaussian noise: A meta derivative identity with applications. *IEEE Transactions on Information Theory*, 69 (3):1883–1898, 2023. doi: 10.1109/TIT.2022.3216012.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Samar Hadou, Yigit Berkay Uslu, and Alejandro Ribeiro. Constrained diffusion models with primal-dual inference. *arXiv preprint arXiv:2606.17192*, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing*, 2022.

William Huang, Yifeng Jiang, Tom Van Wouwe, and C. Karen Liu. Constrained diffusion with trust sampling. In *Advances in Neural Information Processing*, 2024.

Vivek Jayaram, Ira Kemelmacher-Shlizerman, Steve Seitz, and John Thickstun. Linearly constrained diffusion implicit models. In *Advances in Neural Information Processing*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing*, 2022.

Shervin Khalafi, Dongsheng Ding, and Alejandro Ribeiro. Constrained diffusion models via dual training. In *Advances in Neural Information Processing*, 2024.

Michel Ledoux. Four talagrand inequalities under the same umbrella. *arXiv preprint arXiv:1909.00363*, 2019.

Xingchao Liu, Xin Tong, and Qiang Liu. Sampling with trustworthy constraints: A variational gradient framework. In *Advances in Neural Information Processing*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1YO4EE3SPB.

Badr Moufad, Yazid Janati, Lisa Bedin, Alain Oliviero Durmus, Randal Douc, Eric Moulines, and Jimmy Olsson. Variational diffusion posterior sampling with midpoint guidance. In *International Conference on Learning Representations*, 2025.

Angelia Nedić and Asuman Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009.

Alain Oliviero-Durmus, Yazid Janati, Eric Moulines, Marcelo Pereyra, and Sebastian Reich. Generative modelling meets bayesian inference: a new paradigm for inverse problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 383(2299): 20240334, 2025.

pymlex. celeba-gan-xai. Hugging Face model repository, https://huggingface.co/pymlex/celeba-gan-xai, 2026. GPL-3.0 license. Accessed September 2026.

C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

L. C. G. Rogers and David Williams. *Diffusions, Markov Processes, and Martingales*, volume 1. Cambridge University Press, 2000.

Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Beyond first-order Tweedie: Solving inverse problems using latent diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9472–9481, 2024.

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

Yifei Shen, Xinyang Jiang, Yifan Yang, Yezhen Wang, Dongqi Han, and Dongsheng Li. Understanding and improving training-free loss-based diffusion guidance. In *Advances in Neural Information Processing*, 2024.

Marta Gentiloni Silveri and Antonio Ocello. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in W2-distance. In *International Conference on Machine Learning*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023a.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, 2023b.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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

Phong Tran, Anh Tuan Tran, Quynh Phung, and Minh Hoai. Explore image deblurring via encoded blur kernel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11956–11965, 2021.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing*, 2023.

Lingxiao Yang, Shutong Ding, Yifan Cai, Jingyi Yu, Jingya Wang, and Ye Shi. Guidance with spherical Gaussian constraint for conditional diffusion. In *International Conference on Machine Learning*, 2024.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing*, 2024.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Stefano Zampini, Jacob K. Christopher, Luca Oneto, Davide Anguita, and Ferdinando Fioretto. Training-free constrained generation with stable diffusion models. In *Advances in Neural Information Processing*, 2025.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20895–20905, June 2025a.

Jichen Zhang, Liqun Zhao, Antonis Papachristodoulou, and Jack Umenberger. Constrained diffusers for safe planning and control. In *Advances in Neural Information Processing*, 2025b.

Qijie Zhu, Zeqi Ye, Han Liu, Zhaoran Wang, and Minshuo Chen. Training-free adaptation of diffusion models via doob's h-transform. In *International Conference on Machine Learning*, 2026.

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

APPENDIX

A Related work **15**

B DDM Background and Notation **16**

B.1 Forward noising process 16

B.2 Reverse DDM transitions 16

C Proofs of Sec. 3 **17**

C.1 Proof of Proposition 3.1 17

C.2 Proof of Proposition 3.2 18

D Proofs of Sec. 4 **19**

D.1 Preliminary Lemmas 19

D.2 Proofs of the main guarantees 24

D.3 Control of the readout and posterior/Tweedie terms 28

D.4 Characterization of the signed sampler-law mismatch 31

D.5 Gaussian bridge and schedule forgetting 52

D.6 Schedule summability and geometric forgetting 63

E Experiments **64**

E.1 Implementation details 64

E.2 Baseline implementation 65

E.3 Extended results 66

E.4 Runtime and memory 67

E.5 Dual dynamics and constraint feasibility 67

E.6 Additional experiments 71

E.7 Sample images 76

A Related work

Training-free guidance. Loss and energy-based guidance methods provide a natural way to steer a fixed pretrained diffusion model using external differentiable objectives. Methods based on conditioning (Chung et al., 2023; Didi et al., 2023; Denker et al., 2024; Zhu et al., 2026; Song et al., 2023a; Moufad et al., 2025), particle-based approaches (Wu et al., 2023), or more general loss- or energy-based mechanisms (Dhariwal & Nichol, 2021; Song et al., 2023b; Yu et al., 2023; Bansal et al., 2024b; Ye et al., 2024) have been proposed. In order to evaluate the exact guidance (i.e., Doob's h-transform) these approaches rely on approximations based on atomic distributions or (pretrained) parametrizations (Didi et al., 2023; Yu et al., 2023; Bansal et al., 2024b; Denker et al., 2024; Shen et al., 2024; Ye et al., 2024; Zhu et al., 2026), Gaussian approximation (Ho et al., 2022; Song et al., 2023a), and variational methods (e.g., (Song et al., 2023b; Moufad et al., 2025)). Training-free guidance often relies on method-specific guidance scales, schedules, or other hyperparameters; TFG explicitly formalizes this design space and introduces a dedicated hyperparameter search procedure (Ye et al., 2024). Many of these methods were developed specifically to handle inverse problem, such as DPS (Chung et al., 2023), DAPS (Zhang et al., 2025a), DiffPIR (Zhu et al., 2023), ΠGDM (Song et al., 2023a), and RED-diff (Mardani et al., 2024).

Optimization-inspired guidance methods have also been proposed. They adapt classical constrained optimization approaches to the diffusion setting, using projections (Lugmayr et al., 2022; Christopher et al., 2024), trust regions (Huang et al., 2024), proximal maps (Jayaram et al., 2025; Zampini et al., 2025), and variable splitting methods (Blanke et al., 2026). Many are limited in the requirements they can impose to functions for which projections or proximal maps exist in closed form. Otherwise, they need to solve an optimization problem at each inference step.

Distributionally constrained sampling. Our constrained diffusion problem is (PI) is closely related to the one studied in (Liu et al., 2021; Chamon et al., 2024), with the distinction that these works consider a Markov Chain Monte Carlo (MCMC) setting where $p_0(x) \propto e^{-f(x)}$ for a known f. In these settings, sampling from p_0 can be done using, e.g., Langevin Monte Carlo (LMC) techniques. Both works develop primal-dual variants of LMC that they analyze under typical MCMC assumptions, such as when p_0 satisfy log-concavity or log-Sobolev inequalities.

Constrained diffusion models. Several recent works incorporate explicit constraints into diffusion modeling. Dual Training couples dual optimization with score-model training, thereby learning a diffusion model adapted to the constrained objective (Khalafi et al., 2024), while Constrained Diffusers introduces constrained inference mechanisms for diffusion-based planning and control (Zhang et al., 2025b). Most closely related to our formulation, Primal-Dual Inference (PDI) considers entropy-regularized optimization under average constraints and jointly adapts the generated distribution and the dual variables during reverse diffusion (Hadou et al., 2026). Its starting point, however, assumes an explicitly specified pointwise objective together with the constraint functions. For each value of the dual variable, these functions define a Gibbs distribution whose energy is known up to normalization, and PDI trains a dual-conditioned score network to represent the corresponding family of targets.

Doob-based diffusion adaptation. A complementary line of work studies principled modifications of diffusion dynamics through Doob's h-transform. DEFT uses this framework for conditional generation, learning a lightweight correction associated with the conditioning information while keeping the pretrained unconditional diffusion model fixed (Denker et al., 2024). More recently, DOIT develops a training-free, simulation-based approximation of the Doob correction to transport a pretrained generative distribution toward a high-reward target distribution (Zhu et al., 2026). These works provide principled mechanisms for adapting diffusion dynamics toward a prescribed conditional or reweighted target.

B DDM BACKGROUND AND NOTATION

This section summarizes the discrete-time DDM conventions and notation used throughout the paper. We distinguish the forward noising process $(\tilde{X}_t)_{t=0}^T$, with marginals $\tilde{\pi}_t$ and path law $\widetilde{\mathbb{P}}$, from the reference reverse chain $(X_t)_{t=T}^0$, with marginals π_t and path law $\mathbb{P}$.

B.1 FORWARD NOISING PROCESS

Using the noise schedule and coefficients introduced in Sec. 2.1, the forward process starts from $\tilde{X}_0 \sim \tilde{\pi}_0$ and evolves according to the Gaussian Markov transitions

$$Q_t(\cdot \mid x_{t-1}) \triangleq \mathcal{N}\big(\sqrt{\alpha_t}\, x_{t-1}, \beta_t I\big), \qquad t = 1, \dots, T.$$

These transitions yield the aggregate conditional distribution

$$\widetilde{\mathbb{P}}(\tilde{X}_t \in \cdot \mid \tilde{X}_0 = x_0) = \mathcal{N}(\rho_t x_0, \sigma_t^2 I),$$

corresponding to (2). Writing $x_{0:T} = (x_0, \dots, x_T)$, the forward path law is

$$\widetilde{\mathbb{P}}(dx_{0:T}) = \tilde{\pi}_0(dx_0) \prod_{t=1}^{T} Q_t(dx_t \mid x_{t-1}),$$

B.2 REVERSE DDM TRANSITIONS

The exact reverse conditional associated with the forward path law is

$$K_t(\cdot \mid x_t) \triangleq \widetilde{\mathbb{P}}(\tilde{X}_{t-1} \in \cdot \mid \tilde{X}_t = x_t).$$

$\tilde{s}_t(x) \triangleq \nabla_x \log \tilde{p}_t(x)$ denote the exact score of $\tilde{\pi}_t$. For integrable $\tilde{X}_0$, Tweedie's identity applied to the one-step Gaussian channel gives

$$\mathbb{E}_{\widetilde{\mathbb{P}}}[\tilde{X}_{t-1} \mid \tilde{X}_t = x] = \frac{x + \beta_t \tilde{s}_t(x)}{\sqrt{\alpha_t}}.$$

Applying Tweedie's identity to the aggregate Gaussian channel (2) also gives the exact clean conditional mean

$$m_{0|t}(x) \triangleq \mathbb{E}_{\widetilde{\mathbb{P}}}[\tilde{X}_0 \mid \tilde{X}_t = x] = \frac{x + \sigma_t^2 \tilde{s}_t(x)}{\rho_t}.$$

The DDM in (1) approximates these reverse conditionals using a pretrained score model which we assume gives $s_\theta(x, t) = \tilde{s}_t(x)$ and Gaussian transitions

$$K_{\theta,t}(\cdot \mid x_t) = \mathcal{N}\big(m_{\theta,t}(x_t), \beta_t I\big), \qquad m_{\theta,t}(x) = \frac{x + \beta_t s_\theta(x, t)}{\sqrt{\alpha_t}}.$$

The covariance $\beta_t I$ is one of the fixed-covariance choices considered by Ho et al. (2020). An exact score recovers the exact conditional mean $m_{0|t}(x)$, but does not generally make $K_{\theta,t} = K_t$: the exact reverse conditional need not be Gaussian or have covariance $\beta_t I$.

The reference reverse chain used in the analysis is generated by these learned transitions, with path law

$$\mathbb{P}(dx_{0:T}) = \pi_T(dx_T) \prod_{t=1}^{T} K_{\theta,t}(dx_{t-1} \mid x_t), \qquad \pi_T = \mathcal{N}(0, I).$$

In particular, π_0 is the learned approximation to the target distribution $\tilde{\pi}_0$ (see Theoretical guarantee after ($\widetilde{\text{DI}}$)).

C PROOFS OF SEC. 3

C.1 PROOF OF PROPOSITION 3.1

Proof. Fix $\lambda \in \mathbb{R}^m_+$. Since g is bounded below π_0-almost surely, $e^{-\lambda^\top g}$ is bounded. Together with π_0-integrability of g, this ensures $0 < Z_\lambda < +\infty$ and $g \in L^1(\mu_0^\lambda)$. For any μ with $g \in L^1(\mu)$, the change-of-measure identity gives

$$L(\mu, \lambda) = \mathrm{KL}(\mu \| \mu_0^\lambda) - \log Z_\lambda,$$

with equality understood in the extended-value sense. Nonnegativity of KL therefore makes μ_0^λ the unique Lagrangian minimizer, and

$$d(\lambda) = L(\mu_0^\lambda, \lambda) = -\log Z_\lambda.$$

On compact subsets of $\mathbb{R}^m_+$, $|g_i| e^{-\lambda^\top g}$ is dominated by an integrable multiple of $|g_i|$. Differentiation under the integral yields

$$\nabla d(\lambda) = \mathbb{E}_{\mu_0^\lambda}[g],$$

with derivatives at the boundary understood within the nonnegative orthant.

Slater's condition and $0 \le P^\star \le \mathrm{KL}(\mu^\dagger \| \pi_0) < +\infty$ give $P^\star = D^\star$ and $\Lambda^\star \neq \varnothing$. The asserted radius bound follows from Lemma C.1.

Finally, for any $\lambda^\star \in \Lambda^\star$, the first-order optimality conditions give

$$\mathbb{E}_{\mu_0^{\lambda^\star}}[g] \preceq 0, \qquad (\lambda^\star)^\top \mathbb{E}_{\mu_0^{\lambda^\star}}[g] = 0.$$

Thus $\mu_0^{\lambda^\star}$ is feasible and

$$\mathrm{KL}(\mu_0^{\lambda^\star} \| \pi_0) = L(\mu_0^{\lambda^\star}, \lambda^\star) = d(\lambda^\star) = P^\star.$$

Every primal optimizer also minimizes $L(\cdot, \lambda^\star)$, whose minimizer is unique. Hence $\mu^\star = \mu_0^{\lambda^\star}$ for every $\lambda^\star \in \Lambda^\star$. ■

Lemma C.1 (Dual radius under Slater's condition)**.** *Under the assumptions of Prop. 3.1, every $\lambda^\star \in \Lambda^\star$ satisfies*

$$\|\lambda^\star\|_2 \le \|\lambda^\star\|_1 \le \frac{\mathrm{KL}(\mu^\dagger \| \pi_0)}{\delta}. \tag{15}$$

Consequently, for any finite $R \ge \mathrm{KL}(\mu^\dagger \| \pi_0)/\delta$, the set $\Lambda \triangleq \{\lambda \in \mathbb{R}^m_+ : \|\lambda\|_2 \le R\}$ is convex and compact, contains 0 and every dual optimizer, and satisfies

$$R_\Lambda \triangleq \sup_{\lambda, \lambda' \in \Lambda} \|\lambda - \lambda'\|_2 \le 2R. \tag{16}$$

Proof. For any $\lambda \in \mathbb{R}^m_+$, evaluating the Lagrangian at the strictly feasible distribution $\mu^\dagger$ gives

$$d(\lambda) \le L(\mu^\dagger, \lambda) \le \mathrm{KL}(\mu^\dagger \| \pi_0) - \delta \|\lambda\|_1.$$

Since $d(0) = 0$, every dual optimizer satisfies

$$0 \le d(\lambda^\star) \le \mathrm{KL}(\mu^\dagger \| \pi_0) - \delta \|\lambda^\star\|_1,$$

which proves (15). The set Λ is the intersection of $\mathbb{R}^m_+$ with the closed Euclidean ball of center 0 and radius R, hence is convex and compact that contains 0; the radius bound ensures that it contains every dual optimizer. The triangle inequality gives (16). In particular, iterates initialized at $\lambda_T = 0$ and updated by projection onto Λ satisfy $\|\lambda_t\|_2 \le R$ and $\|\lambda_t - \lambda^\star\|_2 \le R_\Lambda$. ■

C.2 Proof of Proposition 3.2

Proof. Fix $\lambda \in \mathbb{R}_+^m$, recall that $\mathbb{P}$ denote the reference path law, with marginals π_t and kernels $K_{\theta,t}$. By definition,

$$h_t^\lambda(x) = \mathbb{E}_{\mathbb{P}}\Big[e^{-\lambda^\top g(X_0)} \mid X_t = x\Big], \qquad h_0^\lambda(x) = e^{-\lambda^\top g(x)}.$$

Positivity and integrability of $e^{-\lambda^\top g(X_0)}$ imply $0 < h_t^\lambda < +\infty$ π_t-almost everywhere. Moreover, the tower property gives

$$\int h_T^\lambda(x)\, \pi_T(dx) = \mathbb{E}_{\mathbb{P}}\Big[e^{-\lambda^\top g(X_0)}\Big] = Z_\lambda,$$

so $\mu_T^\lambda(dx) = Z_\lambda^{-1} h_T^\lambda(x)\pi_T(dx)$ is a probability measure.

For $t = 1, \dots, T$, the Markov property and the tower property yield

$$\begin{aligned} h_t^\lambda(x_t) &= \mathbb{E}_{\mathbb{P}}\big[h_{t-1}^\lambda(X_{t-1}) \mid X_t = x_t\big] \\ &= \int h_{t-1}^\lambda(x_{t-1})\, K_{\theta,t}(dx_{t-1} \mid x_t). \end{aligned}$$

Consequently,

$$K_{\theta,t}^\lambda(dx_{t-1} \mid x_t) = \frac{h_{t-1}^\lambda(x_{t-1})}{h_t^\lambda(x_t)}\, K_{\theta,t}(dx_{t-1} \mid x_t)$$

is a probability kernel. All conditional identities hold π_t-almost everywhere; the kernels may be assigned any fixed probability measure on the remaining null sets.

The backward chain with this initialization and these kernels has path law

$$\begin{aligned} \mathbb{P}^\lambda(dx_{0:T}) &= \mu_T^\lambda(dx_T) \prod_{t=1}^T K_{\theta,t}^\lambda(dx_{t-1} \mid x_t) \\ &= \frac{h_T^\lambda(x_T)}{Z_\lambda} \left[\prod_{t=1}^T \frac{h_{t-1}^\lambda(x_{t-1})}{h_t^\lambda(x_t)}\right] \mathbb{P}(dx_{0:T}) \\ &= \frac{e^{-\lambda^\top g(x_0)}}{Z_\lambda}\, \mathbb{P}(dx_{0:T}), \end{aligned}$$

where the factors telescope and we use the expression for h_0^λ.

Conditioning this change-of-measure density on X_t gives

$$\frac{\mathrm{d}\mu_t^\lambda}{\mathrm{d}\pi_t}(x) = \frac{1}{Z_\lambda}\, \mathbb{E}_{\mathbb{P}}\Big[e^{-\lambda^\top g(X_0)} \mid X_t = x\Big] = \frac{h_t^\lambda(x)}{Z_\lambda}.$$

Thus $p_t^\lambda(x) = Z_\lambda^{-1} h_t^\lambda(x) p_t(x)$. At $t = 0$, this becomes $p_0^\lambda(x) = Z_\lambda^{-1} e^{-\lambda^\top g(x)} p_0(x)$, proving that the chain produces the required sample $X_0^\lambda \sim \mu_0^\lambda$. ■

D PROOFS OF SEC. 4

This section establishes an explicit finite-horizon bound for the dual iterates of Alg. 1, showing how the approximation errors, multiplier movement, and initialization contribute to the guarantee. We work with the forward-reference dual objective $\tilde{d}$. The proof is given in Sec. D.2.2; the table below identifies the estimates used in its conclusion.

Theorem D.1 (Explicit finite-horizon dual guarantee)**.** *Choose $R_\Lambda > 0$ and Λ as in Lemma C.1, applied to the reference measure $\tilde{\pi}_0$. Using the results of Remark 4.1, under Ass. 4.2 and 4.3, for a horizon $T \geq 1$ and a schedule satisfying $0 < \beta_{\min} \leq \beta_t \leq \beta_{\max} < 1$, with $\beta_{\min}$ and $\beta_{\max}$ independent of T, and constant dual step size*

$$\eta = \frac{2R}{G\sqrt{T}}. \tag{17}$$

For any maximizer $\lambda^\star$ of $\tilde{d}$ over $\mathbb{R}^m_+$, its iterates satisfy

$$\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}\Big[\tilde{d}(\lambda^\star) - \tilde{d}(\lambda_k)\Big] \leq C + \frac{A}{\sqrt{T}} + \frac{B}{T}, \tag{18}$$

where

$$\begin{aligned} C &:= R_\Lambda\big(\varepsilon_{\mathrm{read}} + L_g\varepsilon_{\mathrm{TW}}\big) + \Delta_{\mathrm{ker}}, \\ A &:= 2RG + \frac{2R}{G}A^\chi_{\mathrm{mov}}, \\ B &:= \frac{4R^2}{G^2}B^\chi_{\mathrm{mov}} + C_{\mathrm{init}}. \end{aligned} \tag{19}$$

Where A and B are bounded independently of T. Moreover, $\Delta_{\mathrm{ker}} = O(R(\varepsilon_h + \varepsilon_{\mathrm{GD}}))$, with an implicit constant depending only on $G, d, R_X, \beta_{\min}, \beta_{\max}$. The remaining contributions satisfy $\varepsilon_{\mathrm{read}} \leq 2G$ and $\varepsilon_{\mathrm{TW}} \leq R_X$. The proof is given in Sec. D.2.2, and the component bounds are defined in the propositions listed below.

Quantity	Meaning	Definition and control
$\varepsilon_{\mathrm{read}}, \varepsilon_{\mathrm{TW}}$	Bounds for the readout and posterior approximation errors	Prop. D.5 and D.7
$A^\chi_{\mathrm{mov}}, B^\chi_{\mathrm{mov}}$	Coefficients of the linear and quadratic multiplier-movement terms	Prop. D.16 in Sec. D.4.9
Δ_{ker}	Kernel approximation contribution, including forgetting across non-adjacent steps	Prop. D.16 in Sec. D.4.9
C_{init}	Gaussian initialization contribution	Prop. D.15 in Sec. D.4.8

D.1 PRELIMINARY LEMMAS

This section proves comparison-statistic regularity, the compact projection construction. We use the notation and the laws introduced in the main text. In particular, to simplify notation $\mathbb{E}_\lambda$ denotes expectation under the ideal tilted path law $\tilde{P}^\lambda$ for a fixed multiplier λ, whereas $\mathbb{E}$ denotes expectation under the joint law of the adaptive sampler. The lemmas below establish the regularity of the comparison statistic and a moment bound for the implemented sampler. These results are used in the subsequent proof of the main theorem and in the sampler-law mismatch analysis.

We define the following quantities :

$$f^\lambda_k = g(\widehat{x}^\lambda_{0|k}) \quad \text{and} \quad \chi^\lambda_k = g(\tilde{m}^\lambda_{0|k}) \tag{20}$$

For a probability measure μ, a Markov kernel K, and a measurable test function φ, we write

$$\mu\varphi := \int \varphi(x)\,\mu(\mathrm{d}x), \qquad (K\varphi)(x) := \int \varphi(y)\,K(\mathrm{d}y \mid x), \qquad \mu K\varphi := \int (K\varphi)(x)\,\mu(\mathrm{d}x).$$

D.1.1 PROOFS OF REGULARITY, COMPACT PROJECTION, AND MOMENT CONTROL

Lemma D.2 (Uniform multiplier regularity of the comparison statistic)**.** *For every $k = 1, \dots, T$, every x, and every $\lambda \in \Lambda$,*

$$\boxed{\left\|J_\lambda \tilde{m}^\lambda_{0|k}(x)\right\|_{\mathrm{op}} \le R_X G, \qquad \left\|J_\lambda \chi^\lambda_k(x)\right\|_{\mathrm{op}} \le L_\chi, \quad L_\chi := L_g R_X G.} \tag{21}$$

In particular, L_χ is independent of k, ρ_k, σ_k, and the horizon T.

Proof. We first record the consequences of regularity used throughout the proofs. The Jacobian bound makes g L_g-Lipschitz. Since $f^\lambda_k = g(\widehat{x}^\lambda_{0|k})$ and $\chi^\lambda_k = g(\tilde{m}^\lambda_{0|k})$, the global bound on g gives $\sup_{k,x,\lambda} \left\|f^\lambda_k(x)\right\| \le G$ and $\sup_{k,x,\lambda} \left\|\chi^\lambda_k(x)\right\| \le G$. The support bound on $\tilde{X}_0$ also holds under $\tilde{P}^\lambda$, since $\tilde{P}^\lambda$ is absolutely continuous with respect to $\tilde{P}$.

Fix k and x. We first derive the multiplier derivative of the exact tilted posterior mean $\tilde{m}^\lambda_{0|k}(x) := \mathbb{E}_\lambda[\tilde{X}_0 \mid \tilde{X}_k = x]$. Write the reference posterior at the fixed state x as

$$\nu_{k,x}(\mathrm{d}x_0) := \tilde{P}(\mathrm{d}x_0 \mid \tilde{X}_k = x). \tag{22}$$

Under the tilted path law, conditioning on $\tilde{X}_k = x$ gives

$$\tilde{P}^\lambda(\mathrm{d}x_0 \mid \tilde{X}_k = x) = \frac{e^{-\lambda^\top g(x_0)} \nu_{k,x}(\mathrm{d}x_0)}{\tilde{h}^\lambda_k(x)}, \qquad \tilde{h}^\lambda_k(x) = \int e^{-\lambda^\top g(z)} \nu_{k,x}(\mathrm{d}z). \tag{23}$$

The denominator is precisely the Doob function. We define the unnormalized first moment

$$n^\lambda_{0|k}(x) := \int x_0 e^{-\lambda^\top g(x_0)} \nu_{k,x}(\mathrm{d}x_0). \tag{24}$$

Using the definition of expectation we obtain

$$\boxed{\tilde{m}^\lambda_{0|k}(x) = \frac{n^\lambda_{0|k}(x)}{\tilde{h}^\lambda_k(x)}.} \tag{25}$$

We now differentiate this quotient in an arbitrary multiplier direction $v \in \mathbb{R}^m$. The notation

$$J_\lambda \tilde{m}^\lambda_{0|k}(x) v$$

is the directional derivative. With k and x held fixed, replace λ with $\lambda + sv$ and differentiate at $s = 0$. Since

$$\frac{\mathrm{d}}{\mathrm{d}s} e^{-(\lambda+sv)^\top g(x_0)}\Big|_{s=0} = -\big(v^\top g(x_0)\big) e^{-\lambda^\top g(x_0)},$$

we obtain

$$J_\lambda n^\lambda_{0|k}(x) v = -\int x_0 \left(v^\top g(x_0)\right) e^{-\lambda^\top g(x_0)} \nu_{k,x}(\mathrm{d}x_0), \tag{26}$$

$$(\nabla_\lambda \tilde{h}^\lambda_k(x))^\top v = -\int \big(v^\top g(x_0)\big) e^{-\lambda^\top g(x_0)} \nu_{k,x}(\mathrm{d}x_0). \tag{27}$$

Because g and $\tilde{X}_0$ are bounded under our assumptions, these derivatives may be passed under the integrals by dominated convergence.

Applying the quotient rule to (25) gives

$$\begin{aligned} J_\lambda \tilde{m}^\lambda_{0|k}(x) v &= \frac{J_\lambda n^\lambda_{0|k}(x) v\, \tilde{h}^\lambda_k(x) - n^\lambda_{0|k}(x)\, (\nabla_\lambda \tilde{h}^\lambda_k(x))^\top v}{[\tilde{h}^\lambda_k(x)]^2} \\ &= -\,\mathbb{E}_\lambda\Big[\tilde{X}_0\, v^\top g(\tilde{X}_0) \mid \tilde{X}_k = x\Big] + \mathbb{E}_\lambda[\tilde{X}_0 \mid \tilde{X}_k = x]\ \mathbb{E}_\lambda[v^\top g(\tilde{X}_0) \mid \tilde{X}_k = x] \\ &= -\,\mathrm{Cov}_\lambda\Big(\tilde{X}_0, v^\top g(\tilde{X}_0) \mid \tilde{X}_k = x\Big). \end{aligned} \tag{28}$$

Here the covariance in the last line is a vector in $\mathbb{R}^d$. Indeed it is the covariance between the vector $\tilde{X}_0$ and the scalar $v^\top g(\tilde{X}_0)$. This establishes the derivative identity used below.

The Jacobian $J_\lambda \tilde{m}^\lambda_{0|k}(x)$ is a linear map from multiplier space $\mathbb{R}^m$ to state space $\mathbb{R}^d$, hence it can be viewed as a $d \times m$ matrix. A unit vector $v \in \mathbb{R}^m$ chooses a direction in multiplier space, so

$$J_\lambda \tilde{m}^\lambda_{0|k}(x) v \in \mathbb{R}^d$$

is the derivative of the posterior mean in that direction. To measure the Euclidean norm of the resulting d-vector, we use the dual representation

$$\|z\| = \sup_{\|u\|=1} \left|u^\top z\right|, \qquad z \in \mathbb{R}^d.$$

Consequently,

$$\boxed{\left\|J_\lambda \tilde{m}^\lambda_{0|k}(x)\right\|_{\mathrm{op}} = \sup_{\|v\|=1} \left\|J_\lambda \tilde{m}^\lambda_{0|k}(x) v\right\| = \sup_{\substack{\|u\|=1 \\ \|v\|=1}} \left|u^\top J_\lambda \tilde{m}^\lambda_{0|k}(x) v\right|.} \tag{29}$$

Fixing now any such unit vectors $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^m$. Projecting the vector covariance identity (28) onto u gives

$$\begin{aligned}\left|u^\top J_\lambda \tilde{m}^\lambda_{0|k}(x) v\right| &= \left|\mathrm{Cov}_\lambda(u^\top \tilde{X}_0, v^\top g(\tilde{X}_0) \mid \tilde{X}_k = x)\right| \\ &\le \sqrt{\mathrm{Var}_\lambda(u^\top \tilde{X}_0 \mid \tilde{X}_k = x)} \sqrt{\mathrm{Var}_\lambda(v^\top g(\tilde{X}_0) \mid \tilde{X}_k = x)}.\end{aligned}$$

The last line is conditional Cauchy–Schwarz. Since u is unit length and $\left\|\tilde{X}_0\right\| \le R_X$ almost surely we obtain

$$\left|u^\top \tilde{X}_0\right| \le \|u\| \left\|\tilde{X}_0\right\| \le R_X,$$

and therefore

$$\mathrm{Var}_\lambda(u^\top \tilde{X}_0 \mid \tilde{X}_k = x) \le \mathbb{E}_\lambda[(u^\top \tilde{X}_0)^2 \mid \tilde{X}_k = x] \le R_X^2.$$

Similarly, as v is unit length and $\left\|g(\tilde{X}_0)\right\| \le G$ we obtain

$$\mathrm{Var}_\lambda(v^\top g(\tilde{X}_0) \mid \tilde{X}_k = x) \le G^2.$$

Hence, for every pair of unit vectors u, v

$$\left|u^\top J_\lambda \tilde{m}^\lambda_{0|k}(x) v\right| \le R_X G. \tag{30}$$

The same bound holds for every unit pair and its right-hand side is independent of the directions chosen. Taking the supremum over those pairs in (29) therefore yields

$$\boxed{\left\|J_\lambda \tilde{m}^\lambda_{0|k}(x)\right\|_{\mathrm{op}} \le R_X G.} \tag{31}$$

Finally, recalling the definition

$$\chi^\lambda_k(x) = g(\tilde{m}^\lambda_{0|k}(x)).$$

The chain rule implies yield

$$J_\lambda \chi^\lambda_k(x) = J_g(\tilde{m}^\lambda_{0|k}(x)) J_\lambda \tilde{m}^\lambda_{0|k}(x). \tag{32}$$

Therefore, using submultiplicativity of the operator norm we obtain

$$\left\|J_\lambda \chi^\lambda_k(x)\right\|_{\mathrm{op}} \le \left\|J_g(\tilde{m}^\lambda_{0|k}(x))\right\|_{\mathrm{op}} \left\|J_\lambda \tilde{m}^\lambda_{0|k}(x)\right\|_{\mathrm{op}} \le L_g R_X G = L_\chi.$$

This bound holds for every multiplier in Λ. Integrating the derivative along the segment joining λ' and λ in the convex set Λ gives, for fixed k, x give

$$\begin{aligned}\left\|\tilde{m}^\lambda_{0|k}(x) - \tilde{m}^{\lambda'}_{0|k}(x)\right\| &\le R_X G \left\|\lambda - \lambda'\right\|, \\ \left\|\chi^\lambda_k(x) - \chi^{\lambda'}_k(x)\right\| &\le L_\chi \left\|\lambda - \lambda'\right\|.\end{aligned}$$

The derivative-to-Lipschitz argument is also detailed in Prop. D.25. ■

Lemma D.3 (First-moment bound for the implemented DDPM sampler)**.** *Under the exact forward-score, regularity, and support conditions of Ass. 4.2 and Ass. 4.3, there is an explicitly constructed constant $M_{\mathrm{alg}} < \infty$ such that*

$$\boxed{\max_{1\le r\le T} \mathbb{E}\left\|\widehat{X}_r\right\| \le M_{\mathrm{alg}}.} \tag{33}$$

This constant depends on the prescribed schedule, covariances, d, R_X, L_g, and R_Λ, and can be chosen independently of the dual step size. If $0 < \beta_{\min} \le \beta_t \le \beta_{\max} < 1$ uniformly in t and T, it can also be chosen independently of T.

Proof. For the practical guidance field,

$$s_k^\lambda = -J_g(\widehat{x}_{0|k}(x))^\top \lambda.$$

Hence, for the adaptive multiplier,

$$\left\|s_k^\lambda(\widehat{X}_k)\right\| \le L_g \left\|\lambda_k\right\| \le L_g R_\Lambda \qquad \text{a.s.} \tag{34}$$

Recall that

$$\rho_k = \sqrt{\bar\alpha_k}, \qquad \sigma_k^2 = 1 - \bar\alpha_k, \qquad \bar\alpha_k = \alpha_k \bar\alpha_{k-1}, \qquad \alpha_k = 1 - \beta_k.$$

By the exact reference Tweedie identity,

$$\tilde m^0_{0|k}(x) = \frac{x + \sigma_k^2 s_\theta(x,k)}{\rho_k}, \quad \text{or equivalently,} \quad s_\theta(x,k) = \frac{\rho_k \tilde m^0_{0|k}(x) - x}{\sigma_k^2}.$$

Moreover,

$$\left\|\tilde m^0_{0|k}(x)\right\| \le R_X,$$

since $\tilde m^0_{0|k}(x) = \mathbb{E}_{\tilde P}[\tilde X_0 \mid \tilde X_k = x]$ and $\left\|\tilde X_0\right\| \le R_X$.

We now rewrite the unguided DDPM mean. Substituting the preceding score identity gives

$$\frac{x + \beta_k s_\theta(x,k)}{\sqrt{\alpha_k}} = \frac{1}{\sqrt{\alpha_k}}\left[x + \frac{\beta_k}{\sigma_k^2}\left(\rho_k \tilde m^0_{0|k}(x) - x\right)\right] = \frac{1}{\sqrt{\alpha_k}}\left(1 - \frac{\beta_k}{\sigma_k^2}\right) x + \frac{\beta_k \rho_k}{\sqrt{\alpha_k}\sigma_k^2} \tilde m^0_{0|k}(x).$$

Using

$$\sigma_k^2 = \alpha_k \sigma_{k-1}^2 + \beta_k, \qquad \rho_k = \sqrt{\alpha_k}\rho_{k-1},$$

we have

$$1 - \frac{\beta_k}{\sigma_k^2} = \frac{\alpha_k \sigma_{k-1}^2}{\sigma_k^2},$$

and therefore

$$\frac{x + \beta_k s_\theta(x,k)}{\sqrt{\alpha_k}} = \frac{\sqrt{\alpha_k}\sigma_{k-1}^2}{\sigma_k^2}\, x + \frac{\beta_k \rho_{k-1}}{\sigma_k^2}\, \tilde m^0_{0|k}(x). \tag{35}$$

For $k = T, \dots, 2$, the implemented reverse transition can thus be written

$$\widehat{X}_{k-1} = \frac{\sqrt{\alpha_k}\sigma_{k-1}^2}{\sigma_k^2}\widehat{X}_k + \frac{\beta_k \rho_{k-1}}{\sigma_k^2}\tilde m^0_{0|k}(\widehat{X}_k) + \frac{\beta_k}{\sqrt{\alpha_k}}\, s_k^\lambda(\widehat{X}_k) + \Sigma_k^{1/2} Z_k, \tag{36}$$

where $Z_k \sim \mathcal{N}(0, I_d)$ is fresh Gaussian noise.

For the Gaussian term, Cauchy–Schwarz gives

$$\mathbb{E}\left\|\Sigma_k^{1/2} Z_k\right\| \le \sqrt{\mathbb{E}\left\|\Sigma_k^{1/2} Z_k\right\|^2}.$$

Moreover, using $Z_k^\top \Sigma_k Z_k = \mathrm{tr}\big(\Sigma_k Z_k Z_k^\top\big)$ and $\mathbb{E}[Z_k Z_k^\top] = I_d$, we obtain

$$\mathbb{E}\left\|\Sigma_k^{1/2} Z_k\right\|^2 = \mathbb{E}[Z_k^\top \Sigma_k Z_k] = \mathrm{tr}(\Sigma_k),$$

Hence

$$\mathbb{E}\left\|\Sigma_k^{1/2} Z_k\right\| \le \sqrt{\mathrm{tr}(\Sigma_k)}.$$

Applying the triangle inequality to (36), together with $\left\|\tilde{m}_{0|k}^0\right\| \le R_X$ and (34), yields

$$\mathbb{E}\left\|\widehat{X}_{k-1}\right\| \le \frac{\sqrt{\alpha_k}\sigma_{k-1}^2}{\sigma_k^2}\,\mathbb{E}\left\|\widehat{X}_k\right\| + \frac{\beta_k \rho_{k-1}}{\sigma_k^2} R_X + \frac{\beta_k}{\sqrt{\alpha_k}} L_g R_\Lambda + \sqrt{\mathrm{tr}(\Sigma_k)}. \tag{37}$$

We therefore define the deterministic backward envelope as

$$\begin{aligned} u_T &:= \sqrt{d}, \\ u_{k-1} &:= \frac{\sqrt{\alpha_k}\sigma_{k-1}^2}{\sigma_k^2} u_k + \frac{\beta_k \rho_{k-1}}{\sigma_k^2} R_X + \frac{\beta_k}{\sqrt{\alpha_k}} L_g R_\Lambda + \sqrt{\mathrm{tr}(\Sigma_k)}, \qquad k = T, \dots, 2. \end{aligned} \tag{38}$$

Since $\widehat{X}_T \sim \mathcal{N}(0, I_d)$,

$$\mathbb{E}\left\|\widehat{X}_T\right\| \le \sqrt{\mathbb{E}\left\|\widehat{X}_T\right\|^2} = \sqrt{d} = u_T.$$

If $\mathbb{E}\left\|\widehat{X}_k\right\| \le u_k$, then (37) implies $\mathbb{E}\left\|\widehat{X}_{k-1}\right\| \le u_{k-1}$. Backward induction therefore gives

$$\sup_{1 \le k \le T} \mathbb{E}\left\|\widehat{X}_k\right\| \le M_{\mathrm{alg}}, \qquad M_{\mathrm{alg}} := \max_{1 \le k \le T} u_k < \infty.$$

Finiteness follows from the finite horizon and from the finiteness of the schedule coefficients and covariance traces. Under the additional uniform bounds $0 < \beta_{\min} \le \beta_k \le \beta_{\max} < 1$, the same envelope has a horizon-independent bound. ■

D.2 Proofs of the main guarantees

D.2.1 Preliminary identities for the dual-ascent argument

For the dual argument, set

$$\mathcal{H}_k := \sigma(\lambda_T, \lambda_{T-1}, \dots, \lambda_k) \quad (0 \le k \le T),$$

and define :

$$\widehat{P}_k^+ := \mathrm{Law}(\widehat{X}_k \mid \mathcal{H}_k) \quad (1 \le k \le T), \qquad \text{and} \qquad \widehat{P}_k^- := \mathrm{Law}(\widehat{X}_k \mid \mathcal{H}_{k+1}) \quad (1 \le k < T).$$

At the initial level, $\mathcal{H}_T = \sigma(\lambda_T)$ is trivial since $\lambda_T = 0$ is deterministic, and $\widehat{P}_T^+ = \mathcal{N}(0, I_d)$. No pre-selection law at level T is needed. Thus $\widehat{P}_k^-$ and $\widehat{P}_k^+$ are the pre- and post-selection conditional laws.

The fresh noise in Alg. 1 gives, for $1 \le r < T$,

$$\widehat{P}_r^- = \widehat{P}_{r+1}^+ \widehat{K}_{\theta,r+1}^{\lambda_{r+1}} \quad \text{almost surely.} \tag{39}$$

The stochastic dual increment used by the algorithm is

$$\widehat{G}_k := f_k^{\lambda_k}(\widehat{X}_k) = g\Big(\widehat{x}_{0|k}^{\lambda_k}(\widehat{X}_k)\Big) \in \mathbb{R}^m.$$

It is the constraint statistic evaluated at the current noisy state. Thus, $\widehat{G}_k$ is exactly the vector used in the projected dual-ascent step. Its centered fluctuation is

$$\boxed{\xi_k := \widehat{G}_k - \mathbb{E}[\widehat{G}_k \mid \mathcal{H}_k] = \widehat{G}_k - \widehat{P}_k^+ f_k^{\lambda_k}.} \tag{40}$$

To compare the stochastic increment with the exact dual gradient, we insert the same comparison statistic $\chi_k^{\lambda_k}$ under the practical and ideal laws. Explicitly,

$$\begin{aligned}
\widehat{G}_k - \nabla_\lambda \tilde{d}(\lambda_k) = {} & \underbrace{\big(\widehat{G}_k - \widehat{P}_k^+ f_k^{\lambda_k}\big)}_{\xi_k} \\
& + \underbrace{\big(\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k}\big)\chi_k^{\lambda_k}}_{\text{law mismatch}} \\
& + \underbrace{\widehat{P}_k^+\big(f_k^{\lambda_k} - \chi_k^{\lambda_k}\big)}_{R_k^{\mathrm{read}}} \\
& + \underbrace{\big(\tilde{\pi}_k^{\lambda_k}\chi_k^{\lambda_k} - \nabla_\lambda \tilde{d}(\lambda_k)\big)}_{R_k^{\mathrm{post}}}.
\end{aligned} \tag{41}$$

Hence

$$\boxed{\widehat{G}_k - \nabla_\lambda \tilde{d}(\lambda_k) = \xi_k + \big(\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k}\big)\chi_k^{\lambda_k} + R_k^{\mathrm{read}} + R_k^{\mathrm{post}}.} \tag{42}$$

The four pieces have distinct roles. ξ_k is a conditionally centered fluctuation. The explicit term $(\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k})\chi_k^{\lambda_k}$ is a pure law mismatch, because the practical and ideal laws are tested with the same comparison statistic. R_k^{read} is the practical readout discrepancy. Finally, R_k^{post} is the gap between g of the exact tilted posterior mean and the exact dual signal.

Lemma D.4 (Conditional-centering cancellation)**.** *For every k,*

$$\boxed{\mathbb{E}\,\langle \lambda_k - \lambda^\star, \xi_k \rangle = 0.} \tag{43}$$

Proof. By (40), $\xi_k = \widehat{G}_k - \mathbb{E}[\widehat{G}_k \mid \mathcal{H}_k]$, hence $\mathbb{E}[\xi_k \mid \mathcal{H}_k] = 0$. Since $\lambda_k - \lambda^\star$ is $\mathcal{H}_k$-measurable,

$$\begin{aligned}
\mathbb{E}\,\langle \lambda_k - \lambda^\star, \xi_k \rangle &= \mathbb{E}[\mathbb{E}[\langle \lambda_k - \lambda^\star, \xi_k \rangle | \mathcal{H}_k]] \\
&= \mathbb{E}[\langle \lambda_k - \lambda^\star, \mathbb{E}[\xi_k \mid \mathcal{H}_k] \rangle] = 0.
\end{aligned}$$

■

We also record the multiplier movement directly from the algorithm. Define

$$\Delta_r := \lambda_r - \lambda_{r+1}, \qquad 0 \le r \le T-1. \tag{44}$$

The update at level $r+1$ is

$$\lambda_r = \Pi_\Lambda\Big(\lambda_{r+1} + \eta f_{r+1}^{\lambda_{r+1}}(\widehat{X}_{r+1})\Big).$$

Since $\lambda_{r+1} \in \Lambda$, we have $\Pi_\Lambda(\lambda_{r+1}) = \lambda_{r+1}$. Non-expansiveness of the projection and the statistic bound $\|f_j^\lambda(x)\| \le G$ from 4.2 give

$$\boxed{\|\Delta_r\| \le \eta \left\| f_{r+1}^{\lambda_{r+1}}(\widehat{X}_{r+1}) \right\| \le \eta G.} \tag{45}$$

D.2.2 Proof of the main theorem

Proof of Thm. D.1. Euclidean projection is non-expansive. Since $\lambda^\star \in \Lambda$,

$$\begin{aligned}
\|\lambda_{k-1} - \lambda^\star\|^2 &= \left\| \Pi_\Lambda(\lambda_k + \eta \widehat{G}_k) - \Pi_\Lambda(\lambda^\star) \right\|^2 \\
&\le \left\| \lambda_k + \eta \widehat{G}_k - \lambda^\star \right\|^2 \\
&= \|\lambda_k - \lambda^\star\|^2 + 2\eta \left\langle \widehat{G}_k, \lambda_k - \lambda^\star \right\rangle + \eta^2 \left\| \widehat{G}_k \right\|^2.
\end{aligned} \tag{46}$$

Equation (42) gives

$$\widehat{G}_k = \nabla_\lambda \tilde{d}(\lambda_k) + \xi_k + (\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k})\chi_k^{\lambda_k} + R_k^{\mathrm{read}} + R_k^{\mathrm{post}}.$$

Concavity gives

$$\left\langle \nabla_\lambda \tilde{d}(\lambda_k), \lambda_k - \lambda^\star \right\rangle \le \tilde{d}(\lambda_k) - \tilde{d}(\lambda^\star).$$

Moreover $\left\| \widehat{G}_k \right\| = \left\| f_k^{\lambda_k}(\widehat{X}_k) \right\| \le G$. Rearranging therefore gives

$$\begin{aligned}
2\eta[\tilde{d}(\lambda^\star) - \tilde{d}(\lambda_k)] \le\ & \|\lambda_k - \lambda^\star\|^2 - \|\lambda_{k-1} - \lambda^\star\|^2 \\
&+ 2\eta \langle \xi_k, \lambda_k - \lambda^\star \rangle + 2\eta \left\langle (\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k})\chi_k^{\lambda_k}, \lambda_k - \lambda^\star \right\rangle \\
&+ 2\eta \left\langle R_k^{\mathrm{read}}, \lambda_k - \lambda^\star \right\rangle + 2\eta \left\langle R_k^{\mathrm{post}}, \lambda_k - \lambda^\star \right\rangle + \eta^2 G^2.
\end{aligned} \tag{47}$$

Taking expectation removes the ξ_k term by Lemma D.4. For the two approximation terms,

$$\left\langle \lambda_k - \lambda^\star, R_k^{\mathrm{read}} \right\rangle \le R_\Lambda \left\| R_k^{\mathrm{read}} \right\|, \qquad \left\langle \lambda_k - \lambda^\star, R_k^{\mathrm{post}} \right\rangle \le R_\Lambda \left\| R_k^{\mathrm{post}} \right\|.$$

Although the algorithm runs in decreasing level order, the finite sum of squared-distance differences has the usual telescoping form:

$$\begin{aligned}
&\sum_{k=1}^{T} \left(\|\lambda_k - \lambda^\star\|^2 - \|\lambda_{k-1} - \lambda^\star\|^2 \right) \\
&\qquad = \|\lambda_T - \lambda^\star\|^2 - \|\lambda_0 - \lambda^\star\|^2 \le R_\Lambda^2.
\end{aligned}$$

The negative final-distance term can be discarded, and the initial multiplier λ_T and optimizer $\lambda^\star$ lie in the set of diameter R_Λ. Summing (47), taking expectation, and dividing by $2\eta T$ therefore gives the signed projected-ascent estimate

$$\boxed{\begin{aligned}
\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}[\tilde{d}(\lambda^\star) - \tilde{d}(\lambda_k)] \le\ & \frac{R_\Lambda^2}{2\eta T} + \frac{\eta G^2}{2} \\
&+ \frac{1}{T}\sum_{k=1}^{T} \mathbb{E}\left\langle \lambda_k - \lambda^\star, (\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k})\chi_k^{\lambda_k} \right\rangle \\
&+ \frac{R_\Lambda}{T}\sum_{k=1}^{T} \mathbb{E}\big(\left\| R_k^{\mathrm{read}} \right\| + \left\| R_k^{\mathrm{post}} \right\| \big).
\end{aligned}} \tag{48}$$

Here the conditional-law and integration conventions used below are

$$\widehat{P}_k^+ \varphi^{\lambda_k} = \mathbb{E}[\varphi^{\lambda_k}(\widehat{X}_k) \mid \mathcal{H}_k], \qquad \tilde{\pi}_k^{\lambda_k} \varphi^{\lambda_k} \quad = \int \varphi^{\lambda}(x) \tilde{\pi}_k^{\lambda}(\mathrm{d}x) \Big|_{\lambda=\lambda_k}.$$

For the signed contribution, retain the mismatch

$$\mathcal{L}_T^{\chi} := \frac{1}{T} \sum_{k=1}^{T} \mathbb{E} \left\langle \lambda_k - \lambda^\star, \chi_k^{\lambda_k}(\widehat{X}_k) - \tilde{\pi}_k^{\lambda_k} \chi_k^{\lambda_k} \right\rangle. \tag{49}$$

We now identify this quantity with the conditional-law expression in the projected-ascent bound. Since λ_k is $\mathcal{H}_k$-measurable and $\tilde{\pi}_k^{\lambda_k} \chi_k^{\lambda_k}$ is a deterministic function of λ_k, the tower property gives

$$\begin{aligned} \mathbb{E} \left\langle \lambda_k - \lambda^\star, \chi_k^{\lambda_k}(\widehat{X}_k) - \tilde{\pi}_k^{\lambda_k} \chi_k^{\lambda_k} \right\rangle &= \mathbb{E} \left\langle \lambda_k - \lambda^\star, \widehat{P}_k^+ \chi_k^{\lambda_k} - \tilde{\pi}_k^{\lambda_k} \chi_k^{\lambda_k} \right\rangle \\ &= \mathbb{E} \left\langle \lambda_k - \lambda^\star, (\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k}) \chi_k^{\lambda_k} \right\rangle. \end{aligned}$$

Hence,

$$\boxed{\mathcal{L}_T^{\chi} = \frac{1}{T} \sum_{k=1}^{T} \mathbb{E} \left\langle \lambda_k - \lambda^\star, (\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k}) \chi_k^{\lambda_k} \right\rangle.} \tag{50}$$

This is an exact identification.

Prop. D.7 in Sec. D.3.2 and Prop. D.5 in Sec. D.3.1 respectively give

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E} \left\| R_k^{\mathrm{read}} \right\| \le \varepsilon_{\mathrm{read}}, \qquad \frac{1}{T} \sum_{k=1}^{T} \mathbb{E} \left\| R_k^{\mathrm{post}} \right\| \le L_g \varepsilon_{\mathrm{TW}}.$$

The mismatch estimate of Prop. D.16 gives

$$\mathcal{L}_T^{\chi} \le A_{\mathrm{mov}}^{\chi} \eta + B_{\mathrm{mov}}^{\chi} \eta^2 + \Delta_{\mathrm{ker}} + \frac{C_{\mathrm{init}}}{T}.$$

Substituting these three controls into the projected-ascent inequality obtained above yields

$$\begin{aligned} \frac{1}{T} \sum_{k=1}^{T} \mathbb{E}[\tilde{d}(\lambda^\star) - \tilde{d}(\lambda_k)] \le \frac{R_\Lambda^2}{2\eta T} + \frac{\eta G^2}{2} + A_{\mathrm{mov}}^{\chi} \eta + B_{\mathrm{mov}}^{\chi} \eta^2 \\ + R_\Lambda(\varepsilon_{\mathrm{read}} + L_g \varepsilon_{\mathrm{TW}}) + \Delta_{\mathrm{ker}} + \frac{C_{\mathrm{init}}}{T}, \end{aligned}$$

for any constant step size $\eta > 0$. With $\eta = 2R/(G\sqrt{T})$ and $R_\Lambda \le 2R$,

$$\frac{R_\Lambda^2}{2\eta T} + \frac{\eta G^2}{2} = \frac{G}{\sqrt{T}} \left(\frac{R_\Lambda^2}{4R} + R \right) \le \frac{2RG}{\sqrt{T}}.$$

The movement terms become

$$A_{\mathrm{mov}}^{\chi} \eta = \frac{2R A_{\mathrm{mov}}^{\chi}}{G\sqrt{T}}, \qquad B_{\mathrm{mov}}^{\chi} \eta^2 = \frac{4R^2 B_{\mathrm{mov}}^{\chi}}{G^2 T}.$$

Substitution therefore yields

$$\begin{aligned} \frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\Big[\tilde{d}(\lambda^\star) - \tilde{d}(\lambda_k)\Big] &\le R_\Lambda\big(\varepsilon_{\mathrm{read}} + L_g \varepsilon_{\mathrm{TW}}\big) + \Delta_{\mathrm{ker}} \\ &\quad + \frac{2RG + (2R/G) A_{\mathrm{mov}}^{\chi}}{\sqrt{T}} + \frac{(4R^2/G^2) B_{\mathrm{mov}}^{\chi} + C_{\mathrm{init}}}{T} \\ &= C + \frac{A}{\sqrt{T}} + \frac{B}{T}, \end{aligned}$$

with C, A, B as defined in (19), proving (18).

We now bound the coefficients uniformly in T. By Prop. D.26,

$$C_{\mathrm{col}} \le \left(\sqrt{\frac{d}{\beta_{\min}}} + \frac{R_X}{\beta_{\min}}\right) \sum_{\ell=1}^{\infty} (1-\beta_{\min})^{\ell/2}$$
$$= \left(\sqrt{\frac{d}{\beta_{\min}}} + \frac{R_X}{\beta_{\min}}\right) \frac{\sqrt{1-\beta_{\min}}}{1-\sqrt{1-\beta_{\min}}} = O(1).$$

For the DDPM covariances $\Sigma_t = \beta_t I_d$, we have $C_{\mathrm{guide}} \le (1-\beta_{\max})^{-1/2}$. Hence (169), together with $R_\Lambda \le 2R$, gives

$$\Delta_{\mathrm{ker}} \le R_\Lambda G(2 + C_{\mathrm{col}}) \left(\frac{\varepsilon_h}{\sqrt{1-\beta_{\max}}} + \varepsilon_{\mathrm{GD}}\right)$$
$$= O(R(\varepsilon_h + \varepsilon_{\mathrm{GD}}))\,.$$

Moreover, Prop. D.5 and D.7 give $\varepsilon_{\mathrm{read}} \le 2G$ and $\varepsilon_{\mathrm{TW}} \le R_X$, so

$$C = R_\Lambda(\varepsilon_{\mathrm{read}} + L_g \varepsilon_{\mathrm{TW}}) + \Delta_{\mathrm{ker}} \le R_\Lambda(2G + L_g R_X) + \Delta_{\mathrm{ker}} = O(R(1 + \varepsilon_h + \varepsilon_{\mathrm{GD}}))\,.$$

Under the same uniform schedule bounds, Prop. D.15, D.16 and D.26 give $C_{\mathrm{init}}, A^{\chi}_{\mathrm{mov}}, B^{\chi}_{\mathrm{mov}} = O(R)$. Consequently, $A, B = O(R)$ and

$$\frac{A}{\sqrt{T}} + \frac{B}{T} = O(T^{-1/2}),$$

which recovers the optimization term stated in Thm. 4.4. All bounds are uniform in T for fixed problem parameters, approximation tolerances, and schedule bounds $\beta_{\min}, \beta_{\max}$. ■

D.3 Control of the readout and posterior/Tweedie terms

The decomposition (42) left two direct approximation terms:

$$R_k^{\mathrm{read}} = \widehat{P}_k^{+}(f_k^{\lambda_k} - \chi_k^{\lambda_k}), \qquad R_k^{\mathrm{post}} = \tilde{\pi}_k^{\lambda_k}\chi_k^{\lambda_k} - \nabla_\lambda \tilde{d}(\lambda_k).$$

They are treated separately because they live under different laws. R_k^{read} is evaluated under the implemented conditional law $\widehat{P}_k^{+}$, while R_k^{post} is an ideal posterior quantity under $\tilde{P}^{\lambda_k}$.

D.3.1 The exact posterior/Tweedie budget

Proposition D.5 (Posterior/Tweedie budget at the prescribed horizon)**.** *Under Ass. 4.2, define*

$$\boxed{\varepsilon_{\mathrm{TW}} := \frac{1}{T}\sum_{k=1}^{T} \min\left\{R_X, \frac{\sqrt{d}\,\sigma_k}{\rho_k}\right\} \le R_X.} \tag{51}$$

Then the ideal posterior term satisfies

$$\boxed{\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\|R_k^{\mathrm{post}}\right\| \le \frac{L_g}{T}\sum_{k=1}^{T}\min\left\{R_X, \frac{\sqrt{d}\,\sigma_k}{\rho_k}\right\} = L_g\varepsilon_{\mathrm{TW}}.} \tag{52}$$

The proof uses the tilted Tweedie identity proved in B.2 and recalled right after, and the posterior-uncertainty Lemma D.6.

We first recall the tilted-score identity:

$$\boxed{\nabla_x \log \tilde{p}_k^{\lambda}(x) = s_\theta(x,k) + \nabla_x \log \tilde{h}_k^{\lambda}(x).} \tag{53}$$

For $k = 1, \dots, T$ (so $\rho_k > 0$ and $\sigma_k > 0$), recall that the tilted Tweedie identity is

$$\boxed{\tilde{m}_{0|k}^{\lambda}(x) = \frac{x + \sigma_k^2 \nabla_x \log \tilde{p}_k^{\lambda}(x)}{\rho_k} = \frac{x + \sigma_k^2[s_\theta(x,k) + \nabla_x \log \tilde{h}_k^{\lambda}(x)]}{\rho_k}.} \tag{54}$$

Posterior uncertainty

Lemma D.6 (Uniformly bounded tilted posterior uncertainty)**.** *For $k = 1, \dots, T$ and $\lambda \in \Lambda$,*

$$\boxed{\mathbb{E}_\lambda\left\|\tilde{X}_0 - \tilde{m}_{0|k}^{\lambda}(\tilde{X}_k)\right\| \le \min\left\{R_X, \frac{\sqrt{d}\,\sigma_k}{\rho_k}\right\}.} \tag{55}$$

Proof. There are two independent bounds.

First, we recall the orthogonality behind the posterior mean. By its definition,

$$\mathbb{E}_\lambda[\tilde{X}_0 - \tilde{m}_{0|k}^{\lambda}(\tilde{X}_k) \mid \tilde{X}_k] = 0.$$

For any square-integrable estimator $a(\tilde{X}_k)$, both $\tilde{m}_{0|k}^{\lambda}(\tilde{X}_k)$ and $a(\tilde{X}_k)$ are measurable with respect to $\tilde{X}_k$. Conditioning the cross term therefore gives

$$\mathbb{E}_\lambda\left[\left\langle \tilde{X}_0 - \tilde{m}_{0|k}^{\lambda}(\tilde{X}_k), \tilde{m}_{0|k}^{\lambda}(\tilde{X}_k) - a(\tilde{X}_k)\right\rangle \middle| \tilde{X}_k\right] = 0.$$

Expanding the squared norm of $\tilde{X}_0 - a(\tilde{X}_k) = (\tilde{X}_0 - \tilde{m}_{0|k}^{\lambda}(\tilde{X}_k)) + (\tilde{m}_{0|k}^{\lambda}(\tilde{X}_k) - a(\tilde{X}_k))$ and taking expectation yields

$$\begin{aligned}\mathbb{E}_\lambda\left\|\tilde{X}_0 - a(\tilde{X}_k)\right\|^2 = \mathbb{E}_\lambda\left\|\tilde{X}_0 - \tilde{m}_{0|k}^{\lambda}(\tilde{X}_k)\right\|^2 \\ + \mathbb{E}_\lambda\left\|\tilde{m}_{0|k}^{\lambda}(\tilde{X}_k) - a(\tilde{X}_k)\right\|^2.\end{aligned} \tag{56}$$

This proves directly that the conditional mean minimizes mean-squared error. Taking $a = 0$ and rearranging gives

$$\mathbb{E}_\lambda \left\| \tilde{X}_0 - \tilde{m}^\lambda_{0|k}(\tilde{X}_k) \right\|^2 = \mathbb{E}_\lambda \left\| \tilde{X}_0 \right\|^2 - \mathbb{E}_\lambda \left\| \tilde{m}^\lambda_{0|k}(\tilde{X}_k) \right\|^2 \leq \mathbb{E}_\lambda \left\| \tilde{X}_0 \right\|^2 \leq R_X^2.$$

Cauchy–Schwarz then gives

$$\mathbb{E}_\lambda \left\| \tilde{X}_0 - \tilde{m}^\lambda_{0|k}(\tilde{X}_k) \right\| \leq R_X. \tag{57}$$

Second, apply (56) with the estimator $a(\tilde{X}_k) = \tilde{X}_k/\rho_k$. The last squared-error term is nonnegative, so

$$\mathbb{E}_\lambda \left\| \tilde{X}_0 - \tilde{m}^\lambda_{0|k}(\tilde{X}_k) \right\|^2 \leq \mathbb{E}_\lambda \left\| \tilde{X}_0 - \tilde{X}_k/\rho_k \right\|^2.$$

Under $\tilde{P}^\lambda$, $\tilde{X}_k = \rho_k \tilde{X}_0 + \sigma_k \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I_d)$ independent of $\tilde{X}_0$, so the right-hand side is $d\sigma_k^2/\rho_k^2$. Another Cauchy–Schwarz step gives

$$\mathbb{E}_\lambda \left\| \tilde{X}_0 - \tilde{m}^\lambda_{0|k}(\tilde{X}_k) \right\| \leq \frac{\sqrt{d}\,\sigma_k}{\rho_k}. \tag{58}$$

Taking the minimum of (57) and (58) proves the lemma. ■

Proof of Prop. D.5. Because $\chi_k^\lambda = g(\tilde{m}^\lambda_{0|k})$ and $\nabla_\lambda \tilde{d}(\lambda) = \mathbb{E}_\lambda[g(\tilde{X}_0)]$,

$$\left\| \tilde{\pi}_k^\lambda \chi_k^\lambda - \nabla_\lambda \tilde{d}(\lambda) \right\| = \left\| \mathbb{E}_\lambda[g(\tilde{m}^\lambda_{0|k}(\tilde{X}_k)) - g(\tilde{X}_0)] \right\| \leq L_g\, \mathbb{E}_\lambda \left\| \tilde{m}^\lambda_{0|k}(\tilde{X}_k) - \tilde{X}_0 \right\|. \tag{59}$$

Using Lemma D.6, evaluating at the random λ_k, taking the outer expectation and averaging over k gives

$$\boxed{\frac{1}{T}\sum_{k=1}^{T} \mathbb{E} \left\| R_k^{\mathrm{post}} \right\| \leq \frac{L_g}{T} \sum_{k=1}^{T} \min\left\{ R_X, \frac{\sqrt{d}\,\sigma_k}{\rho_k} \right\} \leq L_g \varepsilon_{\mathrm{TW}}.} \tag{60}$$

■

D.3.2 The practical readout budget under the implemented law

Recall the pointwise guidance-field error

$$\Delta^\lambda_{h,k}(x) := s_k^\lambda(x) - \nabla_x \log \tilde{h}_k^\lambda(x).$$

Proposition D.7 (Readout budget under the implemented law)**.** *Under Ass. 4.2,*

$$\boxed{\frac{1}{T}\sum_{k=1}^{T} \mathbb{E} \left\| R_k^{\mathrm{read}} \right\| \leq \varepsilon_{\mathrm{read}} \leq 2G,} \tag{61}$$

where

$$\boxed{\varepsilon_{\mathrm{read}} := \frac{1}{T}\sum_{k=1}^{T} \mathbb{E}\left[\widehat{P}_k^+ \min\left\{ 2G, L_g \frac{\sigma_k^2}{\rho_k} \left\| \Delta^{\lambda_k}_{h,k} \right\| \right\} \right].} \tag{62}$$

The minimum is evaluated pointwise before integration under the implemented conditional law. In particular, 4.3 gives the deterministic bound

$$\varepsilon_{\mathrm{read}} \leq \frac{1}{T}\sum_{k=1}^{T} \min\left\{ 2G, L_g \frac{\sigma_k^2}{\rho_k} \varepsilon_h \right\} \leq 2G. \tag{63}$$

Proof. By definition,

$$R_k^{\mathrm{read}} = \widehat{P}_k^{+}(f_k^{\lambda_k} - \chi_k^{\lambda_k}).$$

Recall the norm inequality for integration:

$$\|\mu\varphi\| = \left\| \int \varphi(x)\mu(\mathrm{d}x) \right\| \le \int \|\varphi(x)\|\, \mu(\mathrm{d}x) = \mu\, \|\varphi\|\,.$$

Applying it to the conditional probability measure $\widehat{P}_k^{+}$, and then using the tower property, yields

$$\begin{aligned}
\mathbb{E}\left\|R_k^{\mathrm{read}}\right\| &\le \mathbb{E}\Big[\widehat{P}_k^{+}\left\|f_k^{\lambda_k} - \chi_k^{\lambda_k}\right\|\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\Big\|f_k^{\lambda_k}(\widehat{X}_k) - \chi_k^{\lambda_k}(\widehat{X}_k)\Big\|\Big|\mathcal{H}_k\Big]\Big] \\
&= \mathbb{E}\left\|f_k^{\lambda_k}(\widehat{X}_k) - \chi_k^{\lambda_k}(\widehat{X}_k)\right\|.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\|R_k^{\mathrm{read}}\right\| &\le \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\|f_k^{\lambda_k}(\widehat{X}_k) - \chi_k^{\lambda_k}(\widehat{X}_k)\right\| \\
&= \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\Big[\widehat{P}_k^{+}\left\|f_k^{\lambda_k} - \chi_k^{\lambda_k}\right\|\Big].
\end{aligned} \tag{64}$$

The error is measured in the bounded dual statistic itself.

The relation with the practical guidance discrepancy is pointwise. By (54),

$$\tilde{m}_{0|k}^{\lambda}(x) = \frac{x + \sigma_k^2[s_\theta(x,k) + \nabla_x \log \tilde{h}_k^{\lambda}(x)]}{\rho_k},$$

whereas the practical guided readout is

$$\widehat{x}_{0|k}^{\lambda}(x) = \frac{x + \sigma_k^2[s_\theta(x,k) + s_k^{\lambda}(x)]}{\rho_k}.$$

Subtracting these two identities gives, for every x,

$$\boxed{\widehat{x}_{0|k}^{\lambda}(x) - \tilde{m}_{0|k}^{\lambda}(x) = \frac{\sigma_k^2}{\rho_k}\Delta_{h,k}^{\lambda}(x).} \tag{65}$$

Since g is L_g-Lipschitz,

$$\left\|f_k^{\lambda}(x) - \chi_k^{\lambda}(x)\right\| \le L_g \frac{\sigma_k^2}{\rho_k}\left\|\Delta_{h,k}^{\lambda}(x)\right\|.$$

Each statistic is also bounded by G. Combining the two estimates gives

$$\boxed{\left\|f_k^{\lambda}(x) - \chi_k^{\lambda}(x)\right\| \le \min\left\{2G, L_g\frac{\sigma_k^2}{\rho_k}\left\|\Delta_{h,k}^{\lambda}(x)\right\|\right\}.} \tag{66}$$

At very noisy levels, the factor σ_k^2/ρ_k can produce a large clean-space bound. The global constraint bound still gives a dual-statistic error at most $2G$, as recorded by the minimum above.

Integrating (66) under $\widehat{P}_k^{+}$ at $\lambda = \lambda_k$, then taking the outer expectation and averaging, gives

$$\begin{aligned}
\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\|R_k^{\mathrm{read}}\right\| &\le \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\Big[\widehat{P}_k^{+}\left\|f_k^{\lambda_k} - \chi_k^{\lambda_k}\right\|\Big] \\
&\le \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\widehat{P}_k^{+}\min\left\{2G, L_g\frac{\sigma_k^2}{\rho_k}\left\|\Delta_{h,k}^{\lambda_k}\right\|\right\}\right] \le \varepsilon_{\mathrm{read}}.
\end{aligned}$$

The pointwise minimum is at most $2G$, so its average also satisfies $\varepsilon_{\mathrm{read}} \le 2G$ which lead Prop. D.7.

■

D.4 Characterization of the signed sampler-law mismatch

The exact signed recursion is controlled term by term below and assembled in Prop. D.16. The forgetting inputs are proved in Sec. D.5.

D.4.1 Continuation and centered continuation

For $2 \le t \le T$, the three kernels from level t to $t-1$ are

$$\begin{aligned} \widehat{K}^{\lambda}_{\theta,t} &:= \text{implemented Gaussian reverse kernel with practical guidance},\\ K^{\lambda,\mathrm{G}}_{t} &:= \text{Gaussian reverse kernel with exact guidance } \nabla_x \log \tilde{h}^{\lambda}_t,\\ \tilde{K}^{\lambda}_{t} &:= \text{exact Doob reverse kernel.} \end{aligned} \tag{67}$$

For $1 \le k \le r \le T$, first define the ideal mean of the level-k comparison statistic by

$$T^{\chi}_k(\lambda) := \tilde{\pi}^{\lambda}_k \chi^{\lambda}_k = \mathbb{E}_{\lambda}[\chi^{\lambda}_k(\tilde{X}_k)] \quad \text{where} \chi^{\lambda}_k = g(\tilde{m}^{\lambda}_{0|k}) \tag{68}$$

We define the *continuation* of this statistic from its base level k to a later observation level r and its centered version by

$$\Phi^{\lambda}_{k\to r}(y) := \mathbb{E}_{\lambda}[\chi^{\lambda}_k(\tilde{X}_k) \mid \tilde{X}_r = y]. \tag{69}$$

$$\boxed{\tilde{\Phi}^{\lambda}_{k\to r}(y) := \Phi^{\lambda}_{k\to r}(y) - T^{\chi}_k(\lambda) = \mathbb{E}_{\lambda}\left[\chi^{\lambda}_k(\tilde{X}_k) - T^{\chi}_k(\lambda) \,\middle|\, \tilde{X}_r = y\right].} \tag{70}$$

The arrow $k \to r$ records the direction of propagation. Indeed the observable is fixed at level k, while r is the later noisy level from which that same observable is viewed. The symbol Φ denotes the conditional expectation; $\tilde{\Phi}$ denotes its centered version. In this notation the tilde on Φ records centering, both continuations are defined under $\tilde{P}^{\lambda}$. Boundedness of χ^{λ}_k gives

$$\|\Phi^{\lambda}_{k\to r}\|_{\infty} \le G, \qquad \|\tilde{\Phi}^{\lambda}_{k\to r}\|_{\infty} \le 2G.$$

At the base level and under the ideal level-r marginal, the tower property gives

$$\Phi^{\lambda}_{k\to k} = \chi^{\lambda}_k, \qquad \tilde{\pi}^{\lambda}_r \Phi^{\lambda}_{k\to r} = T^{\chi}_k(\lambda), \tag{71}$$

$$\tilde{\Phi}^{\lambda}_{k\to k} = \chi^{\lambda}_k - T^{\chi}_k(\lambda), \tag{72}$$

$$\tilde{\pi}^{\lambda}_r \tilde{\Phi}^{\lambda}_{k\to r} = 0. \tag{73}$$

Indeed,

$$\begin{aligned} \tilde{\pi}^{\lambda}_r \tilde{\Phi}^{\lambda}_{k\to r} &= \mathbb{E}_{\lambda}\left[\mathbb{E}_{\lambda}[\chi^{\lambda}_k(\tilde{X}_k) - T^{\chi}_k(\lambda) \mid \tilde{X}_r]\right]\\ &= \mathbb{E}_{\lambda}[\chi^{\lambda}_k(\tilde{X}_k)] - T^{\chi}_k(\lambda) = 0. \end{aligned}$$

For $r > k$, Prop. D.23 and D.25 in Sec. D.5 give, uniformly in $\lambda, \lambda' \in \Lambda$,

$$\left\|\tilde{\Phi}^{\lambda}_{k\to r}(y)\right\| \le G\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}), \tag{74}$$

$$\left\|\tilde{\Phi}^{\lambda}_{k\to r}(y) - \tilde{\Phi}^{\lambda'}_{k\to r}(y)\right\| \le C^{\chi}_{\mathrm{der}}\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}})\,\|\lambda - \lambda'\|\,, \tag{75}$$

$$\mathrm{Lip}(\tilde{\Phi}^{\lambda}_{k\to r}) \le G\kappa_{k\to r}. \tag{76}$$

Here $\kappa_{k\to r}$, M_{ref}, and C^{χ}_{der} are the coefficients and constants defined in the same appendix .

D.4.2 EXACT SIGNED WEAK RECURSION

For $1 \le k \le r \le T$,since by the base identity (72) and the fact that the difference of two probability measures annihilates the constant vector $T_k^\chi(\lambda_k)$ we have :

$$\boxed{\mathcal{R}_{k\to r} := \mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^+ - \tilde{\pi}_r^{\lambda_r})\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle.} \tag{77}$$

and then :

$$\boxed{\mathcal{R}_{k\to k} = \mathbb{E}\left\langle \lambda_k - \lambda^\star, (\widehat{P}_k^+ - \tilde{\pi}_k^{\lambda_k})\chi_k^{\lambda_k}\right\rangle.} \tag{78}$$

Thus the signed law-mismatch contribution in the dual-ascent proof is the base value of this recursion.

Proposition D.8 (Exact signed weak recursion)**.** *For every* $1 \le k \le r < T$*,*

$$\boxed{\mathcal{R}_{k\to r} = \mathsf{Sel}_{k\to r} + \mathsf{Rec}_{k\to r} + \mathsf{Guide}_{k\to r+1} + \mathsf{GD}_{k\to r+1} + \mathcal{R}_{k\to r+1}.} \tag{79}$$

where the four local contributions are

$$\mathsf{Sel}_{k\to r} := \mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^+ - \widehat{P}_r^-)\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle, \tag{80}$$

$$\begin{aligned}\mathsf{Rec}_{k\to r} := {}& \mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_r})\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle \\ & - \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_{r+1}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\rangle,\end{aligned} \tag{81}$$

$$\mathsf{Guide}_{k\to r+1} := \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \widehat{P}_{r+1}^+(\widehat{K}_{\theta,r+1}^{\lambda_{r+1}} - K_{r+1}^{\lambda_{r+1},\mathrm{G}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\rangle, \tag{82}$$

$$\mathsf{GD}_{k\to r+1} := \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \widehat{P}_{r+1}^+(K_{r+1}^{\lambda_{r+1},\mathrm{G}} - \tilde{K}_{r+1}^{\lambda_{r+1}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\rangle. \tag{83}$$

Proof. We recall that the base level k remains fixed throughout the recursion and $\tilde{\Phi}_{k\to r}^\lambda$ is the conditional view from level r of the same centered statistic $\chi_k^\lambda(\tilde{X}_k) - T_k^\chi(\lambda)$. At the base, (72) gives $\tilde{\Phi}_{k\to k}^\lambda = \chi_k^\lambda - T_k^\chi(\lambda)$. Under the exact tilted reverse dynamics,

$$\tilde{K}_{r+1}^\lambda \Phi_{k\to r}^\lambda = \Phi_{k\to r+1}^\lambda.$$

Subtracting the constant $T_k^\chi(\lambda)$ gives

$$\boxed{\tilde{K}_{r+1}^\lambda \tilde{\Phi}_{k\to r}^\lambda = \tilde{\Phi}_{k\to r+1}^\lambda.} \tag{84}$$

At x_{r+1} using by definition that for any bounded measurable f $\big(\widetilde{K}_{r+1}^\lambda f\big)(x_{r+1}) = \mathbb{E}_\lambda\Big[f(\widetilde{X}_r)\,\Big|\,\widetilde{X}_{r+1} = x_{r+1}\Big]$yields,

$$\begin{aligned}\tilde{K}_{r+1}^\lambda \tilde{\Phi}_{k\to r}^\lambda(x_{r+1}) &= \mathbb{E}_\lambda[\tilde{\Phi}_{k\to r}^\lambda(\tilde{X}_r) \mid \tilde{X}_{r+1} = x_{r+1}] \\ &= \mathbb{E}_\lambda\Big[\mathbb{E}_\lambda[\chi_k^\lambda(\tilde{X}_k) - T_k^\chi(\lambda) \mid \tilde{X}_r]\Big|\tilde{X}_{r+1} = x_{r+1}\Big] \\ &= \mathbb{E}_\lambda[\chi_k^\lambda(\tilde{X}_k) - T_k^\chi(\lambda) \mid \tilde{X}_{r+1} = x_{r+1}] \\ &= \tilde{\Phi}_{k\to r+1}^\lambda(x_{r+1}).\end{aligned}$$

To justify the passage from the second to the third line, remark that since $k \le r$ the Markov property gives

$$\mathbb{E}_\lambda\Big[\chi_k^\lambda(\tilde{X}_k) - T_k^\chi(\lambda) \mid \tilde{X}_r\Big] = \mathbb{E}_\lambda\Big[\chi_k^\lambda(\tilde{X}_k) - T_k^\chi(\lambda) \mid \tilde{X}_r, \tilde{X}_{r+1}\Big].$$

Therefore, by the tower property,

$$\begin{aligned}&\mathbb{E}_\lambda\Big[\mathbb{E}_\lambda\Big[\chi_k^\lambda(\tilde{X}_k) - T_k^\chi(\lambda) \mid \tilde{X}_r\Big]\Big|\tilde{X}_{r+1}\Big] \\ &\quad = \mathbb{E}_\lambda\Big[\mathbb{E}_\lambda\Big[\chi_k^\lambda(\tilde{X}_k) - T_k^\chi(\lambda) \mid \tilde{X}_r, \tilde{X}_{r+1}\Big]\Big|\tilde{X}_{r+1}\Big] \\ &\quad = \mathbb{E}_\lambda\Big[\chi_k^\lambda(\tilde{X}_k) - T_k^\chi(\lambda) \mid \tilde{X}_{r+1}\Big].\end{aligned}$$

First we separate the selection effect at level r. By definition,

$$\mathcal{R}_{k\to r} = \mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^+ - \tilde{\pi}_r^{\lambda_r})\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle. \tag{85}$$

The law $\widehat{P}_r^+$ is the post-selection law $\mathrm{Law}(\widehat{X}_r \mid \mathcal{H}_r)$, whereas $\widehat{P}_r^-$ is the pre-selection law $\mathrm{Law}(\widehat{X}_r \mid \mathcal{H}_{r+1})$. We insert this pre-selection law between $\widehat{P}_r^+$ and $\tilde{\pi}_r^{\lambda_r}$ through the identity

$$\widehat{P}_r^+ - \tilde{\pi}_r^{\lambda_r} = (\widehat{P}_r^+ - \widehat{P}_r^-) + (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_r}). \tag{86}$$

Substituting (86) into (85) gives

$$\mathcal{R}_{k\to r} = \underbrace{\mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^+ - \widehat{P}_r^-)\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle}_{\mathsf{Sel}_{k\to r}} + \mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_r})\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle. \tag{87}$$

The first line is exactly the selection contribution $\mathsf{Sel}_{k\to r}$ defined in (80). The second line is the mismatch that remains before selection.

We now insert the multiplier λ_{r+1} and isolate the re-centering term. By the sampler convention, the state transition $\widehat{X}_{r+1} \to \widehat{X}_r$ uses λ_{r+1}. The new multiplier λ_r is selected from the level-$(r+1)$ statistic and is used at the next reverse transition. Therefore the local sampler identity at level r reads

$$\widehat{P}_r^- = \widehat{P}_{r+1}^+ \widehat{K}_{\theta,r+1}^{\lambda_{r+1}}. \tag{88}$$

To use this identity, the remaining ideal comparison must also be written at the same multiplier λ_{r+1}. We therefore compare the current quantity

$$\mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_r})\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle \tag{89}$$

with the corresponding quantity evaluated entirely at λ_{r+1},

$$\mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_{r+1}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\rangle. \tag{90}$$

Their signed difference is, by definition, $\mathsf{Rec}_{k\to r}$ in (81). Equivalently, the identity defining $\mathsf{Rec}_{k\to r}$ can be written as

$$\begin{aligned}&\mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_r})\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle \\ &\quad = \mathsf{Rec}_{k\to r} + \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_{r+1}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\rangle.\end{aligned} \tag{91}$$

Thus λ_{r+1} is introduced because it is the multiplier used by the actual transition $r+1 \to r$, and $\mathsf{Rec}_{k\to r}$ records exactly the cost of replacing the selected λ_r by this previous multiplier.

Combining (87) and (91), we now have

$$\begin{aligned}\mathcal{R}_{k\to r} &= \mathsf{Sel}_{k\to r} + \mathsf{Rec}_{k\to r} \\ &\quad + \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, (\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_{r+1}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\rangle.\end{aligned} \tag{92}$$

At this stage, the only remaining task is to propagate the last mismatch from level r to level $r+1$.

Propagate the pre-selection mismatch from level r to level $r+1$. We use two one-step identities, both recalled here explicitly. The practical sampler identity recalled in (88), gives

$$\widehat{P}_r^- = \widehat{P}_{r+1}^+ \widehat{K}_{\theta,r+1}^{\lambda_{r+1}},$$

while the exact tilted reverse dynamics satisfy the Doob transport identity

$$\tilde{\pi}_r^{\lambda_{r+1}} = \tilde{\pi}_{r+1}^{\lambda_{r+1}} \tilde{K}_{r+1}^{\lambda_{r+1}}.$$

Substituting these two identities into the measure difference appearing in (92) yields the following measure–kernel–test compositions. Here $\mu K\varphi$ means: first apply the kernel to the test, $(K\varphi)(x) = \int \varphi(y)K(\mathrm{d}y \mid x)$, and then integrate the resulting function against μ. Thus, for example,

$$\widehat{P}_{r+1}^+ \widehat{K}_{\theta,r+1}^{\lambda_{r+1}} \tilde{\Phi}_{k\to r}^{\lambda_{r+1}} = \int \widehat{P}_{r+1}^+(\mathrm{d}x_{r+1}) \int \widehat{K}_{\theta,r+1}^{\lambda_{r+1}}(\mathrm{d}x_r \mid x_{r+1})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}(x_r).$$

With this interpretation,

$$(\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_{r+1}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}} = \widehat{P}_{r+1}^+ \widehat{K}_{\theta,r+1}^{\lambda_{r+1}} \tilde{\Phi}_{k\to r}^{\lambda_{r+1}} - \tilde{\pi}_{r+1}^{\lambda_{r+1}} \tilde{K}_{r+1}^{\lambda_{r+1}} \tilde{\Phi}_{k\to r}^{\lambda_{r+1}}. \tag{93}$$

Inserting the explicit intermediate quantity $\widehat{P}_{r+1}^+ \tilde{K}_{r+1}^{\lambda_{r+1}} \tilde{\Phi}_{k\to r}^{\lambda_{r+1}}$ gives

$$\begin{aligned}
&(\widehat{P}_r^- - \tilde{\pi}_r^{\lambda_{r+1}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}} \\
&= \underbrace{\widehat{P}_{r+1}^+ \Big(\widehat{K}_{\theta,r+1}^{\lambda_{r+1}} - \tilde{K}_{r+1}^{\lambda_{r+1}}\Big)\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}}_{\text{local kernel discrepancy on the practical level-}(r+1)\text{ law}} \\
&\quad + \underbrace{(\widehat{P}_{r+1}^+ - \tilde{\pi}_{r+1}^{\lambda_{r+1}})\tilde{K}_{r+1}^{\lambda_{r+1}} \tilde{\Phi}_{k\to r}^{\lambda_{r+1}}}_{\text{remaining law mismatch propagated to level } r+1} .
\end{aligned} \tag{94}$$

The first term in (94) is a one-step kernel error, while the second has the same practical-versus-ideal law mismatch structure as the original quantity but one level higher.

The propagated mismatch can now be identified as $\mathcal{R}_{k\to r+1}$. The continuation transport identity (84) states that, for the same fixed multiplier λ_{r+1},

$$\tilde{K}_{r+1}^{\lambda_{r+1}} \tilde{\Phi}_{k\to r}^{\lambda_{r+1}} = \tilde{\Phi}_{k\to r+1}^{\lambda_{r+1}}. \tag{95}$$

Therefore the second term in (94), after multiplication by $\lambda_{r+1} - \lambda^\star$ and outer expectation, becomes

$$\begin{aligned}
&\mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, (\widehat{P}_{r+1}^+ - \tilde{\pi}_{r+1}^{\lambda_{r+1}})\tilde{K}_{r+1}^{\lambda_{r+1}} \tilde{\Phi}_{k\to r}^{\lambda_{r+1}} \right\rangle \\
&= \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, (\widehat{P}_{r+1}^+ - \tilde{\pi}_{r+1}^{\lambda_{r+1}})\tilde{\Phi}_{k\to r+1}^{\lambda_{r+1}} \right\rangle \\
&= \mathcal{R}_{k\to r+1},
\end{aligned} \tag{96}$$

where the last equality is exactly the definition (77) with r replaced by $r+1$.

Finally split the local kernel discrepancy into guidance and Gaussian–Doob errors. The first term in (94) still compares the implemented kernel directly with the exact Doob kernel. To separate the two approximation mechanisms, we insert the exact-guidance Gaussian kernel $K_{r+1}^{\lambda_{r+1},\mathrm{G}}$ explicitly:

$$\begin{aligned}
&\widehat{K}_{\theta,r+1}^{\lambda_{r+1}} - \tilde{K}_{r+1}^{\lambda_{r+1}} \\
&= \Big(\widehat{K}_{\theta,r+1}^{\lambda_{r+1}} - K_{r+1}^{\lambda_{r+1},\mathrm{G}}\Big) + \Big(K_{r+1}^{\lambda_{r+1},\mathrm{G}} - \tilde{K}_{r+1}^{\lambda_{r+1}}\Big).
\end{aligned} \tag{97}$$

After left multiplication by $\widehat{P}_{r+1}^+$, application to $\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}$, multiplication by $\lambda_{r+1} - \lambda^\star$, and outer expectation, the first bracket is exactly $\mathsf{Guide}_{k\to r+1}$, while the second is exactly $\mathsf{GD}_{k\to r+1}$.

More explicitly, the two terms are

$$\begin{aligned}
\mathsf{Guide}_{k\to r+1} &:= \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \widehat{P}_{r+1}^+ (\widehat{K}_{\theta,r+1}^{\lambda_{r+1}} - K_{r+1}^{\lambda_{r+1},\mathrm{G}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}} \right\rangle, \\
\mathsf{GD}_{k\to r+1} &:= \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \widehat{P}_{r+1}^+ (K_{r+1}^{\lambda_{r+1},\mathrm{G}} - \tilde{K}_{r+1}^{\lambda_{r+1}})\tilde{\Phi}_{k\to r}^{\lambda_{r+1}} \right\rangle.
\end{aligned}$$

These are precisely the contributions defined in (82) and (83).

Substituting the propagated residual and the complete local kernel discrepancy into (92) gives

$$\boxed{\mathcal{R}_{k\to r} = \mathsf{Sel}_{k\to r} + \mathsf{Rec}_{k\to r} + \mathsf{Guide}_{k\to r+1} + \mathsf{GD}_{k\to r+1} + \mathcal{R}_{k\to r+1},}$$

which is exactly (79). Each of its four local contributions is controlled separately below. ■

D.4.3 Selection term $\mathsf{Sel}_{k\to r}$

Proposition D.9. *For $1 \le k < r < T$,*

$$\boxed{|\mathsf{Sel}_{k\to r}| \le 2M_{\mathrm{tot}}\kappa_{k\to r}\left[\eta G(G + R_\Lambda C_{\mathrm{der}}^\chi) + \eta^2 G^2 C_{\mathrm{der}}^\chi\right].} \tag{98}$$

where

$$M_{\mathrm{tot}} := M_{\mathrm{alg}} + M_{\mathrm{ref}}. \tag{99}$$

Proof. For $1 \le r < T$, define the selection operator

$$\mathsf{S}_r(\varphi) := (\widehat{P}_r^{+} - \widehat{P}_r^{-})\varphi. \tag{100}$$

Recall the multiplier increment from (44)

$$\boxed{\Delta_r := \lambda_r - \lambda_{r+1}.} \tag{101}$$

The multiplier λ_{r+1} is $\mathcal{H}_{r+1}$-measurable because it belongs to the multiplier history available before adjoining λ_r. The update at level $r+1$ selects λ_r using additional state information, so λ_r is generally only $\mathcal{H}_r$-measurable.

Let $\varphi^{\lambda_{r+1}}$ be a bounded test function evaluated at the previous multiplier λ_{r+1}, which is $\mathcal{H}_{r+1}$-measurable. Then

$$\boxed{\mathbb{E}[\mathsf{S}_r(\varphi^{\lambda_{r+1}}) \mid \mathcal{H}_{r+1}] = 0.} \tag{102}$$

Consequently,

$$\boxed{\mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \mathsf{S}_r(\varphi^{\lambda_{r+1}})\right\rangle = 0.} \tag{103}$$

The two conditional laws have been defined as

$$\widehat{P}_r^{+} = \mathrm{Law}(\widehat{X}_r \mid \mathcal{H}_r), \qquad \widehat{P}_r^{-} = \mathrm{Law}(\widehat{X}_r \mid \mathcal{H}_{r+1}).$$

Therefore, for any bounded test function whose parameter is already known at the previous-history stage,

$$\widehat{P}_r^{+}\varphi^{\lambda_{r+1}} = \mathbb{E}\Big[\varphi^{\lambda_{r+1}}(\widehat{X}_r) \mid \mathcal{H}_r\Big], \qquad \widehat{P}_r^{-}\varphi^{\lambda_{r+1}} = \mathbb{E}\Big[\varphi^{\lambda_{r+1}}(\widehat{X}_r) \mid \mathcal{H}_{r+1}\Big]. \tag{104}$$

This is the point at which predictability matters: λ_{r+1} is $\mathcal{H}_{r+1}$-measurable, so the map $x \mapsto \varphi^{\lambda_{r+1}}(x)$ is already fixed given $\mathcal{H}_{r+1}$. A test function evaluated at λ_r depends on the newly selected multiplier; its change is accounted for explicitly below.

We expand the conditional expectation of the selection operator using $\mathsf{S}_r = \widehat{P}_r^{+} - \widehat{P}_r^{-}$ and (104):

$$\mathbb{E}\big[\mathsf{S}_r(\varphi^{\lambda_{r+1}}) \mid \mathcal{H}_{r+1}\big] = \mathbb{E}\Big[\mathbb{E}[\varphi^{\lambda_{r+1}}(\widehat{X}_r) \mid \mathcal{H}_r]\Big|\mathcal{H}_{r+1}\Big] - \mathbb{E}[\varphi^{\lambda_{r+1}}(\widehat{X}_r) \mid \mathcal{H}_{r+1}]. \tag{105}$$

Because $\mathcal{H}_{r+1} \subset \mathcal{H}_r$, the tower property gives

$$\mathbb{E}\Big[\mathbb{E}[\varphi^{\lambda_{r+1}}(\widehat{X}_r) \mid \mathcal{H}_r]\Big|\mathcal{H}_{r+1}\Big] = \mathbb{E}[\varphi^{\lambda_{r+1}}(\widehat{X}_r) \mid \mathcal{H}_{r+1}].$$

The two terms in (105) are therefore identical and cancel, which proves

$$\mathbb{E}[\mathsf{S}_r(\varphi^{\lambda_{r+1}}) \mid \mathcal{H}_{r+1}] = 0.$$

For the second statement, $\lambda_{r+1} - \lambda^\star$ is also $\mathcal{H}_{r+1}$-measurable. Conditioning the signed contribution on $\mathcal{H}_{r+1}$ therefore allows this factor to be taken outside the inner conditional expectation:

$$\mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \mathsf{S}_r(\varphi^{\lambda_{r+1}})\right\rangle = \mathbb{E}\big[\big\langle \lambda_{r+1} - \lambda^\star, \mathbb{E}[\mathsf{S}_r(\varphi^{\lambda_{r+1}}) \mid \mathcal{H}_{r+1}]\big\rangle\big] = 0.$$

We now apply this cancellation to the centered continuation introduced in (70). For each fixed multiplier, $\tilde{\Phi}_{k\to r}^{\lambda} = \Phi_{k\to r}^{\lambda} - T_k^{\chi}(\lambda)$. The mean $T_k^{\chi}(\lambda)$ is constant in the state variable, so the selection operator annihilates it. Keeping the centered version in the calculation allows us to apply the forgetting bounds directly.

Write

$$\Delta_r = \lambda_r - \lambda_{r+1}.$$

and define

$$\delta\tilde{\Phi}_{k\to r} := \tilde{\Phi}_{k\to r}^{\lambda_r} - \tilde{\Phi}_{k\to r}^{\lambda_{r+1}}.$$

Expanding both the multiplier and the centered continuation gives the exact four-term decomposition

$$\mathsf{Sel}_{k\to r} := \mathbb{E}\left\langle \lambda_r - \lambda^\star, (\widehat{P}_r^+ - \widehat{P}_r^-)\tilde{\Phi}_{k\to r}^{\lambda_r}\right\rangle = \underbrace{\mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \mathsf{S}_r(\tilde{\Phi}_{k\to r}^{\lambda_{r+1}})\right\rangle}_{T_1}$$
$$+ \underbrace{\mathbb{E}\left\langle \Delta_r, \mathsf{S}_r(\tilde{\Phi}_{k\to r}^{\lambda_{r+1}})\right\rangle}_{T_2}$$
$$+ \underbrace{\mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \mathsf{S}_r(\delta\tilde{\Phi}_{k\to r})\right\rangle}_{T_3}$$
$$+ \underbrace{\mathbb{E}\left\langle \Delta_r, \mathsf{S}_r(\delta\tilde{\Phi}_{k\to r})\right\rangle}_{T_4}. \tag{106}$$

For T_1, the test is exactly $\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}$, whose multiplier λ_{r+1} is $\mathcal{H}_{r+1}$-measurable. Equation (103) therefore reads

$$T_1 = \mathbb{E}\left\langle \lambda_{r+1} - \lambda^\star, \mathsf{S}_r(\tilde{\Phi}_{k\to r}^{\lambda_{r+1}})\right\rangle = 0.$$

For T_2, recall first what the selection operator means:

$$\mathsf{S}_r(\psi) = \widehat{P}_r^+\psi - \widehat{P}_r^-\psi.$$

The triangle inequality for vector-valued integrals, $\|(P-Q)\psi\| \le P\,\|\psi\| + Q\,\|\psi\|$, therefore gives pathwise

$$\left\|\mathsf{S}_r(\tilde{\Phi}_{k\to r}^{\lambda_{r+1}})\right\| = \left\|\widehat{P}_r^+\tilde{\Phi}_{k\to r}^{\lambda_{r+1}} - \widehat{P}_r^-\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\|$$
$$\le \widehat{P}_r^+\left\|\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\| + \widehat{P}_r^-\left\|\tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\|. \tag{107}$$

Now use the first forgetting estimate (74), which states pointwise that

$$\left\|\tilde{\Phi}_{k\to r}^{\lambda}(x)\right\| \le G\kappa_{k\to r}(\|x\| + M_{\mathrm{ref}}).$$

Applying this bound inside each of the two probability integrals in (107) yields

$$\left\|\mathsf{S}_r(\tilde{\Phi}_{k\to r}^{\lambda_{r+1}})\right\| \le G\kappa_{k\to r}\left[\widehat{P}_r^+(\|x\| + M_{\mathrm{ref}}) + \widehat{P}_r^-(\|x\| + M_{\mathrm{ref}})\right]. \tag{108}$$

Taking the outer expectation gives

$$\mathbb{E}\left\|\mathsf{S}_r(\tilde{\Phi}_{k\to r}^{\lambda_{r+1}})\right\| \le G\kappa_{k\to r}\,\mathbb{E}\left[\widehat{P}_r^+(\|x\| + M_{\mathrm{ref}}) + \widehat{P}_r^-(\|x\| + M_{\mathrm{ref}})\right]. \tag{109}$$

The terms $\widehat{P}_r^+\,\|x\|$ and $\widehat{P}_r^-\,\|x\|$ are conditional first moments

$$\widehat{P}_r^+\,\|x\| = \mathbb{E}[\left\|\widehat{X}_r\right\| \mid \mathcal{H}_r], \qquad \widehat{P}_r^-\,\|x\| = \mathbb{E}[\left\|\widehat{X}_r\right\| \mid \mathcal{H}_{r+1}].$$

After taking the outer expectation, the tower property applied to each conditional law gives

$$\mathbb{E}[\widehat{P}_r^+\,\|x\|] = \mathbb{E}[\widehat{P}_r^-\,\|x\|] = \mathbb{E}\left\|\widehat{X}_r\right\|.$$

Using the moment bound of Lemma D.3, $\sup_r \mathbb{E}\left\|\widehat{X}_r\right\| \le M_{\mathrm{alg}}$, and the abbreviation $M_{\mathrm{tot}} = M_{\mathrm{alg}} + M_{\mathrm{ref}}$, we obtain

$$\boxed{\mathbb{E}\left\|\mathsf{S}_r(\tilde{\Phi}_{k\to r}^{\lambda_{r+1}})\right\| \le 2M_{\mathrm{tot}}G\kappa_{k\to r}.} \tag{110}$$

Together with the movement lemma (45), $\|\Delta_r\| \le \eta G$,

$$|T_2| \le 2M_{\mathrm{tot}}\eta G^2\kappa_{k\to r}. \tag{111}$$

For T_3 and T_4, recall the multiplier-Lipschitz forgetting estimate (75):

$$\left\|\tilde{\Phi}^{\lambda}_{k\to r}(x) - \tilde{\Phi}^{\lambda'}_{k\to r}(x)\right\| \le C^{\chi}_{\mathrm{der}}\kappa_{k\to r}(\|x\| + M_{\mathrm{ref}})\,\|\lambda - \lambda'\|\,.$$

With $\lambda = \lambda_r$, $\lambda' = \lambda_{r+1}$ and $\|\Delta_r\| \le \eta G$,

$$\left\|\delta\tilde{\Phi}_{k\to r}(x)\right\| \le C^{\chi}_{\mathrm{der}}\kappa_{k\to r}(\|x\| + M_{\mathrm{ref}})\eta G. \tag{112}$$

We again expand the action of the selection operator before bounding it. From $\mathsf{S}_r(\delta\tilde{\Phi}_{k\to r}) = \widehat{P}^{+}_r\delta\tilde{\Phi}_{k\to r} - \widehat{P}^{-}_r\delta\tilde{\Phi}_{k\to r}$ and the same inequality $\|(P-Q)\psi\| \le P\,\|\psi\| + Q\,\|\psi\|$,

$$\left\|\mathsf{S}_r(\delta\tilde{\Phi}_{k\to r})\right\| \le \widehat{P}^{+}_r\left\|\delta\tilde{\Phi}_{k\to r}\right\| + \widehat{P}^{-}_r\left\|\delta\tilde{\Phi}_{k\to r}\right\|.$$

The pointwise estimate (112) then gives

$$\left\|\mathsf{S}_r(\delta\tilde{\Phi}_{k\to r})\right\| \le C^{\chi}_{\mathrm{der}}\kappa_{k\to r}\eta G\left[\widehat{P}^{+}_r(\|x\| + M_{\mathrm{ref}}) + \widehat{P}^{-}_r(\|x\| + M_{\mathrm{ref}})\right].$$

Taking outer expectation and using $\mathbb{E}[\widehat{P}^{\pm}_r\,\|x\|] = \mathbb{E}\left\|\widehat{X}_r\right\| \le M_{\mathrm{alg}}$ exactly as above yields

$$\boxed{\mathbb{E}\left\|\mathsf{S}_r(\delta\tilde{\Phi}_{k\to r})\right\| \le 2M_{\mathrm{tot}}C^{\chi}_{\mathrm{der}}\kappa_{k\to r}\eta G.} \tag{113}$$

Therefore, since $\|\lambda_{r+1} - \lambda^\star\| \le R_\Lambda$,

$$|T_3| \le 2M_{\mathrm{tot}}R_\Lambda C^{\chi}_{\mathrm{der}}\kappa_{k\to r}\eta G, \tag{114}$$

and using one more factor $\|\Delta_r\| \le \eta G$,

$$|T_4| \le 2M_{\mathrm{tot}}C^{\chi}_{\mathrm{der}}\kappa_{k\to r}\eta^2 G^2. \tag{115}$$

Since $T_1 = 0$ and (111)–(115) give

$$|T_2| \le 2M_{\mathrm{tot}}\eta G^2\kappa_{k\to r}, \qquad |T_3| \le 2M_{\mathrm{tot}}R_\Lambda C^{\chi}_{\mathrm{der}}\eta G\kappa_{k\to r},$$
$$|T_4| \le 2M_{\mathrm{tot}}C^{\chi}_{\mathrm{der}}\eta^2 G^2\kappa_{k\to r},$$

their sum is exactly the right-hand side of (98). ■

D.4.4 RE-CENTERING TERM $\mathsf{Rec}_{k\to r}$

Proposition D.10. *For $1 \le k < r < T$,*

$$\boxed{|\mathsf{Rec}_{k\to r}| \le M_{\mathrm{tot}}\kappa_{k\to r}\eta G(G + R_\Lambda C^{\chi}_{\mathrm{der}}).} \tag{116}$$

Proof. For $1 \le k \le r < T$ and a deterministic λ, define the random vector-valued function (random through $\widehat{P}^{-}_r$)

$$Q_{k\to r}(\lambda) := (\widehat{P}^{-}_r - \tilde{\pi}^{\lambda}_r)\tilde{\Phi}^{\lambda}_{k\to r}. \tag{117}$$

The purpose of this term is to align the ideal comparison with the multiplier λ_{r+1} that actually generated the practical transition from level $r+1$ to r. Because the continuation is centered by definition, (73) gives $\tilde{\pi}^{\lambda}_r\tilde{\Phi}^{\lambda}_{k\to r} = 0$. Consequently,

$$\boxed{Q_{k\to r}(\lambda) = \widehat{P}^{-}_r\tilde{\Phi}^{\lambda}_{k\to r}.} \tag{118}$$

Since $\Delta_r = \lambda_r - \lambda_{r+1}$,

$$\begin{aligned}&\langle\lambda_r - \lambda^\star, Q_{k\to r}(\lambda_r)\rangle - \langle\lambda_{r+1} - \lambda^\star, Q_{k\to r}(\lambda_{r+1})\rangle\\ &\quad= \langle\Delta_r, Q_{k\to r}(\lambda_r)\rangle + \langle\lambda_{r+1} - \lambda^\star, Q_{k\to r}(\lambda_r) - Q_{k\to r}(\lambda_{r+1})\rangle\,.\end{aligned} \tag{119}$$

By definition (81), the outer expectation of the left-hand side is $\mathsf{Rec}_{k\to r}$.

For the first term, we make the measure action explicit. Equation (118) states

$$Q_{k\to r}(\lambda_r) = \widehat{P}_r^{-}\tilde{\Phi}_{k\to r}^{\lambda_r}.$$

Here the right-hand side is the pathwise integral of the centered continuation against the pre-selection conditional law. Using directly the triangle inequality $\|P\psi\| \le P\,\|\psi\|$,

$$\begin{aligned}\|Q_{k\to r}(\lambda_r)\| &= \left\|\widehat{P}_r^{-}\tilde{\Phi}_{k\to r}^{\lambda_r}\right\| \\ &\le \widehat{P}_r^{-}\left\|\tilde{\Phi}_{k\to r}^{\lambda_r}\right\|. \end{aligned} \tag{120}$$

The forgetting estimate (74) states pointwise that

$$\left\|\tilde{\Phi}_{k\to r}^{\lambda}(x)\right\| \le G\kappa_{k\to r}(\|x\| + M_{\mathrm{ref}}).$$

Substituting this pointwise bound inside the $\widehat{P}_r^{-}$ integral gives

$$\begin{aligned}\|Q_{k\to r}(\lambda_r)\| &\le G\kappa_{k\to r}\widehat{P}_r^{-}(\|x\| + M_{\mathrm{ref}}) \\ &= G\kappa_{k\to r}\left(\widehat{P}_r^{-}\,\|x\| + M_{\mathrm{ref}}\right). \end{aligned} \tag{121}$$

Since

$$\widehat{P}_r^{-}\,\|x\| = \mathbb{E}[\left\|\widehat{X}_r\right\| \mid \mathcal{H}_{r+1}],$$

taking outer expectation and using the tower property yields

$$\begin{aligned}\mathbb{E}\,\|Q_{k\to r}(\lambda_r)\| &\le G\kappa_{k\to r}\left(\mathbb{E}\left\|\widehat{X}_r\right\| + M_{\mathrm{ref}}\right) \\ &\le G\kappa_{k\to r}(M_{\mathrm{alg}} + M_{\mathrm{ref}}).\end{aligned}$$

With $M_{\mathrm{tot}} := M_{\mathrm{alg}} + M_{\mathrm{ref}}$,

$$\boxed{\mathbb{E}\,\|Q_{k\to r}(\lambda_r)\| \le M_{\mathrm{tot}}G\kappa_{k\to r}.} \tag{122}$$

With the movement bound (45), $\|\Delta_r\| \le \eta G$,

$$|\mathbb{E}\,\langle\Delta_r, Q_{k\to r}(\lambda_r)\rangle| \le M_{\mathrm{tot}}\eta G^2\kappa_{k\to r}. \tag{123}$$

For the second term, equation (118) gives the exact difference

$$Q_{k\to r}(\lambda_r) - Q_{k\to r}(\lambda_{r+1}) = \widehat{P}_r^{-}\left(\tilde{\Phi}_{k\to r}^{\lambda_r} - \tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right). \tag{124}$$

Applying again the elementary inequality $\|P\psi\| \le P\,\|\psi\|$ gives

$$\|Q_{k\to r}(\lambda_r) - Q_{k\to r}(\lambda_{r+1})\| \le \widehat{P}_r^{-}\left\|\tilde{\Phi}_{k\to r}^{\lambda_r} - \tilde{\Phi}_{k\to r}^{\lambda_{r+1}}\right\|.$$

The multiplier-Lipschitz forgetting estimate (75) states that, for every x,

$$\left\|\tilde{\Phi}_{k\to r}^{\lambda_r}(x) - \tilde{\Phi}_{k\to r}^{\lambda_{r+1}}(x)\right\| \le C_{\mathrm{der}}^{\chi}\kappa_{k\to r}(\|x\| + M_{\mathrm{ref}})\,\|\Delta_r\|.$$

Therefore

$$\|Q_{k\to r}(\lambda_r) - Q_{k\to r}(\lambda_{r+1})\| \le C_{\mathrm{der}}^{\chi}\kappa_{k\to r}\,\|\Delta_r\|\left(\widehat{P}_r^{-}\,\|x\| + M_{\mathrm{ref}}\right).$$

Taking outer expectation, using $\mathbb{E}[\widehat{P}_r^{-}\,\|x\|] = \mathbb{E}\left\|\widehat{X}_r\right\| \le M_{\mathrm{alg}}$, and then the movement bound (45), $\|\Delta_r\| \le \eta G$, gives

$$\boxed{\mathbb{E}\,\|Q_{k\to r}(\lambda_r) - Q_{k\to r}(\lambda_{r+1})\| \le M_{\mathrm{tot}}C_{\mathrm{der}}^{\chi}\kappa_{k\to r}\eta G.} \tag{125}$$

Since $\|\lambda_{r+1} - \lambda^\star\| \le R_\Lambda$,

$$|\mathbb{E}\,\langle\lambda_{r+1} - \lambda^\star, Q_{k\to r}(\lambda_r) - Q_{k\to r}(\lambda_{r+1})\rangle| \le M_{\mathrm{tot}}R_\Lambda C_{\mathrm{der}}^{\chi}\kappa_{k\to r}\eta G. \tag{126}$$

The two terms in (119) are therefore bounded by

$$M_{\mathrm{tot}}\eta G^2\kappa_{k\to r} \quad \text{and} \quad M_{\mathrm{tot}}R_\Lambda C_{\mathrm{der}}^{\chi}\eta G\kappa_{k\to r},$$

respectively. Their sum is

$$\boxed{M_{\mathrm{tot}}\kappa_{k\to r}\eta G(G + R_\Lambda C_{\mathrm{der}}^{\chi}),}$$

■

D.4.5 DIAGONAL CONTRIBUTION $\mathsf{Sel}_{k\to k} + \mathsf{Rec}_{k\to k}$

Proposition D.11. *For each $k = 1, \dots, T-1$, the diagonal contribution satisfies $|\mathsf{Sel}_{k\to k}| + |\mathsf{Rec}_{k\to k}| \le R_{\mathrm{base}}(\eta)$, where*

$$\boxed{R_{\mathrm{base}}(\eta) := \eta G(6G + 3R_\Lambda C^\chi_{\mathrm{diag}}) + 2C^\chi_{\mathrm{diag}}\eta^2 G^2.} \tag{127}$$

The diagonal regularity constant is

$$C^\chi_{\mathrm{diag}} := 2L_\chi + G^2. \tag{128}$$

Proof. At $r = k$, the coefficient $\kappa_{k\to r}$ is not defined, since the forgetting estimates apply only when $r > k$. We therefore control the unique diagonal contribution associated with each base index k directly through uniform bounds on the centered continuation and its multiplier derivative.

Diagonal regularity bounds. At the base level, At the base,

$$\tilde\Phi^\lambda_{k\to k} = \chi^\lambda_k - T^\chi_k(\lambda). \tag{129}$$

Since both terms have norm at most G,

$$\left\|\tilde\Phi^\lambda_{k\to k}\right\|_\infty \le 2G. \tag{130}$$

Let J_λ denote the Jacobian with respect to λ. By Lemma D.24,

$$J_\lambda T^\chi_k(\lambda) = \mathbb{E}_\lambda[J_\lambda \chi^\lambda_k(\tilde X_k)] - \mathrm{Cov}_\lambda(\chi^\lambda_k(\tilde X_k), g(\tilde X_0)). \tag{131}$$

For any unit $u, v \in \mathbb{R}^m$,

$$\begin{aligned}
&\left|u^\top \mathrm{Cov}_\lambda(\chi^\lambda_k(\tilde X_k), g(\tilde X_0))v\right| \\
&\quad = \left|\mathrm{Cov}_\lambda(u^\top\chi^\lambda_k(\tilde X_k), v^\top g(\tilde X_0))\right| \\
&\quad \le \sqrt{\mathrm{Var}_\lambda(u^\top\chi^\lambda_k(\tilde X_k))}\sqrt{\mathrm{Var}_\lambda(v^\top g(\tilde X_0))} \le G^2.
\end{aligned}$$

The last inequality uses the fact that both scalar variables have absolute value at most G. Taking the supremum over unit u, v gives a covariance operator norm at most G^2. Also, the norm inequality for the matrix expectation gives

$$\left\|\mathbb{E}_\lambda[J_\lambda\chi^\lambda_k(\tilde X_k)]\right\|_{\mathrm{op}} \le \mathbb{E}_\lambda\left\|J_\lambda\chi^\lambda_k(\tilde X_k)\right\|_{\mathrm{op}} \le L_\chi.$$

The triangle inequality in (131) yields

$$\|J_\lambda T^\chi_k(\lambda)\|_{\mathrm{op}} \le L_\chi + G^2, \tag{132}$$

The derivative of the centered base statistic is $J_\lambda\tilde\Phi^\lambda_{k\to k}(x) = J_\lambda\chi^\lambda_k(x) - J_\lambda T^\chi_k(\lambda)$. Its two terms are bounded by L_χ and $L_\chi + G^2$, respectively, so

$$\boxed{\sup_x \left\|J_\lambda\tilde\Phi^\lambda_{k\to k}(x)\right\|_{\mathrm{op}} \le C^\chi_{\mathrm{diag}}, \qquad C^\chi_{\mathrm{diag}} = 2L_\chi + G^2.} \tag{133}$$

At $r = k$, we use the uniform bounds (130) and (133).

Control of $\mathsf{Sel}_{k\to k}$. Here, we reuse the decomposition (106) with $r = k$. The definitions of $T_1, \dots, T_4$ remain unchanged; only the bounds are adapted to the base level, where no forgetting estimate is available.

The predictable term T_1 is still zero by (103). For T_2, the selection operator is $\mathsf{S}_k(\psi) = \widehat P^+_k\psi - \widehat P^-_k\psi$. Hence using $\|(P-Q)\psi\| \le P\|\psi\| + Q\|\psi\|$ gives

$$\begin{aligned}
\left\|\mathsf{S}_k(\tilde\Phi^{\lambda_{k+1}}_{k\to k})\right\| &\le \widehat P^+_k\left\|\tilde\Phi^{\lambda_{k+1}}_{k\to k}\right\| + \widehat P^-_k\left\|\tilde\Phi^{\lambda_{k+1}}_{k\to k}\right\| \\
&\le 2G + 2G = 4G,
\end{aligned}$$

where the last line uses $\left\|\tilde{\Phi}^{\lambda}_{k\to k}\right\|_{\infty} \le 2G$. Since $\|\Delta_k\| \le \eta G$

$$|T_2| \le \mathbb{E}[\|\Delta_k\| \left\|\mathsf{S}_k(\tilde{\Phi}^{\lambda_{k+1}}_{k\to k})\right\|] \le 4\eta G^2.$$

For the change in the centered continuation, the Jacobian bound (133) and the fundamental theorem of calculus along the segment between λ_{k+1} and λ_k give

$$\left\|\tilde{\Phi}^{\lambda_k}_{k\to k}(x) - \tilde{\Phi}^{\lambda_{k+1}}_{k\to k}(x)\right\| \le C^{\chi}_{\mathrm{diag}} \|\Delta_k\| \le C^{\chi}_{\mathrm{diag}}\eta G.$$

Applying $\|(P-Q)\psi\| \le P\|\psi\| + Q\|\psi\|$ once more,

$$\left\|\mathsf{S}_k(\tilde{\Phi}^{\lambda_k}_{k\to k} - \tilde{\Phi}^{\lambda_{k+1}}_{k\to k})\right\| \le 2C^{\chi}_{\mathrm{diag}}\eta G.$$

Therefore

$$|T_3| \le R_\Lambda\, 2C^{\chi}_{\mathrm{diag}}\eta G, \qquad |T_4| \le \eta G\, 2C^{\chi}_{\mathrm{diag}}\eta G = 2C^{\chi}_{\mathrm{diag}}\eta^2 G^2.$$

Control of $\mathsf{Rec}_{k\to k}$ For the re-centering term, equation (118) gives $Q_{k\to k}(\lambda) = \widehat{P}^{-}_k \tilde{\Phi}^{\lambda}_{k\to k}$. The inequality $\|P\psi\| \le P\|\psi\|$ and $\left\|\tilde{\Phi}^{\lambda}_{k\to k}\right\|_{\infty} \le 2G$ imply

$$\|Q_{k\to k}(\lambda)\| \le \widehat{P}^{-}_k \left\|\tilde{\Phi}^{\lambda}_{k\to k}\right\| \le 2G.$$

Thus the movement of the multiplier prefactor contributes at most

$$\mathbb{E}\,|\langle \Delta_k, Q_{k\to k}(\lambda_k)\rangle| \le (\eta G)(2G) = 2\eta G^2.$$

Likewise,

$$\begin{aligned}\|Q_{k\to k}(\lambda_k) - Q_{k\to k}(\lambda_{k+1})\| &\le \widehat{P}^{-}_k \left\|\tilde{\Phi}^{\lambda_k}_{k\to k} - \tilde{\Phi}^{\lambda_{k+1}}_{k\to k}\right\| \\ &\le C^{\chi}_{\mathrm{diag}}\eta G,\end{aligned}$$

so the second re-centering contribution is at most $R_\Lambda C^{\chi}_{\mathrm{diag}}\eta G$. Summing the four selection contributions and the two re-centering contributions yields (127).

■

D.4.6 Practical-guidance term $\mathsf{Guide}_{k\to q}$

In the local kernel estimates, q denotes the transition level: the transition is $q \to q-1$ and acts on a continuation observed at level $q-1$. Thus $q = r+1$ in the recursion above. The base level k is unchanged.

Proposition D.12 (Practical-guidance bounds)**.** *Under Ass. 4.3, set*

$$C_q := \frac{\beta_q}{\sqrt{\alpha_q}} I_d, \qquad H_q := C_q^\top \Sigma_q^{-1} C_q, \qquad \|v\|_{H_q} := \sqrt{v^\top H_q v}, \qquad 2 \le q \le T.$$

For the adjacent step $q = k+1 \le T$,

$$\boxed{\begin{aligned}|\mathsf{Guide}_{k\to k+1}| &\le R_\Lambda G\, \mathbb{E}\left[\widehat{P}^{+}_{k+1}\left\|\Delta^{\lambda_{k+1}}_{h,k+1}\right\|_{H_{k+1}}\right] \\ &\le R_\Lambda G \sqrt{\|H_{k+1}\|_{\mathrm{op}}}\,\varepsilon_h.\end{aligned}} \tag{134}$$

For every non-adjacent step $k+2 \le q \le T$,

$$\boxed{\begin{aligned}|\mathsf{Guide}_{k\to q}| &\le R_\Lambda G \kappa_{k\to q-1}\, \mathbb{E}\left[\widehat{P}^{+}_{q}\left\|C_q \Delta^{\lambda_q}_{h,q}\right\|\right] \\ &\le R_\Lambda G \kappa_{k\to q-1} \|C_q\|_{\mathrm{op}}\, \varepsilon_h.\end{aligned}} \tag{135}$$

All conversion coefficients are finite for the prescribed schedule and covariances.

Proof. Recall that, for $q > k$, the practical-guidance contribution is

$$\mathsf{Guide}_{k\to q} = \mathbb{E}\left[\left\langle \lambda_q - \lambda^\star,\, \widehat{P}^{+}_q\big(\widehat{K}^{\lambda_q}_{\theta,q} - K^{\lambda_q,\mathrm{G}}_q\big)\tilde{\Phi}^{\lambda_q}_{k\to q-1}\right\rangle\right].$$

Pointwise comparison of the Gaussian kernels. We first compare the two kernels at a fixed multiplier λ and source state x, before integrating these comparisons under the implemented conditional law.

Recall that $m_{\theta,q}$ is the exact reference mean as $s_\theta(\cdot, q) = \nabla \log \tilde{p}_q$. The prescribed DDPM mean and guidance coefficient are

$$m_{\theta,q}(x) = \frac{x + \beta_q s_\theta(x, q)}{\sqrt{\alpha_q}}, \qquad C_q = \frac{\beta_q}{\sqrt{\alpha_q}} I_d.$$

The two Gaussian kernels under comparison share the prescribed covariance Σ_q and satisfy

$$\begin{aligned} \widehat{K}^{\lambda}_{\theta,q}(\cdot \mid x) &= \mathcal{N}\big(m_{\theta,q}(x) + C_q s^{\lambda}_t(x), \Sigma_q\big)\,, \\ K^{\lambda,\mathrm{G}}_q(\cdot \mid x) &= \mathcal{N}\Big(m_{\theta,q}(x) + C_q \nabla_x \log \tilde{h}^{\lambda}_q(x), \Sigma_q\Big)\,. \end{aligned} \tag{136}$$

Their difference of means is consequently $C_q \Delta^{\lambda}_{h,q}(x)$, where

$$\Delta^{\lambda}_{h,q}(x) = s^{\lambda}_t(x) - \nabla_x \log \tilde{h}^{\lambda}_q(x).$$

We use the convention $\mathrm{TV}(\mu, \nu) := \sup_A |\mu(A) - \nu(A)|$. For $q \geq 2$, the covariance $\Sigma_q = \beta_q I_d$ is positive definite by the prescribed schedule. The common-covariance Gaussian formula,therefore gives, pointwise in x,

$$\mathrm{KL}\Big(\widehat{K}^{\lambda}_{\theta,q}(\cdot \mid x) \,\Big\|\, K^{\lambda,\mathrm{G}}_q(\cdot \mid x)\Big) = \frac{1}{2}\big(C_q \Delta^{\lambda}_{h,q}(x)\big)^{\top} \Sigma_q^{-1} \big(C_q \Delta^{\lambda}_{h,q}(x)\big) = \frac{1}{2} \left\|\Delta^{\lambda}_{h,q}(x)\right\|^2_{H_q}.$$

Pinsker's inequality yields

$$\mathrm{TV}\Big(\widehat{K}^{\lambda}_{\theta,q}(\cdot \mid x), K^{\lambda,\mathrm{G}}_q(\cdot \mid x)\Big) \leq \frac{1}{2} \left\|\Delta^{\lambda}_{h,q}(x)\right\|_{H_q}. \tag{137}$$

The guidance assumption 4.3 further implies

$$\left\|\Delta^{\lambda}_{h,q}(x)\right\|_{H_q} \leq \sqrt{\|H_q\|_{\mathrm{op}}} \left\|\Delta^{\lambda}_{h,q}(x)\right\| \leq \sqrt{\|H_q\|_{\mathrm{op}}}\, \varepsilon_h.$$

For the Wasserstein comparison, recall that

$$W_1(\mu, \nu) := \inf_{\zeta \in \Gamma(\mu,\nu)} \int \|u - v\|\ \zeta(\mathrm{d}u, \mathrm{d}v),$$

where $\Gamma(\mu, \nu)$ denotes the set of couplings.

The two Gaussian laws in (136) differ only by a translation. Using the standard fact that the W_1 distance between two translations of the same Gaussian law equals the distance between their means we obtain

$$\boxed{W_1\Big(\widehat{K}^{\lambda}_{\theta,q}(\cdot \mid x), K^{\lambda,\mathrm{G}}_q(\cdot \mid x)\Big) = \left\|C_q \Delta^{\lambda}_{h,q}(x)\right\|.} \tag{138}$$

Assumption 4.3 gives,

$$\left\|C_q \Delta^{\lambda}_{h,q}(x)\right\| \leq \|C_q\|_{\mathrm{op}} \left\|\Delta^{\lambda}_{h,q}(x)\right\| \leq \|C_q\|_{\mathrm{op}}\ \varepsilon_h.$$

These comparisons hold for every source state and multiplier. Below, distances between kernels denote the corresponding functions of the source state; for example,

$$\mathrm{TV}(K, K')(x) := \mathrm{TV}(K(\cdot \mid x), K'(\cdot \mid x))\,.$$

The same convention applies to W_1 and kernel KL divergences.

Adjacent transition: total-variation control. We consider first step $q = k+1$. For a vector-valued test function satisfying $\sup_x \|\varphi(x)\| \le G$, the total-variation bound is

$$\boxed{\|\mu\varphi - \nu\varphi\| \le 2G\,\mathrm{TV}(\mu,\nu).} \tag{139}$$

At the adjacent step, the centered continuation is

$$\tilde{\Phi}^{\lambda_{k+1}}_{k\to k} = \chi^{\lambda_{k+1}}_k - T^{\chi}_k(\lambda_{k+1}).$$

The centering constant does not affect a difference of Markov kernels. Indeed, both kernels have total mass one, and therefore

$$\big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1} - K^{\lambda_{k+1},\mathrm{G}}_{k+1}\big)\tilde{\Phi}^{\lambda_{k+1}}_{k\to k} = \big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1} - K^{\lambda_{k+1},\mathrm{G}}_{k+1}\big)\chi^{\lambda_{k+1}}_k. \tag{140}$$

The transition $k+1 \to k$ can consequently be tested directly against the base statistic, which satisfies $\sup_x \left\|\chi^{\lambda_{k+1}}_k(x)\right\| \le G$. Applying (139) pointwise gives

$$\left\|\big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1} - K^{\lambda_{k+1},\mathrm{G}}_{k+1}\big)\chi^{\lambda_{k+1}}_k\right\| \le 2G\,\mathrm{TV}\Big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1}, K^{\lambda_{k+1},\mathrm{G}}_{k+1}\Big).$$

By (140), the centering constant cancels in the difference of the two kernels. Hence using $|\,\mathbb{E}\,Z| \le \mathbb{E}\,|Z|$ gives,

$$\begin{aligned}
|\mathsf{Guide}_{k\to k+1}| &= \left|\mathbb{E}\Big[\Big\langle \lambda_{k+1} - \lambda^\star,\, \widehat{P}^+_{k+1}\big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1} - K^{\lambda_{k+1},\mathrm{G}}_{k+1}\big)\chi^{\lambda_{k+1}}_k\Big\rangle\Big]\right| \\
&\le \mathbb{E}\Big[\Big|\Big\langle \lambda_{k+1} - \lambda^\star,\, \widehat{P}^+_{k+1}\big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1} - K^{\lambda_{k+1},\mathrm{G}}_{k+1}\big)\chi^{\lambda_{k+1}}_k\Big\rangle\Big|\Big].
\end{aligned}$$

By Cauchy–Schwarz and $\|\lambda_{k+1} - \lambda^\star\| \le R_\Lambda$,

$$|\mathsf{Guide}_{k\to k+1}| \le R_\Lambda\ \mathbb{E}\Big[\Big\|\widehat{P}^+_{k+1}\big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1} - K^{\lambda_{k+1},\mathrm{G}}_{k+1}\big)\chi^{\lambda_{k+1}}_k\Big\|\Big].$$

The inequality $\left\|\widehat{P}^+_{k+1}\psi\right\| \le \widehat{P}^+_{k+1}\,\|\psi\|$ holds for each realization of the conditional measure. Applying it before the outer expectation, followed by the pointwise TV bound, gives

$$\begin{aligned}
|\mathsf{Guide}_{k\to k+1}| &\le R_\Lambda\ \mathbb{E}\Big[\widehat{P}^+_{k+1}\Big\|\big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1} - K^{\lambda_{k+1},\mathrm{G}}_{k+1}\big)\chi^{\lambda_{k+1}}_k\Big\|\Big] \\
&\le 2R_\Lambda G\ \mathbb{E}\Big[\widehat{P}^+_{k+1}\mathrm{TV}\Big(\widehat{K}^{\lambda_{k+1}}_{\theta,k+1}, K^{\lambda_{k+1},\mathrm{G}}_{k+1}\Big)\Big].
\end{aligned}$$

Finally, (137) and 4.3 give

$$\begin{aligned}
|\mathsf{Guide}_{k\to k+1}| &\le R_\Lambda G\ \mathbb{E}\left[\widehat{P}^+_{k+1}\left\|\Delta^{\lambda_{k+1}}_{h,k+1}\right\|_{H_{k+1}}\right] \\
&\le R_\Lambda G\sqrt{\|H_{k+1}\|_{\mathrm{op}}}\ \mathbb{E}\Big[\widehat{P}^+_{k+1}\left\|\Delta^{\lambda_{k+1}}_{h,k+1}\right\|\Big] \\
&\le R_\Lambda G\sqrt{\|H_{k+1}\|_{\mathrm{op}}}\,\varepsilon_h.
\end{aligned}$$

This proves (134).

Non-adjacent transitions: Wasserstein control through forgetting. Now we consider steps $q \ge k+2$. Since $q-1 > k$, the spatial forgetting estimate (76) gives

$$\mathrm{Lip}\Big(\tilde{\Phi}^{\lambda_q}_{k\to q-1}\Big) \le G\kappa_{k\to q-1}. \tag{141}$$

For an L_φ-Lipschitz vector-valued test function and any coupling ζ of measures μ, ν with finite first moments,

$$\|\mu\varphi - \nu\varphi\| = \left\|\int \big(\varphi(u) - \varphi(v)\big)\,\zeta(\mathrm{d}u,\mathrm{d}v)\right\| \le L_\varphi \int \|u - v\|\ \zeta(\mathrm{d}u,\mathrm{d}v).$$

Taking the infimum over couplings yields

$$\boxed{\|\mu\varphi - \nu\varphi\| \le L_\varphi W_1(\mu,\nu).} \tag{142}$$

The Gaussian kernels have finite first moments at each source state. Applying (142) with (141), we obtain

$$\begin{aligned}&\left\|\big(\widehat{K}^{\lambda_q}_{\theta,q}(\cdot \mid x_q) - K^{\lambda_q,\mathrm{G}}_q(\cdot \mid x_q)\big)\tilde{\Phi}^{\lambda_q}_{k\to q-1}\right\| \\ &\qquad \le G\kappa_{k\to q-1} W_1\Big(\widehat{K}^{\lambda_q}_{\theta,q}(\cdot \mid x_q), K^{\lambda_q,\mathrm{G}}_q(\cdot \mid x_q)\Big).\end{aligned} \tag{143}$$

Integrating under $\widehat{P}^+_q$, using $\|\lambda_q - \lambda^\star\| \le R_\Lambda$, and taking the outer expectation gives

$$\begin{aligned}|\mathsf{Guide}_{k\to q}| &\le R_\Lambda G\kappa_{k\to q-1}\,\mathbb{E}\Big[\widehat{P}^+_q W_1\Big(\widehat{K}^{\lambda_q}_{\theta,q}, K^{\lambda_q,\mathrm{G}}_q\Big)\Big] \\ &= R_\Lambda G\kappa_{k\to q-1}\,\mathbb{E}\Big[\widehat{P}^+_q\,\Big\|C_q\Delta^{\lambda_q}_{h,q}\Big\|\Big] \\ &\le R_\Lambda G\kappa_{k\to q-1}\,\|C_q\|_{\mathrm{op}}\,\varepsilon_h.\end{aligned}$$

Here the equality follows from (138), and the last inequality follows from 4.3. This proves (135), including at $q = T$. ■

For the subsequent summations, define the averaged local metric errors

$$\delta^{\mathrm{guide}}_q := \mathbb{E}\Big[\widehat{P}^+_q \mathrm{TV}\Big(\widehat{K}^{\lambda_q}_{\theta,q}, K^{\lambda_q,\mathrm{G}}_q\Big)\Big], \qquad 2 \le q \le T, \tag{144}$$

$$e^{\mathrm{guide}}_q := \mathbb{E}\Big[\widehat{P}^+_q W_1\Big(\widehat{K}^{\lambda_q}_{\theta,q}, K^{\lambda_q,\mathrm{G}}_q\Big)\Big], \qquad 3 \le q \le T. \tag{145}$$

The proof of Prop. D.12 gives

$$\delta^{\mathrm{guide}}_q \le \frac{1}{2}\,\mathbb{E}\left[\widehat{P}^+_q\,\left\|\Delta^{\lambda_q}_{h,q}\right\|_{H_q}\right] \le \frac{1}{2}\sqrt{\|H_q\|_{\mathrm{op}}}\,\varepsilon_h,$$

and

$$e^{\mathrm{guide}}_q = \mathbb{E}\Big[\widehat{P}^+_q\,\Big\|C_q\Delta^{\lambda_q}_{h,q}\Big\|\Big] \le \|C_q\|_{\mathrm{op}}\,\varepsilon_h.$$

These quantities are consequences of the pointwise guidance bounds. In this notation, the intermediate bounds established above read

$$|\mathsf{Guide}_{k\to k+1}| \le 2R_\Lambda G\,\delta^{\mathrm{guide}}_{k+1}, \qquad |\mathsf{Guide}_{k\to q}| \le R_\Lambda G\kappa_{k\to q-1}\,e^{\mathrm{guide}}_q \quad (q \ge k+2).$$

Corollary D.13 (Guidance-field control for fixed-large DDPM steps)**.** *Under the assumptions of Prop. D.12, specialize to fixed-large steps $\Sigma_q = \beta_q I_d$. Then*

$$\boxed{\begin{aligned}\delta^{\mathrm{guide}}_q &\le C_{\mathrm{guide}}\varepsilon_h, \quad 2 \le q \le T, \\ e^{\mathrm{guide}}_q &\le C_{\mathrm{guide}}\varepsilon_h, \quad 3 \le q \le T,\end{aligned}} \tag{146}$$

where $\beta_{\max} < 1$ is any prescribed upper bound on the β_q (in particular, one may take their maximum at a fixed horizon), and

$$C_{\mathrm{guide}} := \max\left\{\frac{\beta_{\max}}{\sqrt{1-\beta_{\max}}}, \frac{1}{2}\sqrt{\frac{\beta_{\max}}{1-\beta_{\max}}}\right\} < \infty. \tag{147}$$

Proof. For fixed-large DDPM steps,

$$C_q = \frac{\beta_q}{\sqrt{\alpha_q}} I_d, \qquad \Sigma_q = \beta_q I_d, \qquad H_q = C_q^\top \Sigma_q^{-1} C_q = \frac{\beta_q}{\alpha_q} I_d.$$

Since $\alpha_q = 1 - \beta_q \ge 1 - \beta_{\max} > 0$,

$$\|C_q\|_{\mathrm{op}} = \frac{\beta_q}{\sqrt{\alpha_q}} \le \frac{\beta_{\max}}{\sqrt{1-\beta_{\max}}}, \qquad \frac{1}{2}\sqrt{\|H_q\|_{\mathrm{op}}} = \frac{1}{2}\sqrt{\frac{\beta_q}{\alpha_q}} \le \frac{1}{2}\sqrt{\frac{\beta_{\max}}{1-\beta_{\max}}}.$$

Both conversion factors are therefore bounded by C_{guide} over the prescribed levels.

The Gaussian comparisons established in Prop. D.12, together with Ass. 4.3, give pointwise in the source state

$$\mathrm{TV}\Big(\widehat{K}^{\lambda}_{\theta,q}, K^{\lambda,\mathrm{G}}_{q}\Big) \le \frac{1}{2}\left\|\Delta^{\lambda}_{h,q}\right\|_{H_q} \le \frac{1}{2}\sqrt{\|H_q\|_{\mathrm{op}}}\left\|\Delta^{\lambda}_{h,q}\right\| \le C_{\mathrm{guide}}\varepsilon_h,$$
$$W_1\Big(\widehat{K}^{\lambda}_{\theta,q}, K^{\lambda,\mathrm{G}}_{q}\Big) = \left\|C_q\Delta^{\lambda}_{h,q}\right\| \le \|C_q\|_{\mathrm{op}}\left\|\Delta^{\lambda}_{h,q}\right\| \le C_{\mathrm{guide}}\varepsilon_h.$$

These inequalities hold at the random multiplier λ_q and remain valid after integration under $\widehat{P}^{+}_{q}$ and taking the outer expectation. The definitions (144) and (145) then give (146). Only $q \ge 2$ enters the kernel budgets for the dual average over levels $1, \dots, T$; the non-adjacent budget starts at $q = 3$. ■

D.4.7 Gaussian–Doob term $\mathsf{GD}_{k\to q}$

Proposition D.14 (Adjacent and non-adjacent Gaussian–Doob bounds)**.** *Under Ass. 4.3, for every* $1 \le k < q \le T$*,*

$$\boxed{|\mathsf{GD}_{k\to q}| \le R_\Lambda G \varepsilon_{\mathrm{GD}}.} \tag{148}$$

In particular, this controls the adjacent step $q = k+1$*. For every non-adjacent step* $k+2 \le q \le T$*, we also have*

$$\boxed{|\mathsf{GD}_{k\to q}| \le R_\Lambda G \kappa_{k\to q-1}\sqrt{\|\Sigma_q\|_{\mathrm{op}}}\,\varepsilon_{\mathrm{GD}}.} \tag{149}$$

All covariance coefficients are finite for the prescribed DDPM schedule.

Proof. Recall that the Gaussian–Doob contribution is

$$\mathsf{GD}_{k\to q} = \mathbb{E}\Big[\Big\langle \lambda_q - \lambda^\star,\, \widehat{P}^{+}_{q}\big(K^{\lambda_q,\mathrm{G}}_{q} - \tilde{K}^{\lambda_q}_{q}\big)\tilde{\Phi}^{\lambda_q}_{k\to q-1}\Big\rangle\Big].$$

We first compare the kernels at a fixed source state x and multiplier λ, and then apply these comparisons to the continuation.

Total-variation bounds. Write

$$D := \tilde{K}^{\lambda}_{q}(\cdot \mid x), \qquad Q := K^{\lambda,\mathrm{G}}_{q}(\cdot \mid x) = \mathcal{N}\Big(m_{\theta,q}(x) + C_q\nabla_x \log \tilde{h}^{\lambda}_{q}(x), \Sigma_q\Big).$$

Here $m_{\theta,q}$ is the unguided reverse mean introduced in Sec. D.4.6, evaluated under the exact-score assumption. Since $q \ge 2$, $\Sigma_q = \beta_q I_d$ is positive definite by the prescribed schedule. The Gaussian-Doob error assumption 4.3 gives

$$\mathrm{KL}(D\|Q) \le \frac{\varepsilon^2_{\mathrm{GD}}}{2}.$$

With the total-variation convention used in Sec. D.4.6, Pinsker's inequality therefore yields

$$\mathrm{TV}(D,Q) \le \sqrt{\frac{1}{2}\,\mathrm{KL}(D\|Q)} \le \frac{\varepsilon_{\mathrm{GD}}}{2}.$$

Wasserstein bound For the Wasserstein comparison, we use Talagrand's Gaussian inequality (Ledoux, 2019)

$$W_2(\nu, \mathcal{N}(0, I_d))^2 \le 2\,\mathrm{KL}(\nu\|\mathcal{N}(0, I_d)).$$

We introduce the invertible affine map

$$\mathcal{W}^{\lambda}_{q,x}(z) := \Sigma_q^{-1/2}\Big(z - m_{\theta,q}(x) - C_q\nabla_x \log \tilde{h}^{\lambda}_{q}(x)\Big). \tag{150}$$

By construction, $(\mathcal{W}^{\lambda}_{q,x})_{\#}Q = \mathcal{N}(0, I_d)$. Hence, applying the Gaussian transport inequality to $\nu = (\mathcal{W}^{\lambda}_{q,x})_{\#}D$ gives

$$W_2\big((\mathcal{W}^{\lambda}_{q,x})_{\#}D, (\mathcal{W}^{\lambda}_{q,x})_{\#}Q\big)^2 \le 2\,\mathrm{KL}\big((\mathcal{W}^{\lambda}_{q,x})_{\#}D\big\|(\mathcal{W}^{\lambda}_{q,x})_{\#}Q\big).$$

Since $\mathcal{W}_{q,x}^{\lambda}$ is invertible and is applied to both measures, relative entropy is invariant under this change of variables:

$$\mathrm{KL}\big((\mathcal{W}_{q,x}^{\lambda})_{\#}D\big\|(\mathcal{W}_{q,x}^{\lambda})_{\#}Q\big) = \mathrm{KL}(D\|Q).$$

It remains to return to the original coordinates. For any coupling (U, V) of the two laws, set

$$X := (\mathcal{W}_{q,x}^{\lambda})^{-1}(U), \qquad Y := (\mathcal{W}_{q,x}^{\lambda})^{-1}(V).$$

Then (X, Y) is a coupling of (D, Q). Since

$$(\mathcal{W}_{q,x}^{\lambda})^{-1}(u) = m_{\theta,q}(x) + C_q\nabla_x \log \tilde{h}_q^{\lambda}(x) + \Sigma_q^{1/2}u,$$

we have

$$\left\|(\mathcal{W}_{q,x}^{\lambda})^{-1}(u) - (\mathcal{W}_{q,x}^{\lambda})^{-1}(v)\right\| \le \left\|\Sigma_q^{1/2}\right\|_{\mathrm{op}} \|u - v\|.$$

Since this bound holds for every coupling (U, V) of $(\mathcal{W}_{q,x}^{\lambda})_{\#}D$ and $(\mathcal{W}_{q,x}^{\lambda})_{\#}Q$, and the corresponding (X, Y) is a coupling of D and Q, squaring and taking the infimum over all such couplings gives

$$W_2(D, Q) \le \left\|\Sigma_q^{1/2}\right\|_{\mathrm{op}} W_2\big((\mathcal{W}_{q,x}^{\lambda})_{\#}D, (\mathcal{W}_{q,x}^{\lambda})_{\#}Q\big),$$

and hence

$$W_2(D, Q)^2 \le 2\|\Sigma_q\|_{\mathrm{op}} \mathrm{KL}(D\|Q).$$

Finally, Cauchy–Schwarz applied to each coupling gives $W_1(D, Q) \le W_2(D, Q)$. Consequently,

$$\begin{aligned} W_1(D, Q) &\le W_2(D, Q) \\ &\le \sqrt{2\|\Sigma_q\|_{\mathrm{op}} \mathrm{KL}(D\|Q)} \\ &\le \sqrt{\|\Sigma_q\|_{\mathrm{op}}}\, \varepsilon_{\mathrm{GD}}. \end{aligned} \tag{151}$$

Thus both the TV and Wasserstein discrepancies are finite and both comparisons hold uniformly in x and λ.

Total-variation control of the Gaussian–Doob contribution. We first use the TV comparison. The difference of two Markov kernels annihilates the centering constant, so it acts identically on $\tilde{\Phi}$ and Φ. Hence, for every $q > k$,

$$\begin{aligned} &\big(K_q^{\lambda_q,\mathrm{G}} - \tilde{K}_q^{\lambda_q}\big)\tilde{\Phi}_{k\to q-1}^{\lambda_q} \\ &\quad = \big(K_q^{\lambda_q,\mathrm{G}} - \tilde{K}_q^{\lambda_q}\big)\big(\tilde{\Phi}_{k\to q-1}^{\lambda_q} + T_k^{\chi}(\lambda_q)\big) \\ &\quad = \big(K_q^{\lambda_q,\mathrm{G}} - \tilde{K}_q^{\lambda_q}\big)\Phi_{k\to q-1}^{\lambda_q}. \end{aligned}$$

By the definition of the centered continuation,

$$\tilde{\Phi}_{k\to q-1}^{\lambda_q}(z) + T_k^{\chi}(\lambda_q) = \Phi_{k\to q-1}^{\lambda_q}(z) = \mathbb{E}_{\lambda_q}[\chi_k^{\lambda_q}(\tilde{X}_k) \mid \tilde{X}_{q-1} = z],$$

whose norm is at most G. The vector-valued TV inequality (139) therefore applies to $\Phi_{k\to q-1}^{\lambda_q}$ with constant $2G$. At each source state x_q, it gives

$$\begin{aligned} &\left\|\big(K_q^{\lambda_q,\mathrm{G}}(\cdot \mid x_q) - \tilde{K}_q^{\lambda_q}(\cdot \mid x_q)\big)\tilde{\Phi}_{k\to q-1}^{\lambda_q}\right\| \\ &\quad \le 2G\,\mathrm{TV}\Big(K_q^{\lambda_q,\mathrm{G}}(\cdot \mid x_q), \tilde{K}_q^{\lambda_q}(\cdot \mid x_q)\Big). \end{aligned}$$

Integrating under $\widehat{P}_q^{+}$, using $\|P\psi\| \le P\|\psi\|$ and $\|\lambda_q - \lambda^{\star}\| \le R_{\Lambda}$, and taking the outer expectation yields

$$\begin{aligned} |\mathsf{GD}_{k\to q}| &\le 2R_{\Lambda}G\,\mathbb{E}\Big[\widehat{P}_q^{+}\mathrm{TV}\Big(K_q^{\lambda_q,\mathrm{G}}, \tilde{K}_q^{\lambda_q}\Big)\Big] \\ &\le R_{\Lambda}G\varepsilon_{\mathrm{GD}}. \end{aligned}$$

This proves (148) for all $q > k$, including the adjacent case $q = k+1$, where $\tilde{\Phi}_{k\to k}^{\lambda_q} = \chi_k^{\lambda_q} - T_k^{\chi}(\lambda_q)$. The constant $T_k^{\chi}(\lambda_q)$ cancels in the kernel difference.

Non-adjacent control through forgetting. Now suppose $q \geq k+2$. Since $q-1 > k$, the spatial forgetting estimate (76) gives

$$\mathrm{Lip}\Big(\tilde{\Phi}^{\lambda_q}_{k\to q-1}\Big) \leq G\kappa_{k\to q-1}.$$

The finite Wasserstein distance established above allows us to apply the coupling bound (142) at each source state:

$$\begin{aligned}&\Big\|\big(K_q^{\lambda_q,\mathrm{G}}(\cdot \mid x_q) - \tilde{K}_q^{\lambda_q}(\cdot \mid x_q)\big)\tilde{\Phi}^{\lambda_q}_{k\to q-1}\Big\| \\ &\qquad \leq G\kappa_{k\to q-1} W_1\Big(K_q^{\lambda_q,\mathrm{G}}(\cdot \mid x_q), \tilde{K}_q^{\lambda_q}(\cdot \mid x_q)\Big).\end{aligned}$$

Integrating under the same conditional law $\widehat{P}_q^+$, using $\|\lambda_q - \lambda^\star\| \leq R_\Lambda$, and taking the outer expectation, we obtain

$$\begin{aligned}|\mathsf{GD}_{k\to q}| &\leq R_\Lambda G\kappa_{k\to q-1}\,\mathbb{E}\Big[\widehat{P}_q^+ W_1\Big(K_q^{\lambda_q,\mathrm{G}}, \tilde{K}_q^{\lambda_q}\Big)\Big] \\ &\leq R_\Lambda G\kappa_{k\to q-1}\sqrt{\|\Sigma_q\|_{\mathrm{op}}}\,\varepsilon_{\mathrm{GD}}.\end{aligned}$$

The last inequality is (151). This proves (149), including the transition from $q = T$. ■

Averaged local metric errors. For the subsequent summations, define the averaged local metric errors

$$\delta_q^{\mathrm{GD}} := \mathbb{E}\Big[\widehat{P}_q^+ \mathrm{TV}\Big(K_q^{\lambda_q,\mathrm{G}}, \tilde{K}_q^{\lambda_q}\Big)\Big], \qquad 2 \leq q \leq T, \tag{152}$$

$$e_q^{\mathrm{GD}} := \mathbb{E}\Big[\widehat{P}_q^+ W_1\Big(K_q^{\lambda_q,\mathrm{G}}, \tilde{K}_q^{\lambda_q}\Big)\Big], \qquad 3 \leq q \leq T. \tag{153}$$

The pointwise comparisons in the proof of Prop. D.14 give

$$\boxed{\begin{aligned}&\delta_q^{\mathrm{GD}} \leq \frac{\varepsilon_{\mathrm{GD}}}{2}, && 2 \leq q \leq T,\\ &e_q^{\mathrm{GD}} \leq \sqrt{\|\Sigma_q\|_{\mathrm{op}}}\,\varepsilon_{\mathrm{GD}}, && 3 \leq q \leq T.\end{aligned}} \tag{154}$$

The intermediate termwise estimates read

$$|\mathsf{GD}_{k\to q}| \leq 2R_\Lambda G\,\delta_q^{\mathrm{GD}} \quad (q > k), \qquad |\mathsf{GD}_{k\to q}| \leq R_\Lambda G\kappa_{k\to q-1}\,e_q^{\mathrm{GD}} \quad (q \geq k+2).$$

D.4.8 Gaussian initialization remainder $\mathcal{R}_{k\to T}$

Proposition D.15 (Initialization remainder for the fixed schedule)**.** *Under Ass. 4.3, with $\widehat{X}_T \sim \mathcal{N}(0, I_d)$ and deterministic $\lambda_T = 0$,*

$$\boxed{\frac{1}{T}\sum_{k=1}^{T} |\mathcal{R}_{k\to T}| \leq \frac{C_{\mathrm{init}}}{T},} \tag{155}$$

where

$$\boxed{C_{\mathrm{init}} := R_\Lambda G\left[2 + (\sqrt{d} + M_{\mathrm{ref}})C_{\mathrm{col}}\right] < \infty,} \tag{156}$$

and the column aggregate is

$$\boxed{C_{\mathrm{col}} := \sup_{2\leq r\leq T}\sum_{k=1}^{r-1}\kappa_{k\to r} < \infty.} \tag{157}$$

For $T = 1$, the aggregate is defined as zero.

Proof. The recursion stops at level T, where $\lambda_T = 0$ is deterministic and $\widehat{X}_T \sim \mathcal{N}(0, I_d)$, so $\widehat{P}_T^+ = \mathcal{N}(0, I_d)$. Since $\tilde{P}^0 = \tilde{P}$, its boundary value is

$$\boxed{\mathcal{R}_{k\to T} = \Big\langle -\lambda^\star, (\widehat{P}_T^+ - \tilde{\pi}_T)\tilde{\Phi}^0_{k\to T} \Big\rangle, \qquad 1 \le k \le T.} \tag{158}$$

The outer expectation in the definition of $\mathcal{R}_{k\to T}$ disappears here because the multiplier and both measures are deterministic. The boundary term retains the difference between the Gaussian initialization and the forward marginal $\tilde{\pi}_T$.

At $\lambda = 0$, the ideal tilted marginal is the reference marginal: $\tilde{\pi}^0_T = \tilde{\pi}_T$. The centered-mean identity (73) therefore gives $\tilde{\pi}_T\tilde{\Phi}^0_{k\to T} = 0$. Consequently,

$$(\widehat{P}_T^+ - \tilde{\pi}_T)\tilde{\Phi}^0_{k\to T} = \mathbb{E}\big[\tilde{\Phi}^0_{k\to T}(\widehat{X}_T)\big]. \tag{159}$$

The factor $\sqrt{d}$ below comes from the first moment of this Gaussian initialization. Each coordinate $(\widehat{X}_T)_i$ has mean zero and variance one, hence

$$\mathbb{E}\left\|\widehat{X}_T\right\|^2 = \sum_{i=1}^d \mathbb{E}[(\widehat{X}_T)_i^2] = d, \qquad \mathbb{E}\left\|\widehat{X}_T\right\| \le \sqrt{\mathbb{E}\left\|\widehat{X}_T\right\|^2} = \sqrt{d},$$

where the inequality is Cauchy–Schwarz. For $k < T$, apply (74) at $\widehat{X}_T$. Since $0, \lambda^\star \in \Lambda$, $\|\lambda^\star\| \le R_\Lambda$, and thus

$$\begin{aligned} |\mathcal{R}_{k\to T}| &\le R_\Lambda\, \mathbb{E}\left\|\tilde{\Phi}^0_{k\to T}(\widehat{X}_T)\right\| \\ &\le R_\Lambda G\kappa_{k\to T}\big(\mathbb{E}\left\|\widehat{X}_T\right\| + M_{\mathrm{ref}}\big) \\ &\le R_\Lambda G\kappa_{k\to T}(\sqrt{d} + M_{\mathrm{ref}}), \qquad k < T. \end{aligned} \tag{160}$$

For $k = T$, $\tilde{\Phi}^0_{T\to T} = \chi^0_T - T^\chi_T(0)$ has sup-norm at most $2G$, since each term has norm at most G. Thus

$$|\mathcal{R}_{T\to T}| \le 2R_\Lambda G. \tag{161}$$

Summing these bounds and using the column sum at $r = T$ from (157) gives

$$\begin{aligned} \frac{1}{T}\sum_{k=1}^T |\mathcal{R}_{k\to T}| &\le \frac{R_\Lambda G}{T}\left[2 + (\sqrt{d} + M_{\mathrm{ref}})\sum_{k=1}^{T-1}\kappa_{k\to T}\right] \\ &\le \varepsilon_{\mathrm{init},T}, \end{aligned} \tag{162}$$

where we define the explicit initialization budget

$$\boxed{\varepsilon_{\mathrm{init},T} := \frac{R_\Lambda G}{T}\Big[2 + (\sqrt{d} + M_{\mathrm{ref}})C_{\mathrm{col}}\Big] = \frac{C_{\mathrm{init}}}{T}.} \tag{163}$$

For any prescribed schedule, the column aggregate C_{col} in (157) is finite by Prop. D.26. Since $R_\Lambda, G, M_{\mathrm{ref}}$, and d are finite, the displayed C_{init} is finite as well. If $\beta_t \ge \beta_{\min} > 0$ uniformly over the horizons, the geometric estimate (173) also bounds C_{col}, and hence C_{init}, independently of T. ■

D.4.9 BOUND ON THE SIGNED SAMPLER-LAW MISMATCH

Proposition D.16 (Signed sampler-law mismatch at the fixed finite horizon)**.** *Under Ass. 4.3,*

$$\boxed{\mathcal{L}_T^{\chi} \le A_{\mathrm{mov}}^{\chi}\eta + B_{\mathrm{mov}}^{\chi}\eta^2 + \Delta_{\mathrm{ker}} + \frac{C_{\mathrm{init}}}{T}.} \tag{164}$$

Here $\mathcal{L}_T^{\chi}$ is the signed mismatch from Sec. D.2.2, while C_{init} and C_{col} are given by Prop. D.15. Define the row aggregate

$$\boxed{C_{\mathrm{row}} := \sup_{1\le k\le T} \sum_{r=k+1}^{T} \kappa_{k\to r} < \infty.} \tag{165}$$

With M_{tot} and C_{diag}^{χ} from Prop. D.9 and D.11, the movement coefficients are

$$A_{\mathrm{mov}}^{\chi} := G(6G + 3R_\Lambda C_{\mathrm{diag}}^{\chi}) + 3M_{\mathrm{tot}}C_{\mathrm{row}}G(G + R_\Lambda C_{\mathrm{der}}^{\chi}), \tag{166}$$

$$B_{\mathrm{mov}}^{\chi} := 2C_{\mathrm{diag}}^{\chi}G^2 + 2M_{\mathrm{tot}}C_{\mathrm{row}}G^2C_{\mathrm{der}}^{\chi}. \tag{167}$$

With C_q, H_q from Prop. D.12, set

$$\boxed{\begin{aligned}\Delta_{\mathrm{ker}} := \frac{R_\Lambda G\varepsilon_h}{T}\left[\sum_{q=2}^{T}\sqrt{\|H_q\|_{\mathrm{op}}} + C_{\mathrm{col}}\sum_{q=3}^{T}\|C_q\|_{\mathrm{op}}\right] \\ + \frac{R_\Lambda G\varepsilon_{\mathrm{GD}}}{T}\left[(T-1) + C_{\mathrm{col}}\sum_{q=3}^{T}\sqrt{\|\Sigma_q\|_{\mathrm{op}}}\right].\end{aligned}} \tag{168}$$

Empty sums are zero. All these quantities are finite for the prescribed schedule and covariances.

For fixed-large steps $\Sigma_q = \beta_q I_d$, this kernel budget satisfies

$$\boxed{\begin{aligned}\Delta_{\mathrm{ker}} \le R_\Lambda G(2 + C_{\mathrm{col}})C_{\mathrm{guide}}\varepsilon_h \\ + R_\Lambda G(1 + \sqrt{\beta_{\max}}C_{\mathrm{col}})\varepsilon_{\mathrm{GD}},\end{aligned}} \tag{169}$$

where C_{guide} and $\beta_{\max}$ are defined in Cor. D.13.

Proof. At the base of the recursion,

$$\mathcal{R}_{k\to k} = \mathbb{E}\left\langle \lambda_k - \lambda^\star, (\widehat{P}_k^{+} - \tilde{\pi}_k^{\lambda_k})\chi_k^{\lambda_k}\right\rangle, \qquad \mathcal{L}_T^{\chi} = \frac{1}{T}\sum_{k=1}^{T}\mathcal{R}_{k\to k}.$$

For $k < T$, summing the identity of Prop. D.8 over $r = k, \dots, T-1$ gives

$$\mathcal{R}_{k\to k} = \mathcal{R}_{k\to T} + \sum_{r=k}^{T-1}\left(\mathsf{Sel}_{k\to r} + \mathsf{Rec}_{k\to r} + \mathsf{Guide}_{k\to r+1} + \mathsf{GD}_{k\to r+1}\right). \tag{170}$$

We retain the boundary remainder and the separate practical-guidance and Gaussian–Doob contributions. We then separate the diagonal adaptive contribution and adjacent kernel steps from the terms controlled by forgetting. Using Prop. D.11, D.12 and D.14 and the associated local metric budgets, we obtain

$$\boxed{\begin{aligned}|\mathcal{R}_{k\to k}| \le |\mathcal{R}_{k\to T}| + R_{\mathrm{base}}(\eta) \\ + \sum_{r=k+1}^{T-1}\left(|\mathsf{Sel}_{k\to r}| + |\mathsf{Rec}_{k\to r}|\right) \\ + 2R_\Lambda G\left(\delta_{k+1}^{\mathrm{guide}} + \delta_{k+1}^{\mathrm{GD}}\right) \\ + R_\Lambda G\sum_{q=k+2}^{T}\kappa_{k\to q-1}\left(e_q^{\mathrm{guide}} + e_q^{\mathrm{GD}}\right).\end{aligned}} \tag{171}$$

We average (171) over $k = 1, \dots, T-1$ and add $|\mathcal{R}_{T\to T}|/T$. Since by trinagular inequality

$$\mathcal{L}_T^\chi \le \frac{1}{T}\sum_{k=1}^{T} |\mathcal{R}_{k\to k}|,$$

it remains to bound the averaged movement, boundary, and kernel contributions.

For the off-diagonal movement terms, Prop. D.9 and D.10 give, for $1 \le k < r < T$,

$$\begin{aligned} |\mathsf{Sel}_{k\to r}| + |\mathsf{Rec}_{k\to r}| \le{}& 3M_{\mathrm{tot}}\kappa_{k\to r}\eta G(G + R_\Lambda C_{\mathrm{der}}^\chi) \\ &+ 2M_{\mathrm{tot}}\kappa_{k\to r}\eta^2 G^2 C_{\mathrm{der}}^\chi. \end{aligned} \tag{172}$$

For each fixed k, the sum of the weights is bounded by C_{row}:

$$\sum_{r=k+1}^{T-1} \kappa_{k\to r} \le \sum_{r=k+1}^{T} \kappa_{k\to r} \le C_{\mathrm{row}}.$$

Consequently,

$$\begin{aligned} &\frac{1}{T}\sum_{k=1}^{T-1}\sum_{r=k+1}^{T-1} \left(|\mathsf{Sel}_{k\to r}| + |\mathsf{Rec}_{k\to r}|\right) \\ &\quad\le \frac{T-1}{T} C_{\mathrm{row}} \left[3M_{\mathrm{tot}}\eta G(G + R_\Lambda C_{\mathrm{der}}^\chi) + 2M_{\mathrm{tot}}\eta^2 G^2 C_{\mathrm{der}}^\chi\right] \\ &\quad\le C_{\mathrm{row}} \left[3M_{\mathrm{tot}}\eta G(G + R_\Lambda C_{\mathrm{der}}^\chi) + 2M_{\mathrm{tot}}\eta^2 G^2 C_{\mathrm{der}}^\chi\right]. \end{aligned}$$

There is one diagonal selection/re-centering contribution for each $k < T$ and none for $k = T$. By Prop. D.11, their average is bounded by

$$\frac{T-1}{T} R_{\mathrm{base}}(\eta) \le R_{\mathrm{base}}(\eta) = \eta G(6G + 3R_\Lambda C_{\mathrm{diag}}^\chi) + 2C_{\mathrm{diag}}^\chi \eta^2 G^2.$$

Adding the diagonal and off-diagonal bounds and collecting the coefficients of η and η^2 yields

$$\begin{aligned} &\frac{1}{T}\sum_{k=1}^{T-1}\sum_{r=k}^{T-1} \left(|\mathsf{Sel}_{k\to r}| + |\mathsf{Rec}_{k\to r}|\right) \\ &\qquad\le A_{\mathrm{mov}}^\chi \eta + B_{\mathrm{mov}}^\chi \eta^2, \end{aligned}$$

with exactly the coefficients in (166)–(167).

The propagation constants used here and below are finite for the fixed schedule. Indeed, let

$$\beta_\star := \min_{1\le j\le T} \beta_j > 0, \qquad \theta := 1 - \beta_\star \in (0, 1).$$

The geometric estimate of Prop. D.26 in Sec. D.6, applied to this schedule, gives

$$\boxed{\max\{C_{\mathrm{row}}, C_{\mathrm{col}}\} \le \left(\frac{\sqrt{d}}{\sqrt{\beta_\star}} + \frac{R_X}{\beta_\star}\right) \frac{\sqrt{\theta}}{1-\sqrt{\theta}} < \infty.} \tag{173}$$

Moreover, Lemma D.3 gives $M_{\mathrm{tot}} = M_{\mathrm{alg}} + M_{\mathrm{ref}} < \infty$, and Lemma D.2 supplies the finite derivative constants. Thus A_{mov}^χ and B_{mov}^χ are finite. If $\beta_j \ge \beta_{\min} > 0$ uniformly over the horizons, then $\beta_\star \ge \beta_{\min}$ in (173), so both propagation sums are bounded by the same finite expression with $\beta_{\min}$ in place of $\beta_\star$ and $\theta = 1 - \beta_{\min}$. Combined with the uniform moment bound of Lemma D.3 and the horizon-independent constants $L_\chi, C_{\mathrm{diag}}^\chi, C_{\mathrm{der}}^\chi$ and M_{ref}, this gives uniform bounds for A_{mov}^χ and B_{mov}^χ.

For the boundary terms, Prop. D.15 gives directly

$$\frac{1}{T}\sum_{k=1}^{T} |\mathcal{R}_{k\to T}| \le \frac{C_{\mathrm{init}}}{T}.$$

The same finite column aggregate in (173) ensures $C_{\text{init}} < \infty$.

We now collect the kernel contributions. Each adjacent discrepancy appears once, with $q = k + 1$, so

$$\frac{2R_\Lambda G}{T} \sum_{k=1}^{T-1} \left(\delta_{k+1}^{\text{guide}} + \delta_{k+1}^{\text{GD}} \right) = \frac{2R_\Lambda G}{T} \sum_{q=2}^{T} \left(\delta_q^{\text{guide}} + \delta_q^{\text{GD}} \right).$$

For the non-adjacent terms, we reverse the order of the finite sum using

$$\{(k, q) : 1 \le k < T,\ k + 2 \le q \le T\} = \{(k, q) : 3 \le q \le T,\ 1 \le k \le q - 2\}.$$

For a fixed q, both local discrepancies e_q^{guide} and e_q^{GD} do not depend on the base index k. They can therefore be taken outside the inner sum. With $r = q - 1$, the remaining weights form a column sum

$$\sum_{k=1}^{q-2} \kappa_{k \to q-1} = \sum_{k=1}^{r-1} \kappa_{k \to r} \le C_{\text{col}}.$$

It follows that

$$\begin{aligned}
&\frac{R_\Lambda G}{T} \sum_{k=1}^{T-1} \sum_{q=k+2}^{T} \kappa_{k \to q-1} \left(e_q^{\text{guide}} + e_q^{\text{GD}} \right) \\
&= \frac{R_\Lambda G}{T} \sum_{q=3}^{T} \left(e_q^{\text{guide}} + e_q^{\text{GD}} \right) \sum_{k=1}^{q-2} \kappa_{k \to q-1} \\
&\le \frac{R_\Lambda G C_{\text{col}}}{T} \sum_{q=3}^{T} \left(e_q^{\text{guide}} + e_q^{\text{GD}} \right).
\end{aligned}$$

Thus both non-adjacent kernel contributions are controlled by averages of local errors multiplied by the finite column aggregate.

The local characterizations established in Prop. D.12 and D.14, using Ass. 4.3, are

$$\begin{aligned}
\delta_q^{\text{guide}} &\le \tfrac{1}{2} \sqrt{\|H_q\|_{\text{op}}}\, \varepsilon_h, \quad & e_q^{\text{guide}} &\le \|C_q\|_{\text{op}}\, \varepsilon_h, \\
\delta_q^{\text{GD}} &\le \tfrac{1}{2} \varepsilon_{\text{GD}}, & e_q^{\text{GD}} &\le \sqrt{\|\Sigma_q\|_{\text{op}}}\, \varepsilon_{\text{GD}},
\end{aligned}$$

at their respective relevant levels. Substitution into the adjacent and non-adjacent sums gives

$$\begin{aligned}
&\frac{2R_\Lambda G}{T} \sum_{q=2}^{T} \left(\delta_q^{\text{guide}} + \delta_q^{\text{GD}} \right) + \frac{R_\Lambda G C_{\text{col}}}{T} \sum_{q=3}^{T} \left(e_q^{\text{guide}} + e_q^{\text{GD}} \right) \\
&\le \frac{R_\Lambda G \varepsilon_h}{T} \left[\sum_{q=2}^{T} \sqrt{\|H_q\|_{\text{op}}} + C_{\text{col}} \sum_{q=3}^{T} \|C_q\|_{\text{op}} \right] \\
&\quad + \frac{R_\Lambda G \varepsilon_{\text{GD}}}{T} \left[(T - 1) + C_{\text{col}} \sum_{q=3}^{T} \sqrt{\|\Sigma_q\|_{\text{op}}} \right] \\
&= \Delta_{\text{ker}}.
\end{aligned}$$

The error levels are finite by assumption, and $C_q, \Sigma_q, \Sigma_q^{-1}$ are finite at the relevant prescribed levels. Together with $C_{\text{col}} < \infty$, this proves $\Delta_{\text{ker}} < \infty$.

Combining the movement, initialization, and kernel estimates in the averaged version of (171) yields

$$\mathcal{L}_T^\chi \le \frac{1}{T} \sum_{k=1}^{T} |\mathcal{R}_{k \to k}| \le A_{\text{mov}}^\chi \eta + B_{\text{mov}}^\chi \eta^2 + \Delta_{\text{ker}} + \frac{C_{\text{init}}}{T},$$

which proves (164).

Finally, for fixed-large steps, the calculation in Cor. D.13 gives

$$\sqrt{\|H_q\|_{\text{op}}} \le 2C_{\text{guide}}, \qquad \|C_q\|_{\text{op}} \le C_{\text{guide}}, \qquad \sqrt{\|\Sigma_q\|_{\text{op}}} = \sqrt{\beta_q} \le \sqrt{\beta_{\max}}.$$

For $T \geq 2$, counting the terms in each sum gives

$$\frac{1}{T}\sum_{q=2}^{T}\sqrt{\|H_q\|_{\mathrm{op}}} \leq \frac{2(T-1)}{T}C_{\mathrm{guide}} \leq 2C_{\mathrm{guide}},$$

$$\frac{1}{T}\sum_{q=3}^{T}\|C_q\|_{\mathrm{op}} \leq \frac{T-2}{T}C_{\mathrm{guide}} \leq C_{\mathrm{guide}},$$

$$\frac{1}{T}\sum_{q=3}^{T}\sqrt{\|\Sigma_q\|_{\mathrm{op}}} \leq \frac{T-2}{T}\sqrt{\beta_{\max}} \leq \sqrt{\beta_{\max}}.$$

Substituting these bounds into (168), and using $(T-1)/T \leq 1$, yields

$$\begin{aligned}\Delta_{\mathrm{ker}} \leq\ & 2R_\Lambda G C_{\mathrm{guide}}\varepsilon_h + R_\Lambda G C_{\mathrm{col}} C_{\mathrm{guide}}\varepsilon_h \\ & + R_\Lambda G\varepsilon_{\mathrm{GD}} + R_\Lambda G C_{\mathrm{col}}\sqrt{\beta_{\max}}\varepsilon_{\mathrm{GD}}.\end{aligned}$$

Regrouping gives (169). For $T = 1$, there are no kernel contributions: the adjacent kernel sums start at $q = 2$, and the non-adjacent sums start at $q = 3$. Both are therefore empty, so $\Delta_{\mathrm{ker}} = 0$ and the same bound holds. ■

D.5 GAUSSIAN BRIDGE AND SCHEDULE FORGETTING

For $0 \le k < r \le T$,

$$\boxed{\tilde{X}_r = a_{k\to r}\tilde{X}_k + b_{k\to r}Z, \qquad a_{k\to r} := \frac{\rho_r}{\rho_k}, \qquad b^2_{k\to r} := 1 - a^2_{k\to r},} \tag{174}$$

where $Z \sim \mathcal{N}(0, I_d)$ is independent of $\sigma(\tilde{X}_0, \dots, \tilde{X}_k)$. The arrow in $a_{k\to r}$ and $b_{k\to r}$ labels the forward interval from k to r.

For $r > k$, define the bridge variance proxy and forgetting coefficient

$$V_{k\to r} := d\frac{\sigma_k^2 b^2_{k\to r}}{\sigma_r^2} + R_X^2 \frac{\rho_k^2 b^4_{k\to r}}{\sigma_r^4}, \tag{175}$$

$$\kappa_{k\to r} := \frac{a_{k\to r}}{b^2_{k\to r}}\sqrt{V_{k\to r}}. \tag{176}$$

The coefficient $\kappa_{k\to r}$ is used only for $r > k$, where $b_{k\to r} > 0$.

This section derives the two forgetting estimates used in the mismatch proof. For every $r > k$, we will prove

$$\boxed{\left\|\tilde{\Phi}^\lambda_{k\to r}(y)\right\| \le G\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}),} \tag{177}$$

and

$$\boxed{\left\|J_\lambda \tilde{\Phi}^\lambda_{k\to r}(y)\right\|_{\mathrm{op}} \le C^\chi_{\mathrm{der}}\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}), \qquad C^\chi_{\mathrm{der}} = L_\chi + 3G^2.} \tag{178}$$

The first inequality says that, as the channel from $\tilde{X}_k$ to $\tilde{X}_r$ becomes noisier, conditioning on $\tilde{X}_r = y$ gives less information about the base comparison statistic $\chi^\lambda_k(\tilde{X}_k)$. The second says that the same forgetting also applies to the multiplier sensitivity of the centered continuation. These are precisely the two estimates used in the selection and re-centering controls of Sec. D.4.

D.5.1 THE FORWARD CHANNEL REMAINS GAUSSIAN UNDER THE TILTED LAW

Lemma D.17. *For every $r > k$ and every $\lambda \in \Lambda$,*

$$\boxed{\tilde{P}^\lambda(\tilde{X}_r \in \cdot \mid \tilde{X}_k = x) = \tilde{P}(\tilde{X}_r \in \cdot \mid \tilde{X}_k = x) = \mathcal{N}(a_{k\to r}x, b^2_{k\to r}I_d).} \tag{179}$$

Proof. Recall the path tilt

$$\frac{\mathrm{d}\tilde{P}^\lambda}{\mathrm{d}\tilde{P}} \propto e^{-\lambda^\top g(\tilde{X}_0)},$$

so the likelihood ratio depends only on $\tilde{X}_0$. Under the original forward chain we have the Markov relation

$$\tilde{X}_0 \longrightarrow \tilde{X}_k \longrightarrow \tilde{X}_r,$$

which means $\tilde{X}_0$ and $\tilde{X}_r$ are conditionally independent given $\tilde{X}_k$.

For any bounded test function ψ,

$$\begin{aligned}
\mathbb{E}_\lambda[\psi(\tilde{X}_r) \mid \tilde{X}_k] &= \frac{\mathbb{E}_{\tilde{P}}[e^{-\lambda^\top g(\tilde{X}_0)}\psi(\tilde{X}_r) \mid \tilde{X}_k]}{\mathbb{E}_{\tilde{P}}[e^{-\lambda^\top g(\tilde{X}_0)} \mid \tilde{X}_k]} \\
&= \frac{\mathbb{E}_{\tilde{P}}[e^{-\lambda^\top g(\tilde{X}_0)} \mid \tilde{X}_k]\,\mathbb{E}_{\tilde{P}}[\psi(\tilde{X}_r) \mid \tilde{X}_k]}{\mathbb{E}_{\tilde{P}}[e^{-\lambda^\top g(\tilde{X}_0)} \mid \tilde{X}_k]} \\
&= \mathbb{E}_{\tilde{P}}[\psi(\tilde{X}_r) \mid \tilde{X}_k].
\end{aligned} \tag{180}$$

Thus the conditional kernel is unchanged. Recalling the VP forward channel (174),

$$\tilde{X}_r = a_{k\to r}\tilde{X}_k + b_{k\to r}Z, \qquad Z \sim \mathcal{N}(0, I_d),$$

gives (179). ∎

Lemma D.18. *For $r > k$, conditionally on $(\tilde{X}_0 = x_0, \tilde{X}_r = x_r)$,*

$$\tilde{X}_k \sim \mathcal{N}(A_{k\to r} x_0 + B_{k\to r} x_r, v_{k\to r} I_d)\,, \tag{181}$$

where

$$\boxed{A_{k\to r} = \frac{\rho_k b_{k\to r}^2}{\sigma_r^2}, \qquad B_{k\to r} = \frac{a_{k\to r}\sigma_k^2}{\sigma_r^2}, \qquad v_{k\to r} = \frac{\sigma_k^2 b_{k\to r}^2}{\sigma_r^2}.} \tag{182}$$

Proof. Condition on $\tilde{X}_0 = x_0$. The tilt depends only on $\tilde{X}_0$, so it does not change the conditional forward law given $\tilde{X}_0$. Recalling the VP channel (174)

$$\tilde{X}_k = \rho_k x_0 + \sigma_k Z_1, \qquad \tilde{X}_r = a_{k\to r}\tilde{X}_k + b_{k\to r} Z_2, \qquad \tilde{X}_r = \rho_r x_0 + a_{k\to r}\sigma_k Z_1 + b_{k\to r} Z_2,$$

where Z_1, Z_2 are independent $\mathcal{N}(0, I_d)$ random vectors. Thus $(\tilde{X}_k, \tilde{X}_r) \mid \tilde{X}_0 = x_0$ is jointly Gaussian, being an affine transformation of (Z_1, Z_2), with marginal laws

$$\tilde{X}_k \mid \tilde{X}_0 = x_0 \sim \mathcal{N}(\rho_k x_0, \sigma_k^2 I_d), \qquad \tilde{X}_r \mid \tilde{X}_0 = x_0 \sim \mathcal{N}(\rho_r x_0, \sigma_r^2 I_d).$$

Moreover, independence of the inter-level noise gives

$$\begin{aligned}\mathrm{Cov}_\lambda(\tilde{X}_k, \tilde{X}_r \mid \tilde{X}_0) &= \mathrm{Cov}_\lambda(\tilde{X}_k, a_{k\to r}\tilde{X}_k + b_{k\to r} Z_2 \mid \tilde{X}_0)\\ &= a_{k\to r}\,\mathrm{Var}_\lambda(\tilde{X}_k \mid \tilde{X}_0)\\ &= a_{k\to r}\sigma_k^2 I_d.\end{aligned}$$

Recall that, for a jointly Gaussian pair (U, V) with means m_U, m_V and covariance blocks $S_{UU}, S_{UV}, S_{VU}, S_{VV}$, where S_{VV} is invertible,

$$U \mid V = y \sim \mathcal{N}\big(m_U + S_{UV}S_{VV}^{-1}(y - m_V), S_{UU} - S_{UV}S_{VV}^{-1}S_{VU}\big)\,.$$

We apply this formula conditionally on $\tilde{X}_0 = x_0$, with $U = \tilde{X}_k$, $V = \tilde{X}_r$, $m_U = \rho_k x_0$, $m_V = \rho_r x_0$, $S_{UU} = \sigma_k^2 I_d$, $S_{UV} = S_{VU} = a_{k\to r}\sigma_k^2 I_d$, and $S_{VV} = \sigma_r^2 I_d$. Since $\sigma_r^2 > 0$, this gives

$$\mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_0 = x_0, \tilde{X}_r = x_r] = \rho_k x_0 + \frac{a_{k\to r}\sigma_k^2}{\sigma_r^2}(x_r - \rho_r x_0), \tag{183}$$

$$\mathrm{Var}_\lambda(\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r) = \sigma_k^2 I_d - \frac{a_{k\to r}^2\sigma_k^4}{\sigma_r^2} I_d. \tag{184}$$

The coefficient of x_r in (183) is

$$B_{k\to r} = \frac{a_{k\to r}\sigma_k^2}{\sigma_r^2}.$$

For the coefficient of x_0, using $\rho_r = a_{k\to r}\rho_k$ yields

$$\rho_k - \frac{a_{k\to r}\sigma_k^2\rho_r}{\sigma_r^2} = \frac{\rho_k}{\sigma_r^2}\left(\sigma_r^2 - a_{k\to r}^2\sigma_k^2\right).$$

Since $\sigma_t^2 = 1 - \rho_t^2$

$$\begin{aligned}\sigma_r^2 - a_{k\to r}^2\sigma_k^2 &= (1 - \rho_r^2) - \frac{\rho_r^2}{\rho_k^2}(1 - \rho_k^2)\\ &= 1 - \frac{\rho_r^2}{\rho_k^2} = 1 - a_{k\to r}^2 = b_{k\to r}^2.\end{aligned} \tag{185}$$

This proves the formula for $A_{k\to r}$. Using the identity above (185) and inserting it into (184) gives

$$v_{k\to r} = \frac{\sigma_k^2}{\sigma_r^2}\left(\sigma_r^2 - a_{k\to r}^2\sigma_k^2\right) = \frac{\sigma_k^2 b_{k\to r}^2}{\sigma_r^2}.$$

■

Lemma D.19. *For every $r > k$ and every $y \in \mathbb{R}^d$,*

$$\boxed{\mathbb{E}_\lambda\left[\left\|\tilde{X}_k - \mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_r]\right\|^2 \middle| \tilde{X}_r = y\right] \le V_{k\to r},} \tag{186}$$

where $V_{k\to r}$ is defined in (175)*. The bound is independent of y and λ.*

Proof. We first derive the conditional variance decomposition, then evaluate its two contributions using lemma Lemma D.18.

Conditional variance decomposition Fix λ and the levels $k < r$. For this proof only, we abbreviate the following conditional means by

$$\mu := \mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_r], \qquad \nu := \mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r].$$

Both are random vectors in $\mathbb{R}^d$. The first is measurable with respect to $\sigma(\tilde{X}_r)$, and the second with respect to the larger sigma-field $\sigma(\tilde{X}_0, \tilde{X}_r)$. All the variables below are square-integrable. Indeed $\tilde{X}_0$ is bounded and the forward noise has finite second moments. The tower property gives

$$\mathbb{E}_\lambda[\nu \mid \tilde{X}_r] = \mathbb{E}_\lambda\Big[\mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r] \mid \tilde{X}_r\Big] = \mu.$$

Adding and subtracting ν gives

$$\tilde{X}_k - \mu = (\tilde{X}_k - \nu) + (\nu - \mu).$$

Expanding the squared Euclidean norm andconditioning on $\tilde{X}_r$ yields

$$\begin{aligned}\mathbb{E}_\lambda[\Big\|\tilde{X}_k - \mu\Big\|^2 \mid \tilde{X}_r] = \; & \mathbb{E}_\lambda[\Big\|\tilde{X}_k - \nu\Big\|^2 \mid \tilde{X}_r] + \mathbb{E}_\lambda[\|\nu - \mu\|^2 \mid \tilde{X}_r] \\ & + 2\,\mathbb{E}_\lambda[\Big\langle \tilde{X}_k - \nu, \nu - \mu \Big\rangle \mid \tilde{X}_r].\end{aligned}$$

To cancel the inner product, first condition on $(\tilde{X}_0, \tilde{X}_r)$. Under this conditioning, $\nu - \mu$ is measurable, while

$$\mathbb{E}_\lambda[\tilde{X}_k - \nu \mid \tilde{X}_0, \tilde{X}_r] = \nu - \nu = 0.$$

We can therefore take the second factor outside the inner conditional expectation. The tower property then gives

$$\begin{aligned}\mathbb{E}_\lambda[\Big\langle \tilde{X}_k - \nu, \nu - \mu \Big\rangle \mid \tilde{X}_r] &= \mathbb{E}_\lambda\Big[\mathbb{E}_\lambda[\Big\langle \tilde{X}_k - \nu, \nu - \mu \Big\rangle \mid \tilde{X}_0, \tilde{X}_r]\Big|\tilde{X}_r\Big] \\ &= \mathbb{E}_\lambda\Big[\Big\langle \mathbb{E}_\lambda[\tilde{X}_k - \nu \mid \tilde{X}_0, \tilde{X}_r], \nu - \mu \Big\rangle\Big|\tilde{X}_r\Big] = 0.\end{aligned}$$

Thus exactly two terms remain:

$$\mathbb{E}_\lambda[\Big\|\tilde{X}_k - \mu\Big\|^2 \mid \tilde{X}_r] = \mathbb{E}_\lambda[\Big\|\tilde{X}_k - \nu\Big\|^2 \mid \tilde{X}_r] + \mathbb{E}_\lambda[\|\nu - \mu\|^2 \mid \tilde{X}_r]. \tag{187}$$

Trace of the conditional covariance The conditional variance of a vector is a $d \times d$ covariance matrix:

$$\mathrm{Var}_\lambda(\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r) = \mathbb{E}_\lambda[(\tilde{X}_k - \nu)(\tilde{X}_k - \nu)^\top \mid \tilde{X}_0, \tilde{X}_r].$$

For any column vector $z \in \mathbb{R}^d$, $\mathrm{tr}(zz^\top) = \sum_{i=1}^d z_i^2 = \|z\|^2$. Therefore taking the trace yields

$$\begin{aligned}\mathrm{tr}\,\mathrm{Var}_\lambda(\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r) &= \mathbb{E}_\lambda[\mathrm{tr}((\tilde{X}_k - \nu)(\tilde{X}_k - \nu)^\top) \mid \tilde{X}_0, \tilde{X}_r] \\ &= \mathbb{E}_\lambda[\Big\|\tilde{X}_k - \nu\Big\|^2 \mid \tilde{X}_0, \tilde{X}_r].\end{aligned}$$

The first term in (187) is conditioned only on $\tilde{X}_r$. Applying the tower property gives

$$\begin{aligned}\mathbb{E}_\lambda[\Big\|\tilde{X}_k - \nu\Big\|^2 \mid \tilde{X}_r] &= \mathbb{E}_\lambda\Big[\mathbb{E}_\lambda[\Big\|\tilde{X}_k - \nu\Big\|^2 \mid \tilde{X}_0, \tilde{X}_r]\Big|\tilde{X}_r\Big] \\ &= \mathbb{E}_\lambda[\mathrm{tr}\,\mathrm{Var}_\lambda(\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r) \mid \tilde{X}_r].\end{aligned}$$

Substituting this identity into the two-term decomposition and restoring the full expressions for μ and ν give

$$\begin{aligned}&\mathbb{E}_\lambda\Big[\Big\|\tilde{X}_k - \mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_r]\Big\|^2\Big|\tilde{X}_r\Big] \\ &\quad = \mathbb{E}_\lambda\Big[\mathrm{tr}\,\mathrm{Var}_\lambda(\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r)\Big|\tilde{X}_r\Big] + \mathbb{E}_\lambda\Big[\Big\|\mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r] - \mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_r]\Big\|^2\Big|\tilde{X}_r\Big].\end{aligned} \tag{188}$$

Bound the conditional variance By (181) $\tilde{X}_k$ has conditional covariance $v_{k\to r} I_d$, with $v_{k\to r} = \sigma_k^2 b_{k\to r}^2/\sigma_r^2$. Its trace is $v_{k\to r}\,\mathrm{tr}(I_d) = d\, v_{k\to r}$. Wich depends only on the levels and the schedule, so taking its conditional expectation leaves it unchanged. Therefore the first term in 188 is

$$d\, v_{k\to r} = d\frac{\sigma_k^2 b_{k\to r}^2}{\sigma_r^2}.$$

For the second term,

$$\mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r] = A_{k\to r}\tilde{X}_0 + B_{k\to r}\tilde{X}_r.$$

Conditioning on $\tilde{X}_r$ and using the tower property gives the coarser mean. The bridge coefficients are deterministic, and $\tilde{X}_r$ is measurable under this conditioning therefore

$$\mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_r] = A_{k\to r}\,\mathbb{E}_\lambda[\tilde{X}_0 \mid \tilde{X}_r] + B_{k\to r}\tilde{X}_r.$$

Hence their difference is

$$A_{k\to r}\big(\tilde{X}_0 - \mathbb{E}_\lambda[\tilde{X}_0 \mid \tilde{X}_r]\big).$$

Therefore

$$\begin{aligned}\mathbb{E}_\lambda\left[\left\|\mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_0, \tilde{X}_r] - \mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_r]\right\|^2 \middle| \tilde{X}_r\right] \\ = A_{k\to r}^2\,\mathbb{E}_\lambda\left[\left\|\tilde{X}_0 - \mathbb{E}_\lambda[\tilde{X}_0 \mid \tilde{X}_r]\right\|^2 \middle| \tilde{X}_r\right].\end{aligned}$$

To bound the remaining conditional fluctuation of $\tilde{X}_0$, we expand its squared norm. The cross term becomes the squared norm of its conditional mean, because that mean is $\sigma(\tilde{X}_r)$-measurable. Using $\left\|\tilde{X}_0\right\| \le R_X$ therefore gives

$$\begin{aligned}\mathbb{E}_\lambda[\left\|\tilde{X}_0 - \mathbb{E}_\lambda[\tilde{X}_0 \mid \tilde{X}_r]\right\|^2 \mid \tilde{X}_r] &= \mathbb{E}_\lambda[\left\|\tilde{X}_0\right\|^2 \mid \tilde{X}_r] - \left\|\mathbb{E}_\lambda[\tilde{X}_0 \mid \tilde{X}_r]\right\|^2 \\ &\le R_X^2.\end{aligned}$$

Since the bridge coefficient in (182) is $A_{k\to r} = \rho_k b_{k\to r}^2/\sigma_r^2$, the second term in the total-variance decomposition (188) is bounded by

$$R_X^2\,\frac{\rho_k^2 b_{k\to r}^4}{\sigma_r^4}. \tag{189}$$

Adding these two explicit contributions gives the bound in (175). ■

D.5.2 Generic centered forgetting through the Gaussian channel

Let U be a bounded scalar random variable measurable with respect to the path up to level k. Define

$$H_U(y) := \mathbb{E}_\lambda[U \mid \tilde{X}_r = y]. \tag{190}$$

Lemma D.20. *For $r > k$,*

$$\boxed{\nabla_y H_U(y) = \frac{a_{k\to r}}{b_{k\to r}^2}\,\mathrm{Cov}_\lambda(\tilde{X}_k, U \mid \tilde{X}_r = y).} \tag{191}$$

Proof. We write $a = a_{k\to r}$ and $b = b_{k\to r}$ to lighten notation. Equation (179) states explicitly that

$$\tilde{X}_r \mid \tilde{X}_k = x \sim \mathcal{N}(ax, b^2 I_d) \qquad \text{under } \tilde{P}^\lambda.$$

Hence the conditional density of $\tilde{X}_r = y$ given $\tilde{X}_k = x$ is

$$q(y \mid x) = \frac{1}{(2\pi b^2)^{d/2}}\exp\left(-\frac{\|y - ax\|^2}{2b^2}\right).$$

Let $\mu_{U,k}^\lambda(\mathrm{d}u, \mathrm{d}x)$ denote the joint law of $(U, \tilde{X}_k)$ under $\tilde{P}^\lambda$. The joint law of $(U, \tilde{X}_k, \tilde{X}_r)$ then factorizes as

$$\mu_{U,k}^\lambda(\mathrm{d}u, \mathrm{d}x)\, q(y \mid x)\, \mathrm{d}y.$$

Consequently the marginal density of $\tilde{X}_r$ is

$$\tilde{p}_r^{\lambda}(y) = \int q(y \mid x)\, \mu_{U,k}^{\lambda}(\mathrm{d}u, \mathrm{d}x). \tag{192}$$

Bayes' formula therefore gives

$$\tilde{P}^{\lambda}(\mathrm{d}u, \mathrm{d}x \mid \tilde{X}_r = y) = \frac{q(y \mid x)\mu_{U,k}^{\lambda}(\mathrm{d}u, \mathrm{d}x)}{\int q(y \mid x')\mu_{U,k}^{\lambda}(\mathrm{d}u', \mathrm{d}x')}. \tag{193}$$

Integrating u against this law yields

$$H_U(y) = \frac{\int u\, q(y \mid x)\mu_{U,k}^{\lambda}(\mathrm{d}u, \mathrm{d}x)}{\int q(y \mid x)\mu_{U,k}^{\lambda}(\mathrm{d}u, \mathrm{d}x)}. \tag{194}$$

Set

$$n_U(y) := \int u\, q(y \mid x)\mu_{U,k}^{\lambda}(\mathrm{d}u, \mathrm{d}x), \qquad \tilde{p}_r^{\lambda}(y) = \int q(y \mid x)\mu_{U,k}^{\lambda}(\mathrm{d}u, \mathrm{d}x).$$

Then $H_U(y) = n_U(y)/\tilde{p}_r^{\lambda}(y)$. The denominator is the marginal density of $\tilde{X}_r$.

We now justify differentiation under the integral. Since $r > k$, $b > 0$. The Gaussian kernel is smooth and

$$\nabla_y q(y \mid x) = -\frac{y - ax}{b^2} q(y \mid x). \tag{195}$$

Then using $te^{-t^2/(2b^2)} \le be^{-1/2}$ we obtain

$$\begin{aligned} \|\nabla_y q(y \mid x)\| &= \frac{1}{(2\pi b^2)^{d/2}} \frac{\|y - ax\|}{b^2} \exp\left(-\frac{\|y - ax\|^2}{2b^2}\right) \\ &\le \frac{e^{-1/2}}{b(2\pi b^2)^{d/2}}, \end{aligned}$$

This bound is independent of x and y. Since U is bounded and $\mu_{U,k}^{\lambda}$ is a probability measure, dominated convergence allow differentiation under both integrals in (194). Hence

$$\nabla_y H_U(y) = \frac{\nabla_y n_U(y)}{\tilde{p}_r^{\lambda}(y)} - \frac{n_U(y)}{\tilde{p}_r^{\lambda}(y)} \frac{\nabla_y \tilde{p}_r^{\lambda}(y)}{\tilde{p}_r^{\lambda}(y)}. \tag{196}$$

Using $\nabla_y q = q \nabla_y \log q$ and the Bayes law (193) give,

$$\begin{aligned} \frac{\nabla_y n_U(y)}{\tilde{p}_r^{\lambda}(y)} &= \mathbb{E}_{\lambda}[U \nabla_y \log q(y \mid \tilde{X}_k) \mid \tilde{X}_r = y], \\ \frac{\nabla_y \tilde{p}_r^{\lambda}(y)}{\tilde{p}_r^{\lambda}(y)} &= \mathbb{E}_{\lambda}[\nabla_y \log q(y \mid \tilde{X}_k) \mid \tilde{X}_r = y]. \end{aligned}$$

Substituting these expressions into (196) gives

$$\nabla_y H_U(y) = \mathrm{Cov}_{\lambda}(\nabla_y \log q(y \mid \tilde{X}_k), U \mid \tilde{X}_r = y). \tag{197}$$

The Gaussian channel score is

$$\nabla_y \log q(y \mid x) = -\frac{y - ax}{b^2}.$$

Conditionally on $\tilde{X}_r = y$, the vector y is constant and has zero covariance with U. Therefore

$$\begin{aligned} \nabla_y H_U(y) &= \mathrm{Cov}_{\lambda}\left(-\frac{y - a\tilde{X}_k}{b^2}, U \,\middle|\, \tilde{X}_r = y\right) \\ &= \frac{a}{b^2} \mathrm{Cov}_{\lambda}(\tilde{X}_k, U \mid \tilde{X}_r = y), \end{aligned}$$

which proves (191). ■

Lemma D.21 (Lipschitz estimate)**.** *For bounded scalar U measurable up to level k,*

$$\boxed{\|\nabla_y H_U(y)\| \le \|U\|_\infty \, \kappa_{k\to r}.} \tag{198}$$

Proof. Let

$$A := U - \mathbb{E}_\lambda[U \mid \tilde{X}_r], \qquad B := \tilde{X}_k - \mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_r].$$

Then, conditionally on $\tilde{X}_r = y$,

$$\mathrm{Cov}_\lambda(\tilde{X}_k, U \mid \tilde{X}_r = y) = \mathbb{E}_\lambda[AB \mid \tilde{X}_r = y].$$

Since A is scalar and B is vector-valued we obtain

$$\begin{aligned}\Big\|\mathrm{Cov}_\lambda(\tilde{X}_k, U \mid \tilde{X}_r = y)\Big\| &= \Big\|\mathbb{E}_\lambda[AB \mid \tilde{X}_r = y]\Big\| \\ &\le \mathbb{E}_\lambda[|A| \, \|B\| \mid \tilde{X}_r = y] \\ &\le \sqrt{\mathbb{E}_\lambda[A^2 \mid \tilde{X}_r = y]}\sqrt{\mathbb{E}_\lambda[\|B\|^2 \mid \tilde{X}_r = y]}.\end{aligned}$$

The last line is conditional Cauchy–Schwarz. The first factor is

$$\sqrt{\mathrm{Var}_\lambda(U \mid \tilde{X}_r = y)} \le \sqrt{\mathbb{E}_\lambda[U^2 \mid \tilde{X}_r = y]} \le \|U\|_\infty . \tag{199}$$

The conditional-variance estimate (186) of Lemma D.19,with $V_{k\to r}$ defined in (175), states

$$\mathbb{E}_\lambda[\Big\|\tilde{X}_k - \mathbb{E}_\lambda[\tilde{X}_k \mid \tilde{X}_r]\Big\|^2 \mid \tilde{X}_r = y] \le V_{k\to r},$$

so the second factor is at most $\sqrt{V_{k\to r}}$. Hence

$$\Big\|\mathrm{Cov}_\lambda(\tilde{X}_k, U \mid \tilde{X}_r = y)\Big\| \le \|U\|_\infty \sqrt{V_{k\to r}}.$$

Equation (191) in Lemma D.20 multiplies this covariance by $a_{k\to r}/b_{k\to r}^2$. Since (176) defines $\kappa_{k\to r} = (a_{k\to r}/b_{k\to r}^2)\sqrt{V_{k\to r}}$ we obtain

$$\|\nabla_y H_U(y)\| \le \|U\|_\infty \frac{a_{k\to r}}{b_{k\to r}^2}\sqrt{V_{k\to r}} = \|U\|_\infty \, \kappa_{k\to r}. \tag{200}$$

This gradient bound is uniform in the observation. Fix $y, y' \in \mathbb{R}^d$ and integrate along the segment $y_s = y' + s(y - y')$; the fundamental theorem of calculus yields

$$\begin{aligned}|H_U(y) - H_U(y')| &= \left|\int_0^1 \nabla_y H_U(y_s)^\top (y - y') \, \mathrm{d}s\right| \\ &\le \int_0^1 \|\nabla_y H_U(y_s)\| \, \|y - y'\| \, \mathrm{d}s \\ &\le \|U\|_\infty \, \kappa_{k\to r} \, \|y - y'\| .\end{aligned}$$

■

Lemma D.22 (Centered forgetting for any bounded observable)**.** *Fix λ. Let U be a bounded scalar observable measurable with respect to $\sigma(\tilde{X}_0, \dots, \tilde{X}_k)$. Then, for $r > k$,*

$$\boxed{\Big|\mathbb{E}_\lambda[U \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[U]\Big| \le \|U\|_\infty \, \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}),} \tag{201}$$

where

$$M_{\mathrm{ref}} = R_X + \sqrt{d}. \tag{202}$$

Proof. By the tower property,

$$\mathbb{E}_\lambda[H_U(\tilde{X}_r)] = \mathbb{E}_\lambda[\mathbb{E}_\lambda[U \mid \tilde{X}_r]] = \mathbb{E}_\lambda[U].$$

Since y is fixed, $H_U(y)$ is constant with respect to the randomness of $\tilde{X}_r$. Jensen's inequality therefore gives

$$\begin{aligned} |H_U(y) - \mathbb{E}_\lambda U| &= \left| H_U(y) - \mathbb{E}_\lambda H_U(\tilde{X}_r) \right| \\ &\le \mathbb{E}_\lambda \left| H_U(y) - H_U(\tilde{X}_r) \right| \\ &\le \|U\|_\infty \kappa_{k\to r} \, \mathbb{E}_\lambda \left\| y - \tilde{X}_r \right\|, \end{aligned} \tag{203}$$

where the last line uses the Lipschitz estimate (198),

$$\|\nabla_y H_U(z)\| \le \|U\|_\infty \kappa_{k\to r} \quad \text{for every } z,$$

so $|H_U(y) - H_U(\tilde{X}_r)| \le \|U\|_\infty \kappa_{k\to r} \left\| y - \tilde{X}_r \right\|$. Next,

$$\mathbb{E}_\lambda \left\| y - \tilde{X}_r \right\| \le \|y\| + \mathbb{E}_\lambda \left\| \tilde{X}_r \right\|.$$

Recalling that 179 gives

$$\tilde{X}_r = \rho_r \tilde{X}_0 + \sigma_r Z, \qquad Z \sim \mathcal{N}(0, I_d),$$

with the Gaussian noise independent of $\tilde{X}_0$. Hence

$$\begin{aligned} \mathbb{E}_\lambda \left\| \tilde{X}_r \right\| &\le \rho_r \, \mathbb{E}_\lambda \left\| \tilde{X}_0 \right\| + \sigma_r \, \mathbb{E}_\lambda \|Z\| \\ &\le R_X + \sqrt{\mathbb{E}_\lambda \|Z\|^2} = R_X + \sqrt{d} = M_{\mathrm{ref}}. \end{aligned} \tag{204}$$

Substitute into (203) gives the result. ■

D.5.3 Application to the centered continuation and its multiplier derivative

Proposition D.23 (Forgetting of $\tilde{\Phi}^\lambda_{k\to r}$)**.** *For every $r > k$,*

$$\boxed{\left\| \tilde{\Phi}^\lambda_{k\to r}(y) \right\| \le G\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}).} \tag{205}$$

Moreover, both continuations have the same spatial Lipschitz bound:

$$\boxed{\mathrm{Lip}(\Phi^\lambda_{k\to r}) = \mathrm{Lip}(\tilde{\Phi}^\lambda_{k\to r}) \le G\kappa_{k\to r}.} \tag{206}$$

Proof. The forgetting Lemma D.22 is scalar, whereas the continuation is $\mathbb{R}^m$-valued. We therefore project into an arbitrary unit output direction $u \in \mathbb{R}^m$, using the vector duality. Define the scalar bounded observable

$$U := u^\top \chi^\lambda_k(\tilde{X}_k).$$

Because $\left\|\chi^\lambda_k(x)\right\| \le G$, we have $|U| \le G$. The definition of the continuation yields

$$u^\top \tilde{\Phi}^\lambda_{k\to r}(y) = \mathbb{E}_\lambda[u^\top \chi^\lambda_k(\tilde{X}_k) \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[u^\top \chi^\lambda_k(\tilde{X}_k)].$$

The generic centered forgetting estimate (201) therefore gives

$$\left| u^\top \tilde{\Phi}^\lambda_{k\to r}(y) \right| \le G\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}).$$

Taking the supremum over unit u proves (205).

For the spatial derivative, applying Lemma D.21 to the same scalar observable U gives

$$\left\| \nabla_y \, \mathbb{E}_\lambda[u^\top \chi^\lambda_k(\tilde{X}_k) \mid \tilde{X}_r = y] \right\| \le G\kappa_{k\to r}.$$

The scalar continuation differs from this conditional expectation only by the constant $\mathbb{E}_\lambda[u^\top \chi^\lambda_k(\tilde{X}_k)]$. This constant cancels when two observations are compared. Therefore, for any y, y' we obtain

$$|u^\top(\tilde{\Phi}^\lambda_{k\to r}(y) - \tilde{\Phi}^\lambda_{k\to r}(y'))| \le G\kappa_{k\to r} \|y - y'\|.$$

Taking the supremum over unit u yields the bound for $\tilde{\Phi}$. Since Φ and $\tilde{\Phi}$ differ by the state-independent vector $T^\chi_k(\lambda)$, their Lipschitz constants coincide, proving (206). ■

Lemma D.24 (Derivative under the exponential tilt)**.** *Let $U^\lambda \in \mathbb{R}^\ell$ be differentiable in λ and integrable under the tilted law, with bounds permitting differentiation under the expectation. Let $\mathcal{C}$ be a sigma-field that does not depend on λ. Then*

$$\boxed{J_\lambda \, \mathbb{E}_\lambda[U^\lambda \mid \mathcal{C}] = \mathbb{E}_\lambda[J_\lambda U^\lambda \mid \mathcal{C}] - \mathrm{Cov}_\lambda(U^\lambda, g(\tilde{X}_0) \mid \mathcal{C}).} \tag{207}$$

The Jacobian and covariance on the right are $\ell \times m$ matrices. In our Gaussian conditioning applications, this identity can be evaluated at every observation y.

Proof. Fix a multiplier direction $v \in \mathbb{R}^m$ and set $\lambda_s = \lambda + sv$.

Write $w_s := e^{-\lambda_s^\top g(\tilde{X}_0)}$ for the unnormalized exponential weight. Conditional change of measure gives

$$\mathbb{E}_{\lambda_s}[U^{\lambda_s} \mid \mathcal{C}] = \frac{\mathbb{E}_{\tilde{P}}[U^{\lambda_s} w_s \mid \mathcal{C}]}{\mathbb{E}_{\tilde{P}}[w_s \mid \mathcal{C}]}.$$

We first differentiate the two quantities separately. The weight satisfies

$$\frac{\mathrm{d}w_s}{\mathrm{d}s} = -\big(v^\top g(\tilde{X}_0)\big) w_s, \qquad \frac{\mathrm{d}U^{\lambda_s}}{\mathrm{d}s} = J_\lambda U^{\lambda_s} v.$$

The product rule and differentiation under the fixed reference expectation therefore give

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s} \mathbb{E}_{\tilde{P}}[U^{\lambda_s} w_s \mid \mathcal{C}] &= \mathbb{E}_{\tilde{P}}\Big[\big(J_\lambda U^{\lambda_s} v - U^{\lambda_s} v^\top g(\tilde{X}_0)\big) w_s \mid \mathcal{C}\Big], \\ \frac{\mathrm{d}}{\mathrm{d}s} \mathbb{E}_{\tilde{P}}[w_s \mid \mathcal{C}] &= -\,\mathbb{E}_{\tilde{P}}[(v^\top g(\tilde{X}_0)) w_s \mid \mathcal{C}]. \end{aligned}$$

These interchanges are justified by the fact that g is bounded, and w_s together with its derivative is bounded locally uniformly in s. All the quantities inside these two expectations are therefore dominated by integrable bounds under $\tilde{P}$.

Appying the quotient yields

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s} \mathbb{E}_{\lambda_s}[U^{\lambda_s} \mid \mathcal{C}] = {} & \frac{\mathbb{E}_{\tilde{P}}[(J_\lambda U^{\lambda_s} v) w_s \mid \mathcal{C}]}{\mathbb{E}_{\tilde{P}}[w_s \mid \mathcal{C}]} \\ & - \frac{\mathbb{E}_{\tilde{P}}[U^{\lambda_s} (v^\top g(\tilde{X}_0)) w_s \mid \mathcal{C}]}{\mathbb{E}_{\tilde{P}}[w_s \mid \mathcal{C}]} \\ & + \frac{\mathbb{E}_{\tilde{P}}[U^{\lambda_s} w_s \mid \mathcal{C}]}{\mathbb{E}_{\tilde{P}}[w_s \mid \mathcal{C}]} \frac{\mathbb{E}_{\tilde{P}}[(v^\top g(\tilde{X}_0)) w_s \mid \mathcal{C}]}{\mathbb{E}_{\tilde{P}}[w_s \mid \mathcal{C}]}. \end{aligned}$$

Each normalized ratio is a conditional expectation under $\tilde{P}^{\lambda_s}$. Rewriting them under that law gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s} \mathbb{E}_{\lambda_s}[U^{\lambda_s} \mid \mathcal{C}] = {} & \mathbb{E}_{\lambda_s}[J_\lambda U^{\lambda_s} v \mid \mathcal{C}] \\ & - \mathbb{E}_{\lambda_s}[U^{\lambda_s} v^\top g(\tilde{X}_0) \mid \mathcal{C}] \\ & + \mathbb{E}_{\lambda_s}[U^{\lambda_s} \mid \mathcal{C}] \, \mathbb{E}_{\lambda_s}[v^\top g(\tilde{X}_0) \mid \mathcal{C}]. \end{aligned}$$

By the definition of covariance, the second and third lines combine into a negative covariance between a vector and a scalar. Hence

$$\frac{\mathrm{d}}{\mathrm{d}s} \mathbb{E}_{\lambda_s}[U^{\lambda_s} \mid \mathcal{C}] = \mathbb{E}_{\lambda_s}[J_\lambda U^{\lambda_s} v \mid \mathcal{C}] - \mathrm{Cov}_{\lambda_s}(U^{\lambda_s}, v^\top g(\tilde{X}_0) \mid \mathcal{C}).$$

At $s = 0$, the left side is the action of the multiplier Jacobian on v. Moreover,

$$\mathrm{Cov}_\lambda(U^\lambda, v^\top g(\tilde{X}_0) \mid \mathcal{C}) = \mathrm{Cov}_\lambda(U^\lambda, g(\tilde{X}_0) \mid \mathcal{C}) v.$$

Thus the identity specifies the action of the two $\ell \times m$ matrices in (207) on every $v \in \mathbb{R}^m$, proving their equality.

We apply the lemma to $U^\lambda = \chi_k^\lambda(\tilde{X}_k)$. Its two bounds are $\|U^\lambda\| \le G$ and $\|J_\lambda U^\lambda\|_{\mathrm{op}} \le L_\chi$, the latter supplied by Lemma D.2. First condition on $\mathcal{C} = \sigma(\tilde{X}_r)$; then apply the same identity with the trivial sigma-field $\mathcal{C} = \{\varnothing, \Omega\}$. Because the centered continuation is the conditional mean minus the unconditional mean at the *same* multiplier, subtracting these two derivative identities gives

$$\begin{aligned} J_\lambda \tilde{\Phi}_{k\to r}^\lambda(y) = \; & \mathbb{E}_\lambda[J_\lambda \chi_k^\lambda(\tilde{X}_k) \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[J_\lambda \chi_k^\lambda(\tilde{X}_k)] \\ & - \mathrm{Cov}_\lambda(\chi_k^\lambda(\tilde{X}_k), g(\tilde{X}_0) \mid \tilde{X}_r = y) + \mathrm{Cov}_\lambda(\chi_k^\lambda(\tilde{X}_k), g(\tilde{X}_0)). \end{aligned} \tag{208}$$

■

Proposition D.25 (Multiplier-derivative forgetting)**.** *For every $r > k$,*

$$\boxed{\left\|J_\lambda \tilde{\Phi}_{k\to r}^\lambda(y)\right\|_{\mathrm{op}} \le C_{\mathrm{der}}^\chi \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}), \qquad C_{\mathrm{der}}^\chi = L_\chi + 3G^2.} \tag{209}$$

Consequently, for all $\lambda, \lambda' \in \Lambda$,

$$\boxed{\left\|\tilde{\Phi}_{k\to r}^\lambda(y) - \tilde{\Phi}_{k\to r}^{\lambda'}(y)\right\| \le C_{\mathrm{der}}^\chi \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}})\,\|\lambda - \lambda'\|\,.} \tag{210}$$

Proof. Fix $1 \le k < r \le T$, a state $y \in \mathbb{R}^d$, and a multiplier $\lambda \in \Lambda$. All expectations in the following argument are under the fixed ideal law $\tilde{P}^\lambda$. We first bound the derivative at this multiplier and state.

The centered continuation takes values in $\mathbb{R}^m$ and its parameter also belongs to $\mathbb{R}^m$, so $J_\lambda \tilde{\Phi}_{k\to r}^\lambda(y)$ is an $m \times m$ matrix. A vector $v \in \mathbb{R}^m$ specifies an input perturbation of the multiplier, and multiplying the resulting output by $u^\top$ produces a scalar. The operator-norm identity established in (29) is here

$$\left\|J_\lambda \tilde{\Phi}_{k\to r}^\lambda(y)\right\|_{\mathrm{op}} = \sup_{\substack{u,v\in\mathbb{R}^m \\ \|u\|=\|v\|=1}} \left|u^\top J_\lambda \tilde{\Phi}_{k\to r}^\lambda(y) v\right|. \tag{211}$$

To prove a bound on this supremum, we will prove the same scalar bound for every such pair u, v.

Scalar reduction. For arbitrary unit directions $u, v \in \mathbb{R}^m$, define

$$F := u^\top \chi_k^\lambda(\tilde{X}_k), \qquad H := v^\top g(\tilde{X}_0), \qquad \dot{F} := u^\top J_\lambda \chi_k^\lambda(\tilde{X}_k) v. \tag{212}$$

The dot is a local abbreviation for the multiplier derivative in direction v:

$$\dot{F} = \frac{\mathrm{d}}{\mathrm{d}s} u^\top \chi_k^{\lambda+sv}(\tilde{X}_k)\Big|_{s=0}.$$

The bounds on F and H follow from Cauchy–Schwarz and the global constraint bound

$$|F| \le \|u\| \left\|\chi_k^\lambda(\tilde{X}_k)\right\| \le G, \qquad |H| \le \|v\| \left\|g(\tilde{X}_0)\right\| \le G.$$

Lemma D.2 gives

$$|\dot{F}| = |u^\top J_\lambda \chi_k^\lambda(\tilde{X}_k) v| \le \|u\| \left\|J_\lambda \chi_k^\lambda(\tilde{X}_k)\right\|_{\mathrm{op}} \|v\| \le L_\chi, \qquad L_\chi = L_g R_X G. \tag{213}$$

Consequently,

$$|F| \le G, \qquad |H| \le G, \qquad |\dot{F}| \le L_\chi. \tag{214}$$

F and $\dot{F}$ are functions of $\tilde{X}_k$, while H is a function of $\tilde{X}_0$. Hence $\dot{F}$, FH, F, and H are all measurable with respect to $\sigma(\tilde{X}_0, \ldots, \tilde{X}_k)$. They are bounded, so Lemma D.22 applies to each of them.

Derivative decomposition. Projecting the centered identity (208) onto the output direction u and applying the resulting Jacobian to the input direction v gives directly

$$u^\top J_\lambda \tilde{\Phi}^\lambda_{k\to r}(y)v = \mathbb{E}_\lambda[\dot{F} \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[\dot{F}] - \mathrm{Cov}_\lambda(F, H \mid \tilde{X}_r = y) + \mathrm{Cov}_\lambda(F, H).$$

We now expand both scalar covariances:

$$\mathrm{Cov}_\lambda(F, H \mid \tilde{X}_r = y) = \mathbb{E}_\lambda[FH \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[F \mid \tilde{X}_r = y]\,\mathbb{E}_\lambda[H \mid \tilde{X}_r = y],$$
$$\mathrm{Cov}_\lambda(F, H) = \mathbb{E}_\lambda[FH] - \mathbb{E}_\lambda[F]\,\mathbb{E}_\lambda[H].$$

Substituting these two expansions and grouping conditional-minus-unconditional quantities gives

$$\begin{aligned} u^\top J_\lambda \tilde{\Phi}^\lambda_{k\to r}(y)v = &\underbrace{\mathbb{E}_\lambda[\dot{F} \mid \tilde{X}_r = y] - \mathbb{E}_\lambda \dot{F}}_{(I)} \\ &- \underbrace{\big(\mathbb{E}_\lambda[FH \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[FH]\big)}_{(II)} \\ &+ \underbrace{\big(\mathbb{E}_\lambda[F \mid \tilde{X}_r = y]\,\mathbb{E}_\lambda[H \mid \tilde{X}_r = y] - \mathbb{E}_\lambda F\ \mathbb{E}_\lambda H\big)}_{(III)}. \end{aligned} \tag{215}$$

Forgetting bounds. Recall the exact statement of Lemma D.22: for each bounded scalar observable U measurable up to level k,

$$\big|\,\mathbb{E}_\lambda[U \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[U]\big| \le \|U\|_\infty\, \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}).$$

For term (I), choose $U = \dot{F}$. The bound (213) verifies $\left\|\dot{F}\right\|_\infty \le L_\chi$, so

$$\begin{aligned} |(I)| &= \big|\,\mathbb{E}_\lambda[\dot{F} \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[\dot{F}]\big| \\ &\le \left\|\dot{F}\right\|_\infty \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}) \\ &\le L_\chi \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}). \end{aligned}$$

This is the contribution controlled by Lemma D.2: the bound on the explicit derivative of χ^λ_k becomes a forgetting bound on its centered conditional expectation.

For term (II), choose $U = FH$. The product is a bounded past observable with $\|FH\|_\infty \le G^2$, hence

$$\begin{aligned} |(II)| &= \big|\,\mathbb{E}_\lambda[FH \mid \tilde{X}_r = y] - \mathbb{E}_\lambda[FH]\big| \\ &\le \|FH\|_\infty\, \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}) \\ &\le G^2 \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}). \end{aligned}$$

This term comes from the covariance correction in (208); the bounds on the constraint values control it.

Product term. Term (III) is a difference of two products of expectations. To expose centered expectations, define at the fixed y

$$m_F := \mathbb{E}_\lambda[F \mid \tilde{X}_r = y], \quad \mu_F := \mathbb{E}_\lambda[F], \qquad m_H := \mathbb{E}_\lambda[H \mid \tilde{X}_r = y], \quad \mu_H := \mathbb{E}_\lambda[H].$$

Inserting and subtracting $\mu_F m_H$ gives

$$m_F m_H - \mu_F \mu_H = (m_F - \mu_F) m_H + \mu_F (m_H - \mu_H).$$

By boundedness of F and H, conditional and ordinary expectation give

$$|m_H| \le \mathbb{E}_\lambda[|H| \mid \tilde{X}_r = y] \le G, \qquad |\mu_F| \le \mathbb{E}_\lambda[|F|] \le G.$$

using the generic forgetting lemma (Lemma D.22) separately with $U = F$ and $U = H$ yields

$$|m_F - \mu_F| \le G\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}), \qquad |m_H - \mu_H| \le G\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}).$$

The triangle inequality for the product decomposition therefore gives

$$\begin{aligned} |(III)| &\le |m_F - \mu_F|\,|m_H| + |\mu_F|\,|m_H - \mu_H| \\ &\le G^2 \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}) + G^2 \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}) \\ &= 2G^2 \kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}). \end{aligned}$$

Operator-norm bound Combining the three estimates with (215) yields

$$\begin{aligned}|u^\top J_\lambda \tilde{\Phi}^\lambda_{k\to r}(y)v| &\le |(I)| + |(II)| + |(III)| \\ &\le (L_\chi + G^2 + 2G^2)\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}) \\ &= C^\chi_{\mathrm{der}}\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}}).\end{aligned}$$

The coefficient is the same for every unit u, v. Taking their supremum as in (211) proves (209).

Lipschitz consequence. Let $\lambda, \lambda' \in \Lambda$ and keep k, r, y fixed. Convexity of Λ ensures that the entire segment

$$\lambda_s := \lambda' + s(\lambda - \lambda'), \qquad 0 \le s \le 1,$$

lies in Λ. Using the chain rule for this one-dimensional parameterization we obtain

$$\frac{\mathrm{d}}{\mathrm{d}s}\tilde{\Phi}^{\lambda_s}_{k\to r}(y) = J_\lambda \tilde{\Phi}^{\lambda_s}_{k\to r}(y)(\lambda - \lambda').$$

Integrating from $s = 0$ to $s = 1$ gives the exact difference of the endpoint values. Applying the norm inequality for a vector-valued integral and then the operator norm gives

$$\begin{aligned}\left\|\tilde{\Phi}^\lambda_{k\to r}(y) - \tilde{\Phi}^{\lambda'}_{k\to r}(y)\right\| &= \left\|\int_0^1 J_\lambda \tilde{\Phi}^{\lambda_s}_{k\to r}(y)(\lambda - \lambda')\,\mathrm{d}s\right\| \\ &\le \int_0^1 \left\|J_\lambda \tilde{\Phi}^{\lambda_s}_{k\to r}(y)\right\|_{\mathrm{op}} \|\lambda - \lambda'\|\,\mathrm{d}s \\ &\le C^\chi_{\mathrm{der}}\kappa_{k\to r}(\|y\| + M_{\mathrm{ref}})\,\|\lambda - \lambda'\|\,.\end{aligned}$$

The last bound is uniform along the segment, by (209). This proves (210). ■

D.6 Schedule summability and geometric forgetting

D.6.1 Uniformly positive VP schedules imply row and column geometric forgetting

Proposition D.26 (Geometric bound for a positive per-step noise floor)**.** *Suppose*

$$\beta_j \geq \beta_\star > 0 \tag{216}$$

for all levels $j = 1, \dots, T$, with $\beta_\star$ independent of the horizon. Let $\theta := 1 - \beta_\star < 1$. Then, for every $1 \leq k < r \leq T$,

$$\boxed{\kappa_{k\to r} \leq C_\star \theta^{(r-k)/2}, \qquad C_\star := \frac{\sqrt{d}}{\sqrt{\beta_\star}} + \frac{R_X}{\beta_\star}.} \tag{217}$$

Consequently,

$$\boxed{C_{\mathrm{row}} \leq C_\star \frac{\sqrt{\theta}}{1-\sqrt{\theta}}, \qquad C_{\mathrm{col}} \leq C_\star \frac{\sqrt{\theta}}{1-\sqrt{\theta}}.} \tag{218}$$

Both bounds are uniform in T.

Proof. For $n = r - k \geq 1$,

$$a_{k\to r}^2 = \prod_{j=k+1}^{r} (1-\beta_j) \leq \theta^n, \qquad a_{k\to r} \leq \theta^{n/2}.$$

Moreover,

$$b_{k\to r}^2 = 1 - a_{k\to r}^2 \geq \beta_\star, \qquad \sigma_r^2 \geq \beta_\star, \qquad \sigma_k \leq \sigma_r, \qquad \rho_k \leq 1.$$

Recall

$$\kappa_{k\to r} = \frac{a_{k\to r}}{b_{k\to r}^2} \sqrt{V_{k\to r}},$$

with

$$V_{k\to r} = d \frac{\sigma_k^2 b_{k\to r}^2}{\sigma_r^2} + R_X^2 \frac{\rho_k^2 b_{k\to r}^4}{\sigma_r^4}.$$

Using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ gives

$$\begin{aligned}
\kappa_{k\to r} &\leq \frac{a_{k\to r}}{b_{k\to r}^2} \left[\sqrt{d} \frac{\sigma_k b_{k\to r}}{\sigma_r} + R_X \frac{\rho_k b_{k\to r}^2}{\sigma_r^2} \right] \\
&= a_{k\to r} \left[\frac{\sqrt{d}\,\sigma_k}{b_{k\to r}\sigma_r} + \frac{R_X \rho_k}{\sigma_r^2} \right] \\
&\leq a_{k\to r} \left[\frac{\sqrt{d}}{\sqrt{\beta_\star}} + \frac{R_X}{\beta_\star} \right] \leq C_\star \theta^{(r-k)/2},
\end{aligned}$$

which proves (217). Moreover, by construction, $\theta < 1$.

For the row sum, fix k and set $n = r - k$:

$$\sum_{r=k+1}^{T} \kappa_{k\to r} \leq C_\star \sum_{n=1}^{\infty} \theta^{n/2} = C_\star \frac{\sqrt{\theta}}{1-\sqrt{\theta}}.$$

For the column sum, fix r and set $n = r - k$:

$$\sum_{k=1}^{r-1} \kappa_{k\to r} \leq C_\star \sum_{n=1}^{r-1} \theta^{n/2} \leq C_\star \frac{\sqrt{\theta}}{1-\sqrt{\theta}}.$$

Taking the corresponding suprema proves (218). ∎

E EXPERIMENTS

E.1 IMPLEMENTATION DETAILS

Implementing Algorithm 1 requires choosing (1) the pretrained model and its sampler, (2) the constraint functions g_i, and (3) the dual update rule and its step sizes. The inverse-problem experiments use Algorithm 1 with the projection onto $\mathbb{R}^m_+$ described in Section 5. The semantic-photometric control experiments change the sampler and the dual update as described below.

Inverse problems. We use the pretrained pixel-space DDM of Chung et al. (2023) with $T = 1{,}000$ diffusion steps for all tasks. In all PD-CDS runs, the denoised estimate $\hat{x}_{0|t}$ is clipped to the model's pixel range $[-1, 1]$. Given a measurement $y \in \mathbb{R}^n$, we impose

$$g_{\text{inv}}(x) = \frac{1}{c_{\text{inv}}} \frac{\|\mathcal{A}(x) - y\|_1}{N} - 1 \le 0, \qquad c_{\text{inv}} = 0.01,$$

where N is the number of observed entries (e.g., unmasked pixels for inpainting). When a TV constraint is imposed, we use

$$g_{\text{TV}}(x) = \frac{\text{TV}(x)}{c_{\text{TV}}} - 1 \le 0, \qquad c_{\text{TV}} = 0.1,$$

with the following discrete form of total variation (Rudin et al., 1992):

$$\text{TV}(x) = \frac{1}{N_{\text{pix}}} \sum_{i,j} \sqrt{(x_{i+1,j} - x_{i,j})^2 + (x_{i,j+1} - x_{i,j})^2},$$

where N_{pix} is the number of pixels.

We use the dual step size $\eta_t = \eta\sqrt{\bar{\alpha}_t}$, with $\bar{\alpha}_t$ defined in Appendix B, which gave the most consistent results. We set $\eta = 1$ for all tasks except desolarization, for which $\eta = 2$ was needed to reach feasible solutions. This schedule increases along sampling and limits dual updates at early, high-noise steps, when $\hat{x}_{0|t}$ is less reliable.

Semantic-photometric control. All methods use the same unconditional CelebA-HQ model of Ho et al. (2020). PD-CDS uses 250 respaced DDIM steps. Guidance enters through the predicted noise, $\tilde{\epsilon} = \epsilon_\theta(x_t, t) - \sigma_t s_t^\lambda$, with s_t^λ as in Algorithm 1, and the dual variables are updated once per sampling step.

Semantic guidance uses a pretrained attribute classifier with separate heads for `Smiling` and `Not-Young`, released as `attribute_classifier.pt` in the `pymlex/celeba-gan-xai` repository (pymlex, 2026). For a target attribute y, we require

$$\ell_{\text{sem}}(x) = -\log p_{\text{Pymlex}}(y \mid x) \le c_{\text{sem}}, \qquad c_{\text{sem}} = -\log p^\star,$$

where $p^\star = 0.97$ for `Smiling` and $p^\star = 0.99$ for `Not-Young`. For a photometric quantity M with target $M^\star$, we require

$$\ell_{\text{phot}}(x) = |M(x) - M^\star| \le c_{\text{phot}}, \qquad c_{\text{phot}} = 0.01.$$

As for inverse problems, the dual update uses the normalized constraints $g_{\text{sem}} = \ell_{\text{sem}}/c_{\text{sem}} - 1$ and $g_{\text{phot}} = \ell_{\text{phot}}/c_{\text{phot}} - 1$, which define the same constraints and only rescale the multipliers. The constraint values reported in Appendix E.5 are $\ell - c$.

Photometric quantities are computed after rescaling the image from $[-1, 1]$ to $[0, 1]$. Brightness is the mean pixel value over all channels,

$$b(x) = \frac{1}{3HW} \sum_{c,h,w} x_{c,h,w}, \qquad M^\star \in \{0.65, 0.25\}.$$

Contrast is the RMS contrast of Rec. 709 luminance,

$$C(x) = \sqrt{\frac{1}{HW} \sum_{h,w} (Y_{h,w} - \bar{Y})^2}, \qquad Y_{h,w} = 0.2126 x_{\text{R},h,w} + 0.7152 x_{\text{G},h,w} + 0.0722 x_{\text{B},h,w},$$

where $\bar{Y}$ is the mean luminance and $M^\star \in \{0.28, 0.18\}$.

For the `Smiling` and `Smiling` + brightness tasks, the dual update is projected gradient ascent with step size $\eta = 0.35$. For `Not-Young` + contrast, it is Adam with learning rate 750, followed by projection onto $\mathbb{R}^m_+$. These settings were chosen heuristically to maintain guidance-classifier accuracy without degrading KID too much; within each task, they are shared by the high and low photometric targets (Table 3).

Table 3: Hyperparameters used for semantic-photometric control on CelebA-HQ. PD-CDS uses 250 DDIM steps; the baselines use 100. Each configuration is used for both the high and low photometric targets.

Method	`Smiling`	`Smiling` + brightness	`Not-Young` + contrast
PD-CDS (update, step)	proj. ascent, $\eta = 0.35$	proj. ascent, $\eta = 0.35$	Adam, LR $= 750$
TFG (ρ, μ, σ)	8, 4, 0.01	6, 6, 0.01	12, 8, 0.01
UGD (strength)	12	48	24
LGD (strength)	0.375	1.5	10
FreeDoM (strength)	3	12	20

LGD uses one Monte Carlo sample in all cases.

Each method generates 200 samples per task, and KID is computed against 1,000 CelebA-HQ validation images. Semantic accuracy uses the attribute classifiers of the GAN-Ensembling project (Chai et al., 2021), which are used neither for guidance nor for hyperparameter selection. A sample counts toward joint accuracy if this evaluator predicts the target attribute and $|M(x) - M^\star| \leq 0.01$. Semantic violations use the guidance classifier and threshold $p^\star$.

E.2 Baseline implementation

All baselines are implemented in a unified framework following their reference code. They share the same checkpoint, forward operators, noiseless observations and paired random seeds.

Inverse problems. Where a reference configuration adds measurement noise, we disable it. Table 4 lists the settings shared across tasks and Table 5 those that vary per task. Except where noted below, we use the reference configurations.

Table 4: Baseline settings shared across all inverse-problem tasks.

Method	**Sampler**	**NFE**	**Init.**	**Fixed hyperparameters**
DAPS	200 annealing levels, EDM schedule ($\rho = 7$, σ: 100 $\to$ 0.1), 5 PF-ODE steps per level	1000	prior	$\tau = 0.01$; Langevin step $\gamma \to 0.01\,\gamma$
DiffPIR	100 steps, quadratic spacing	100	prior	5 inner gradient steps
DPS	1000 ancestral steps	1000	prior	–
ΠGDM	100 uniform deterministic DDIM steps	100	prior	guidance 1.0, $\sigma_y = 0$
RED-diff	1000 timesteps, descending, one Adam step each	1000	per task	lr 0.1, $\beta = (0.9, 0.99)$, $\lambda = 0.25$

DAPS. We use prior initialization, the squared ℓ_2 measurement loss, 200 annealing levels, five ODE steps per level, $\tau = 0.01$, $\sigma_{\min} = 0.1$, $\sigma_{\max} = 100$, a minimum learning-rate ratio of 0.01, and timestep power 7. The learning rates for super-resolution, inpainting, Gaussian blur, motion blur, nonlinear blur, and desolarization are 10^{-4}, 5×10^{-5}, 10^{-4}, 5×10^{-5}, 10^{-5}, and 10^{-4}, respectively. We use 100 Langevin steps for all tasks except nonlinear blur, for which we use 200.

DiffPIR. We use 100 quadratically spaced DDPM steps, prior initialization, the squared ℓ_2 measurement loss, and the automatic data-consistency solver with five proximal iterations and proximal step size 1.0. For super-resolution, inpainting, Gaussian blur, motion blur, nonlinear blur, and

Table 5: Task-specific baseline hyperparameters. A dash indicates the method is not applicable.

Method	**Parameter**	Gauss.	Inpaint.	Motion	SR 4×	NL blur	Desol.
DAPS	γ	10^{-4}	$5{\cdot}10^{-5}$	$5{\cdot}10^{-5}$	10^{-4}	10^{-5}	10^{-4}
	Langevin steps	100	100	100	100	200	100
DiffPIR	λ	12.0	6.0	7.0	6.0	7.0	7.0
	ζ	0.4	0.5	0.9	0.3	0.3	0.3
	σ_n	10^{-3}	10^{-3}	10^{-3}	10^{-3}	0.05	10^{-3}
	Data step	grad.	closed form	grad.	grad.	grad.	grad.
DPS	c	0.3	0.5	0.3	0.3	0.3	0.5
ΠGDM	applicable	✓	✓	✓	✓	–	–
RED-diff	σ_v	1	0.866	1	0.25	1	1
	Init.	$\mathcal{A}^\dagger\mathbf{y}$	$\mathcal{A}^\dagger\mathbf{y}$	$\mathcal{A}^\dagger\mathbf{y}$	$\mathcal{A}^\dagger\mathbf{y}$	prior	$\mathbf{y}$

desolarization, respectively, we use

$$\lambda = (6, 6, 12, 7, 7, 7), \qquad \zeta = (0.3, 0.5, 0.4, 0.9, 0.3, 0.3).$$

The assumed noise level is $\sigma_n = 0.001$, except for nonlinear blur, where $\sigma_n = 0.05$. For inpainting, we additionally apply a final projection onto the observed pixels.

DPS. The guidance scale is normalized by the residual, $\xi_i = c/\|\mathbf{y} - \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_i))\|$, with c taken from the released configurations where available: $c = 0.3$ for super-resolution, Gaussian blur, motion blur, and nonlinear blur, and $c = 0.5$ for inpainting and desolarization. For nonlinear blur, the reference operator draws a new random kernel at every call, including inside the guidance gradient. We instead fix one kernel per problem, shared by all methods. On this task, DPS fits the observation on $35/100$ images (28.40 ± 2.14 dB PSNR, median observation MAE 0.015) and fails on the remaining 65 (7.36 ± 1.94 dB, median observation MAE 0.36). No image falls between 12.6 and 21.7 dB, and the failures are spread across all batches. This bimodality accounts for its low mean and large standard deviation in Table 6.

Π*GDM.* We use a 100-step DDIM sampler with uniformly spaced timesteps. The guidance scale is 1.0, the assumed measurement-noise level is $\sigma_y = 0$, and gradients are computed through the full graph without clipping. ΠGDM requires a pseudoinverse and is therefore only run on the linear tasks.

RED-diff. We use a learning rate of 0.1, regularization weight $\lambda = 0.25$, and the squared ℓ_2 measurement loss. The method is initialized from the pseudoinverse solution $\mathcal{A}^\dagger\mathbf{y}$ on the linear tasks, from the observation $\mathbf{y}$ for desolarization, and from the prior for nonlinear blur. We set σ_v to 0.25 for super-resolution, 0.866 for inpainting, and 1.0 for all remaining tasks.

Semantic-photometric control. TFG, UGD, LGD and FreeDoM use 100 DDIM steps and follow the implementations of Ye et al. (2024). Their total guidance loss is the sum of all losses. For a semantic and a photometric requirement, for instance, it reads

$$\mathcal{L}_{\text{guide}}(x) = -\log p_{\text{Pymlex}}(y \mid x) + |M(x) - M^\star|.$$

The baselines therefore see neither the threshold $p^\star$ nor the tolerance c_{phot}, and their violations are computed after sampling, as diagnostics. Their hyperparameters were selected with the same criterion as for PD-CDS (Table 3).

E.3 Extended results

In this section we report the complete results for the inverse problems and the semantic-photometric control tasks, including full metric tables and representative images.

Inverse problems. Table 6 reports the complete results for inverse problems on the first 100 FFHQ validation images, adding PSNR, SSIM and standard deviations. PD-CDS is competitive on the linear tasks and, together with DAPS, is the strongest method on the nonlinear ones. Adding the TV constraint lowers FID on every task except desolarization, and improves all four metrics on inpainting, motion deblurring and nonlinear deblurring. On Gaussian deblurring and super-resolution, the other metrics change little. On desolarization, the TV run does not reach the observation constraint (Appendix E.5), and all metrics degrade.

Semantic-photometric control. Table 7 reports the complete results for the single `Smiling` requirement, including the mean semantic violation, and Table 8 those for the two-constraint tasks. PD-CDS satisfies the photometric constraint on every sample and obtains the highest joint accuracy and lowest KID on all four tasks, while its semantic accuracy is 2 to 13 points below the best baseline.

E.4 Runtime and memory

We report wall-clock sampling time and peak GPU memory for inverse problems (Table 9) and for semantic-photometric control on CelebA-HQ (Table 10), using the configurations of the main experiments. All measurements were run on a machine with an Intel Core i9-10900K CPU (3.70 GHz), two NVIDIA GeForce RTX 2080 Ti GPUs and 64 GiB of RAM. Each run uses a single GPU and batch size 1, and starts with one warm-up sample. We then time 10 samples individually, with GPU synchronization. For inverse problems, we report the mean time for each task and the peak memory over all tasks. For semantic-photometric control, we report the mean time and peak memory on the `Not-Young` + high contrast task.

Peak allocated memory is the maximum memory occupied by tensors (`torch.cuda.max_memory_allocated`). Peak reserved memory also includes memory held by the PyTorch caching allocator (`torch.cuda.max_memory_reserved`). NFE is the number of forward evaluations of the denoiser per sample. It does not count backward passes through the denoiser, which DPS and ΠGDM perform at every step for inverse problems, and all four baselines perform at every step for semantic-photometric control.

E.5 Dual dynamics and constraint feasibility

We report the dual variables along sampling (step 0 is $t = T$) and the constraint values of the final samples; for the baselines, these are post-hoc diagnostics.

For inverse problems, Table 11 and Figure 3 report feasibility at $t = 0$. On the linear tasks, the mean MAE of PD-CDS is 0.0018 - 0.0064, below the tolerance $c_{\text{inv}} = 0.01$, whereas the constraint would be active at the solution of (PI). It is feasible on 65% of the images for nonlinear deblurring, against at most 29% for the baselines, and on 77% for desolarization, against 100% for RED-diff, DiffPIR and DAPS. With TV, 21% and 24% of the images remain TV-infeasible on Gaussian deblurring and super-resolution (Figure 4), and desolarization meets the observation constraint on no image, consistent with its drop in quality in Table 6.

For semantic-photometric control, a multiplier decreases once its constraint is satisfied. With `Smiling` alone, λ_{sem} reaches zero after ≈ 80 of the 250 steps, and the rest of sampling is unguided (Figure 5). The expectation constraint holds, whereas the baselines concentrate at $p \approx 1$, consistent with their higher accuracy and worse KID (Table 7). In contrast, every sample ends at $|M(x) - M^\star| \approx 10^{-4}$, far inside $c_{\text{phot}} = 0.01$ (Figure 6), because the photometric multipliers grow during almost the whole trajectory (Figure 7). Finally, under `Not-Young` + high contrast, nearly every sample meets the guidance-classifier constraint, yet the independent evaluator approves only 86% of them, since a constraint set by one classifier enforces the attribute only insofar as the two classifiers are in agreement.

The dual trajectories of the inverse problems explain the observation margin and the late TV violations: λ_{inv} grows during almost the whole trajectory and its constraint value becomes negative only in the last steps (Figure 8); since $g_{\text{inv}} \geq -1$, it then decreases by at most η_t per step and cannot relax. The TV constraint value, in contrast, often becomes positive again in the last ≈ 100 steps, too late to be corrected (Figure 9).

Table 6: Results on linear and nonlinear inverse problems using the first 100 FFHQ validation images at 256×256. Mean $\pm$ standard deviation over images for PSNR, SSIM and LPIPS; FID is computed on the same 100 images. Best results within each task are shown in bold and second-best results are underlined.

Task	Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	FID ↓
Linear inverse problems					
Gaussian deblurring	PD-CDS	27.40±2.38	0.742±0.098	0.145±0.133	69.38
	PD-CDS + TV	27.17±2.83	0.724±0.132	0.160±0.178	<u>61.10</u>
	DAPS	<u>29.04</u>±1.74	0.786±0.035	0.146±0.047	77.73
	DiffPIR	28.46±1.93	<u>0.814</u>±0.044	<u>0.141</u>±0.049	**60.00**
	DPS	26.24±1.88	0.726±0.060	**0.130**±0.041	78.41
	ΠGDM	27.10±2.01	0.745±0.099	0.159±0.059	79.03
	RED-diff	**29.49**±1.97	**0.839**±0.040	0.194±0.057	73.17
Box inpainting (128×128)	PD-CDS	21.29±2.51	0.830±0.026	0.102±0.032	50.65
	PD-CDS + TV	21.63±2.62	0.842±0.025	0.086±0.029	38.51
	DAPS	**24.75**±2.80	**0.874**±0.023	**0.069**±0.025	32.56
	DiffPIR	22.97±2.94	<u>0.873</u>±0.028	0.076±0.027	**28.45**
	DPS	<u>24.17</u>±2.42	0.818±0.035	0.105±0.032	62.52
	ΠGDM	24.04±2.71	0.871±0.027	<u>0.072</u>±0.025	<u>30.66</u>
	RED-diff	23.84±2.20	0.859±0.022	0.089±0.026	43.82
	Projection	21.22±2.82	0.838±0.029	0.097±0.037	40.62
Motion deblurring	PD-CDS	29.12±1.56	0.825±0.033	<u>0.075</u>±0.018	<u>56.41</u>
	PD-CDS + TV	<u>29.39</u>±1.52	0.832±0.035	**0.061**±0.017	**46.12**
	DAPS	29.04±1.51	0.803±0.035	0.128±0.036	75.54
	DiffPIR	28.96±1.80	<u>0.846</u>±0.038	0.168±0.043	80.29
	DPS	24.16±1.95	0.676±0.070	0.169±0.047	85.67
	ΠGDM	24.27±1.84	0.690±0.081	0.196±0.055	91.53
	RED-diff	**30.37**±1.65	**0.868**±0.029	0.117±0.033	63.94
Super-resolution (bicubic, $\times 4$)	PD-CDS	27.79±2.30	0.765±0.075	<u>0.122</u>±0.104	62.54
	PD-CDS + TV	27.90±2.50	0.763±0.097	0.123±0.137	<u>53.23</u>
	DAPS	<u>29.20</u>±1.70	0.795±0.032	0.131±0.042	75.98
	DiffPIR	28.46±1.84	0.820±0.042	0.146±0.047	61.81
	DPS	24.57±1.93	0.682±0.072	0.190±0.058	93.76
	ΠGDM	29.16±2.15	<u>0.844</u>±0.042	**0.077**±0.028	**43.65**
	RED-diff	**30.53**±2.07	**0.871**±0.034	0.154±0.044	64.54
Nonlinear inverse problems					
Nonlinear deblurring	PD-CDS	32.79±2.82	0.878±0.077	0.150±0.165	<u>35.38</u>
	PD-CDS + TV	**34.12**±3.01	**0.905**±0.068	<u>0.107</u>±0.156	**29.65**
	DAPS	<u>32.96</u>±1.62	<u>0.900</u>±0.020	**0.064**±0.019	42.60
	DiffPIR	26.04±3.14	0.645±0.134	0.386±0.161	96.61
	DPS	14.72±10.23	0.399±0.340	0.626±0.384	284.71
	RED-diff	22.63±5.39	0.668±0.184	0.339±0.189	110.64
Desolarization	PD-CDS	<u>11.51</u>±5.55	<u>0.626</u>±0.222	<u>0.295</u>±0.137	<u>114.02</u>
	PD-CDS + TV	10.03±3.46	0.458±0.142	0.438±0.108	179.56
	DAPS	**14.30**±6.78	**0.769**±0.148	**0.174**±0.100	**79.47**
	DiffPIR	10.76±4.44	0.604±0.190	0.348±0.138	139.13
	DPS	9.42±5.34	0.452±0.281	0.555±0.327	221.31
	RED-diff	10.60±3.86	0.553±0.181	0.342±0.093	157.17
	Projection	10.39±3.73	0.572±0.195	0.373±0.119	146.08

Table 7: CelebA-HQ generation with a single `Smiling` requirement. Violation is the mean positive semantic constraint violation under the guidance classifier (post-hoc for the baselines; see Table 8). Best results are shown in bold and second-best results are underlined.

Method	**Accuracy (%) ↑**	**Violation ↓**	**KID ↓**
PD-CDS	95.00	0.0068	$\mathbf{0.0315 \pm 0.0065}$
TFG	**100.0**	**0.0000**	0.0408 ± 0.0092
UGD	**100.0**	**0.0000**	0.0615 ± 0.0113
LGD	99.50	0.0011	0.0478 ± 0.0099
FreeDoM	**100.0**	**0.0000**	0.0639 ± 0.0114

Table 8: CelebA-HQ under heterogeneous semantic–photometric constraints. Joint accuracy requires semantic correctness and $|M(x) - M^\star| \leq 0.01$. Best results are shown in bold and second-best results are underlined.

Task	**Method**	**Sem. Acc. (%) ↑**	**Sem. Viol. ↓**	**Photo. MAE ↓**	**Photo. Viol. ↓**	**Joint Acc. (%) ↑**	**KID ↓**
Smile + brightness High (0.65)	PD-CDS	98.0	0.0017	**0.0001**	**0.0000**	**98.0**	$\mathbf{0.0312 \pm 0.0083}$
	TFG	99.0	**0.0000**	0.0162	0.0106	60.0	0.0601 ± 0.0112
	UGD	94.0	0.0011	0.0126	0.0069	62.5	0.0549 ± 0.0123
	LGD	**100.0**	**0.0000**	0.1139	0.1043	7.0	0.0744 ± 0.0130
	FreeDoM	93.5	**0.0000**	0.0105	0.0058	75.0	0.0529 ± 0.0120
Smile + brightness Low (0.25)	PD-CDS	95.5	0.0022	**0.0001**	**0.0000**	**95.5**	$\mathbf{0.0273 \pm 0.0049}$
	TFG	99.5	**0.0000**	0.0494	0.0426	45.5	0.0574 ± 0.0122
	UGD	95.5	0.0011	0.0296	0.0229	50.0	0.0527 ± 0.0101
	LGD	**100.0**	**0.0000**	0.1962	0.1864	2.0	0.0708 ± 0.0123
	FreeDoM	93.5	**0.0000**	0.0254	0.0200	70.0	0.0521 ± 0.0104
not-Young + contrast High (0.28)	PD-CDS	86.0	0.0004	**0.0001**	**0.0000**	**86.0**	$\mathbf{0.0444 \pm 0.0101}$
	TFG	93.5	**0.0000**	0.0120	0.0081	74.5	0.0586 ± 0.0137
	UGD	88.5	0.0001	0.0293	0.0210	23.5	0.0818 ± 0.0137
	LGD	**99.0**	0.0002	0.0370	0.0277	16.5	0.1050 ± 0.0170
	FreeDoM	94.5	0.0050	0.0279	0.0233	70.5	0.0620 ± 0.0127
not-Young + contrast Low (0.18)	PD-CDS	92.0	**0.0000**	**0.0001**	**0.0000**	**92.0**	$\mathbf{0.0502 \pm 0.0107}$
	TFG	98.5	**0.0000**	0.0096	0.0048	64.5	0.0642 ± 0.0146
	UGD	94.0	0.0001	0.0328	0.0238	20.0	0.0889 ± 0.0144
	LGD	**99.5**	0.0003	0.0623	0.0524	3.0	0.0988 ± 0.0165
	FreeDoM	95.5	0.0056	0.0180	0.0126	66.5	0.0640 ± 0.0138

Note. Sem. = semantic; Photo. = photometric; Acc. = accuracy; Viol. = mean positive constraint violation; MAE = mean absolute error. Semantic violations are evaluated using the guidance classifier and the corresponding PD-CDS threshold. For the baselines, they are post-hoc diagnostics, since these methods do not explicitly optimize the thresholded semantic constraint.

Table 9: Wall-clock sampling time per image and peak GPU memory over all tasks for inverse problems. For every method run on nonlinear deblurring, the memory peak is reached on that task.

		Time (s/img)						**Peak mem. (MiB)**	
Method	**NFE**	Gauss.	Inpaint.	Motion	SR 4×	NL blur	Desol.	Alloc.	Res.
PD-CDS	1000	48.3	45.5	48.5	45.5	64.8	45.6	1144	1446
PD-CDS + TV	1000	50.9	46.6	48.3	47.1	64.9	46.6	1135	1430
DAPS	1000	102.3	48.7	102.7	61.8	457.5	60.1	1109	1364
DiffPIR	100	5.8	4.4	5.7	4.8	9.7	4.8	1107	1366
DPS	1000	87.2	85.8	87.4	85.7	97.9	85.2	2656	2984
ΠGDM	100	8.7	8.6	8.7	8.5	–	–	2058	2392
RED-diff	1000	45.6	43.7	45.7	34.3	55.7	44.2	1120	1384
Projection	1000	–	43.2	–	–	–	43.5	672	870

Table 10: Sampling time and peak GPU memory for semantic-photometric control on CelebA-HQ (`Not-Young` + high contrast, mean over 10 samples).

Method	**NFE**	**Time (s/img)**	**Peak alloc. (MiB)**	**Peak reserved (MiB)**
PD-CDS	250	15.14	1298	1496
TFG	100	12.23	2605	2858
UGD	100	17.07	2607	2864
LGD	100	16.42	2607	2864
FreeDoM	100	16.33	2607	2864

Table 11: Feasibility at $t = 0$ on the 100 FFHQ images: percentage of images with $g_{\text{inv}} \leq 0$ and, in parentheses, the mean MAE in units of 10^{-2} (the expectation constraint holds if it is at most 1). The last row gives the TV constraint of PD-CDS + TV (percentage with $\text{TV} \leq 0.1$; mean TV in units of 10^{-1}).

Method	Inpainting	SR $\times 4$	Gaussian deblur	Motion deblur	Nonlinear deblur	Desolarization
PD-CDS	100 (0.64)	100 (0.18)	97 (0.28)	98 (0.36)	65 (1.63)	77 (0.87)
PD-CDS + TV	99 (0.69)	100 (0.17)	91 (0.38)	98 (0.38)	65 (1.56)	0 (5.83)
DPS	0 (2.59)	0 (5.82)	36 (1.14)	11 (1.39)	0 (48.2)	0 (22.7)
ΠGDM	100 (0.47)	100 (0.31)	100 (0.48)	61 (0.97)	–	–
RED-diff	100 (0.37)	100 (0.27)	100 (0.26)	100 (0.34)	29 (1.70)	100 (0.46)
DiffPIR	100 (0.00)	0 (1.61)	100 (0.23)	100 (0.36)	0 (4.07)	100 (0.04)
DAPS	100 (0.78)	0 (1.51)	100 (0.30)	100 (0.54)	1 (1.30)	100 (0.58)
PD-CDS + TV, TV constraint	95 (0.72)	76 (0.95)	79 (1.00)	96 (0.72)	74 (1.08)	100 (0.70)

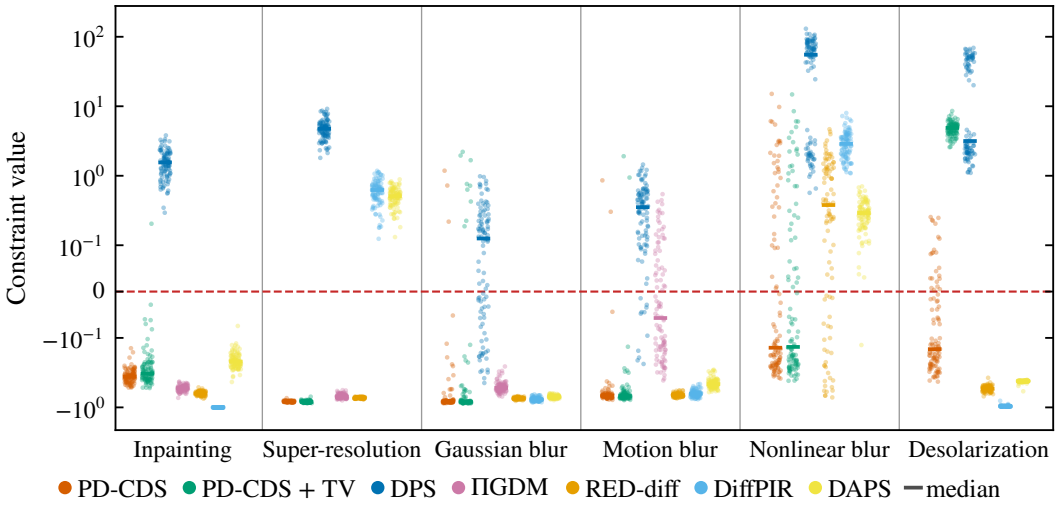


Figure 3: Normalized observation constraint g_{inv} at $t = 0$ for every image and method. Values below the dashed line are feasible; bars mark medians.

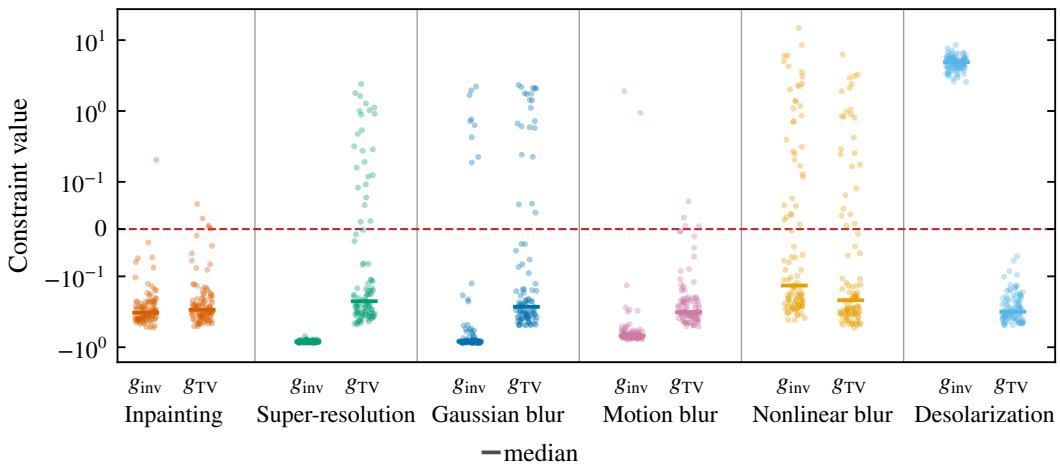


Figure 4: PD-CDS + TV: normalized observation and TV constraints, g_{inv} and g_{TV}, at $t = 0$ for every image. Values below the dashed line are feasible; bars mark medians.

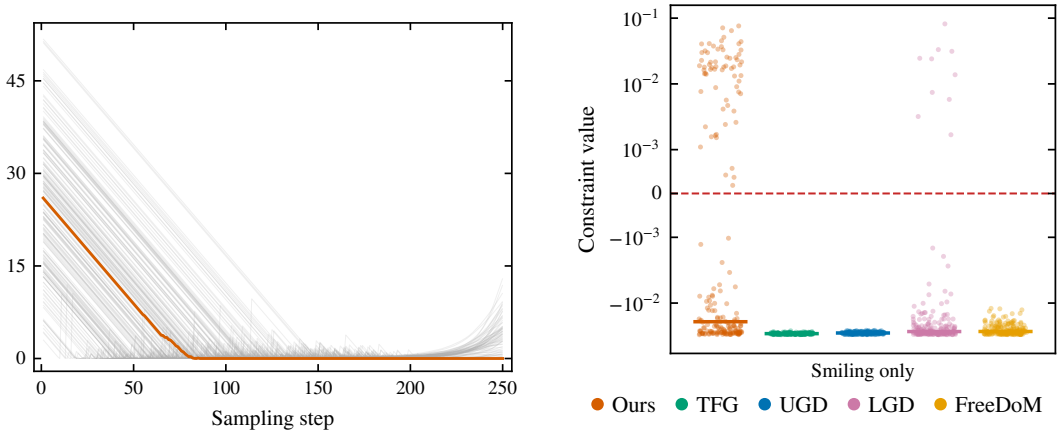


Figure 5: `Smiling` only: PD-CDS semantic multiplier (left) and semantic constraint value at $t = 0$ for all methods (right).

E.6 Additional experiments

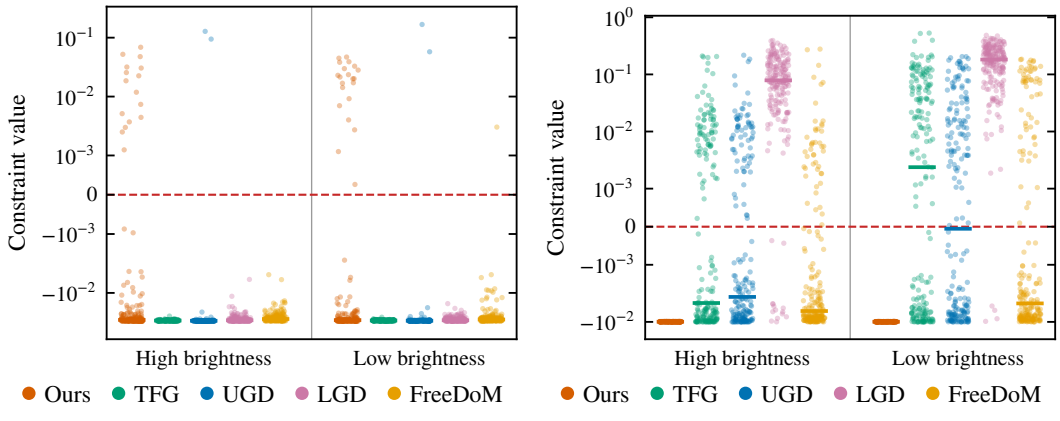
We additionally test PD-CDS on class-conditional generation on CIFAR-10. The task is to generate images from a given class (`Automobile`, `Bird`, `Frog` or `Truck`).

Implementation details. All methods use the same unconditional CIFAR-10 diffusion model, `openai_cifar10.pt`, released with the TFG implementation (Ye et al., 2024). Guidance is computed for all methods with the same ResNet-34 classifier, which we trained on CIFAR-10. In all PD-CDS runs, the denoised estimate $\hat{x}_{0|t}$ is clipped to $[-1, 1]$ before the guidance term is computed. For a target class y, we require the guidance classifier to assign probability at least 0.99 to y:

$$\ell_{\text{cls}}(x) = -\log p_{\text{ResNet}}(y \mid x) \le c_{\text{cls}}, \qquad c_{\text{cls}} = -\log 0.99.$$

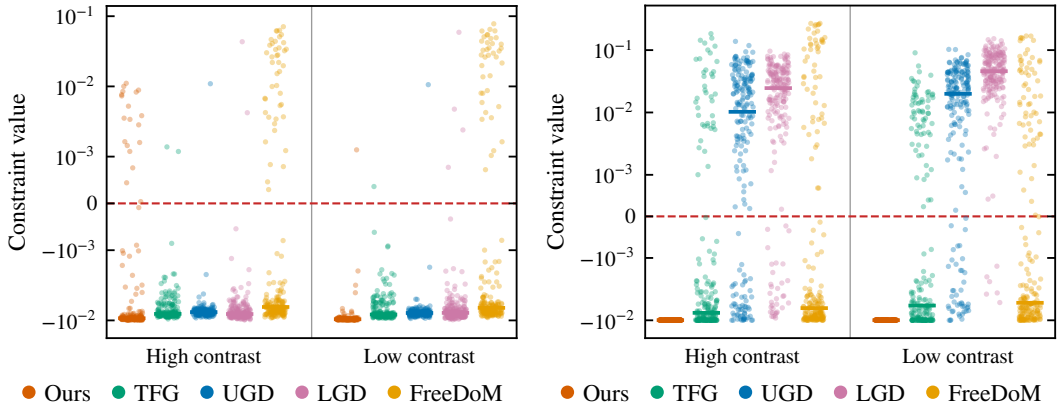
PD-CDS uses 250 respaced DDIM steps. The dual variable is updated with Adam and projected onto $\mathbb{R}_+$ after each step.

Baselines. TFG, UGD, LGD and FreeDoM use 100 DDIM steps and the implementations of Ye et al. (2024). For each method and class, we selected the configuration with the highest accuracy, using KID to choose among configurations with similar accuracy. The selected values are listed in Table 12; ρ, μ and σ follow the notation of Ye et al. (2024).



(a) `Smiling` + brightness: semantic.

(b) `Smiling` + brightness: photometric.



(c) `Not-Young` + contrast: semantic.

(d) `Not-Young` + contrast: photometric.

Figure 6: Constraint values at $t = 0$ on CelebA-HQ for all methods. Values below the dashed line are feasible. Semantic values use the guidance classifier and the PD-CDS thresholds $p^\star$ and are post-hoc diagnostics for the baselines.

Table 12: Hyperparameters used for class-conditional generation on CIFAR-10. PD-CDS uses 250 DDIM steps; the baselines use 100.

Method	Automobile	Bird	Frog	Truck
PD-CDS (Adam LR)	500	450	600	500
TFG (ρ, μ, σ)	0.5, 4, 0.001	1, 0.125, 0.5	0.25, 0.5, 0.25	0.5, 4, 0.001
UGD (strength)	4	2	3	2
LGD (strength)	11	28	16	6
FreeDoM (strength)	0.55	2.5	4	0.4

LGD uses one Monte Carlo sample in all cases.

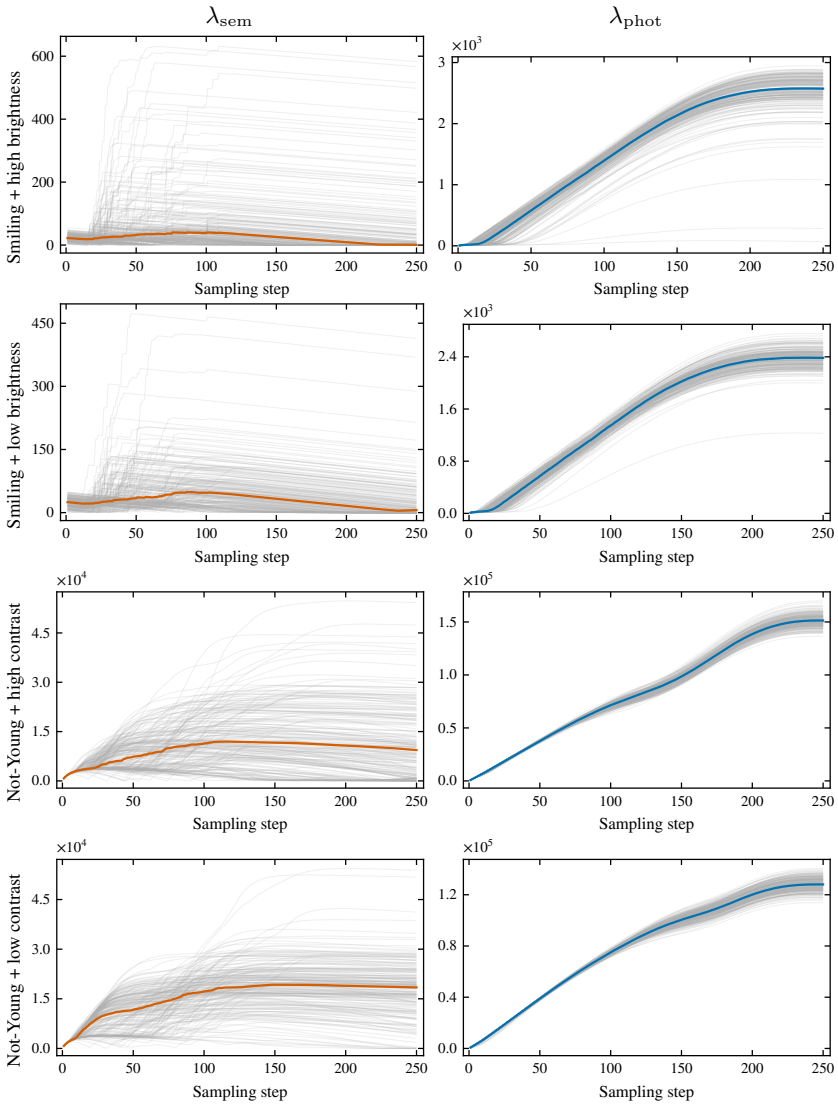


Figure 7: PD-CDS multipliers in the two-constraint CelebA-HQ settings. Grey: individual samples; colored: median. `Smiling` tasks use projected ascent ($\eta = 0.35$); `Not-Young` tasks use Adam (learning rate 750).

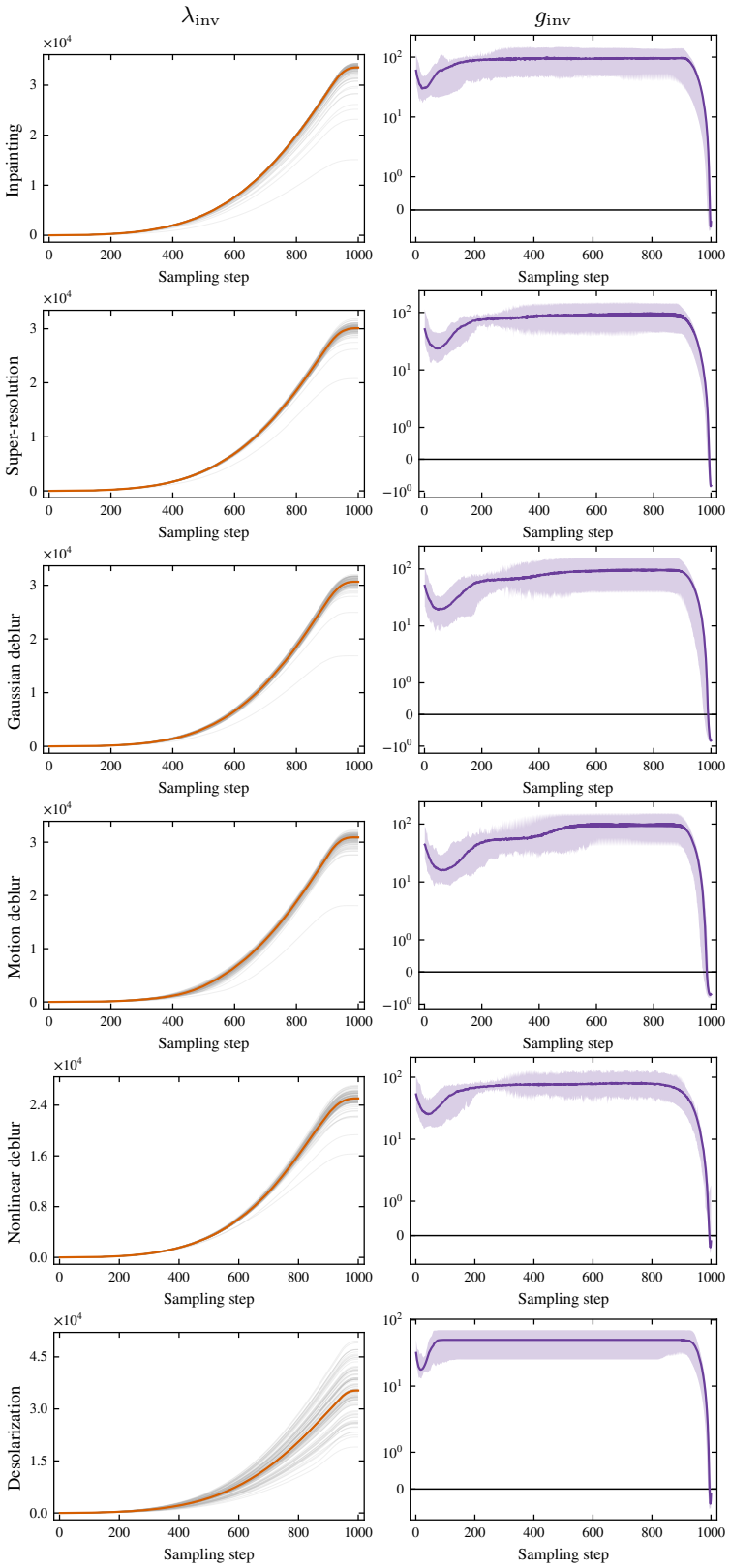


Figure 8: PD-CDS on inverse problems: observation multiplier (left) and the constraint value used in its update (right). Left: individual images (grey) and median (orange). Right: median (line) and spread across images (shaded).

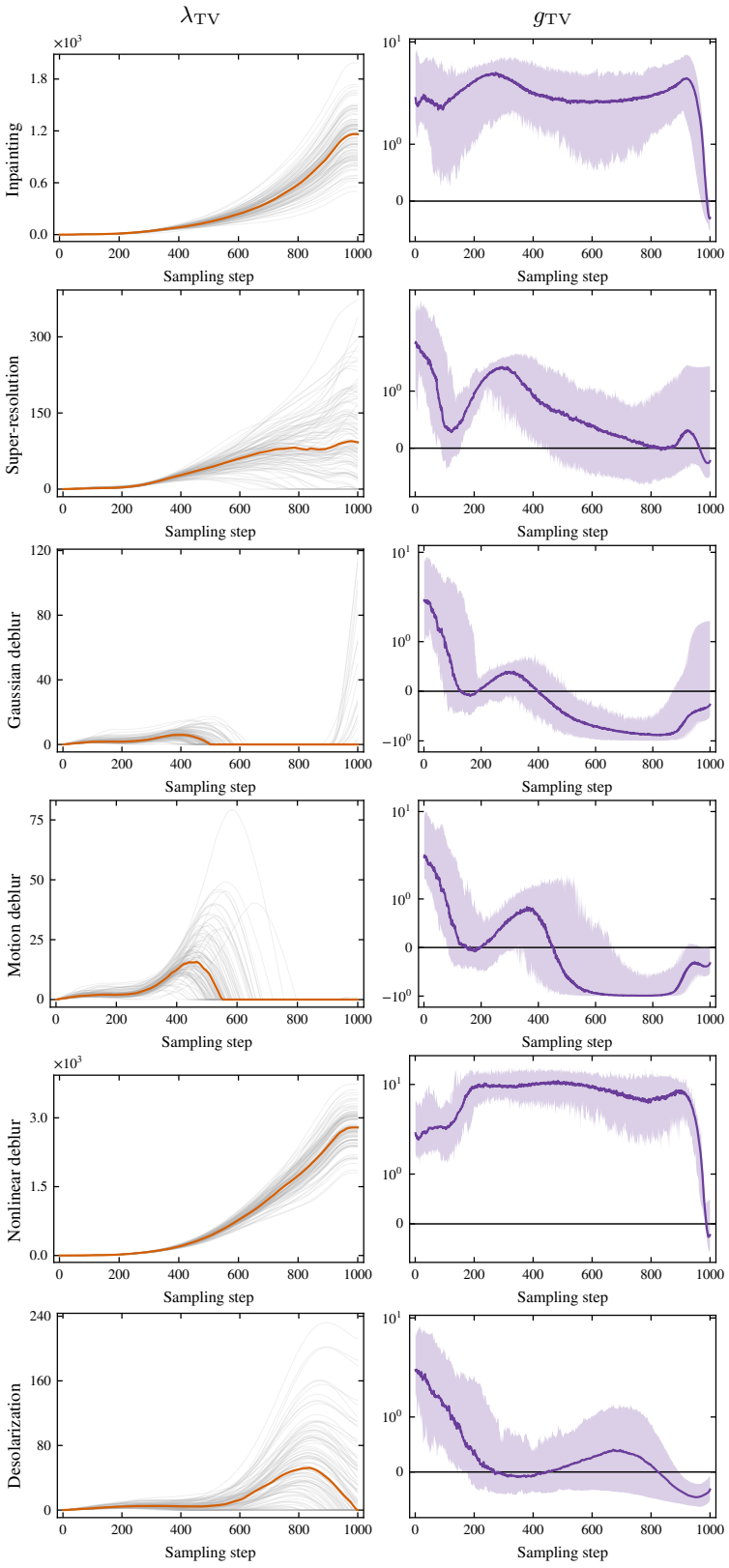


Figure 9: PD-CDS + TV: TV multiplier (left) and TV constraint value (right). Left: individual images (grey) and median (orange). Right: median (line) and spread across images (shaded).

For each target class and method, we generate $2{,}000$ images, pooled from four runs of 500 samples with different seeds. Accuracy is computed with an independent ConvNeXt-Tiny classifier (Liu et al., 2022) fine-tuned on CIFAR-10 (ahsanjavid, 2022), so that evaluation does not use the classifier that guides sampling. KID and FID compare the generated images with the $5{,}000$ training images of the target class. KID is averaged over 100 random subsets of size $1{,}000$, and FID is computed once on the full set of $2{,}000$ images.

Results. Results are reported in Table 13. PD-CDS achieves the highest accuracy and the lowest KID and FID on `Automobile`, `Bird` and `Truck`. On `Frog`, the trade-off differs: UGD reaches the highest external accuracy, while TFG obtains the lowest KID and FID. PD-CDS nevertheless remains close to UGD in accuracy (94.10% against 95.35%).

Table 13: Class-conditional generation on CIFAR-10. Best results within each class are shown in bold and second-best results are underlined.

Class	Method	Accuracy (%) ↑	KID ↓	FID ↓
Automobile	PD-CDS	**94.10**	**0.01072 ± 0.00068**	**27.067**
	TFG	<u>91.30</u>	<u>0.02286 ± 0.00136</u>	<u>33.375</u>
	UGD	84.45	0.02779 ± 0.00132	47.215
	LGD	10.35	0.28004 ± 0.00370	293.197
	FreeDoM	86.85	0.03887 ± 0.00186	50.308
Bird	PD-CDS	**84.00**	**0.02556 ± 0.00112**	**51.600**
	TFG	75.25	0.05553 ± 0.00165	88.105
	UGD	53.35	0.03287 ± 0.00146	<u>57.491</u>
	LGD	57.50	0.47617 ± 0.00250	390.220
	FreeDoM	<u>76.75</u>	<u>0.03124 ± 0.00125</u>	57.843
Frog	PD-CDS	<u>94.10</u>	0.03333 ± 0.00145	61.324
	TFG	85.05	**0.01646 ± 0.00103**	**35.074**
	UGD	**95.35**	<u>0.01978 ± 0.00114</u>	<u>38.605</u>
	LGD	78.70	0.02504 ± 0.00130	46.626
	FreeDoM	88.70	0.07395 ± 0.00251	96.602
Truck	PD-CDS	**95.80**	**0.01054 ± 0.00055**	**26.119**
	TFG	<u>83.70</u>	0.03498 ± 0.00147	42.625
	UGD	62.85	<u>0.02422 ± 0.00102</u>	<u>40.068</u>
	LGD	24.50	0.11560 ± 0.00225	158.045
	FreeDoM	67.65	0.02684 ± 0.00110	42.389

E.7 Sample images

We provide below qualitative results for the inverse problems of Section 5 (Figures 10–11) and for class-conditional generation on CIFAR-10 with PD-CDS and the training-free guidance baselines (Appendix E.6, Figure 12).

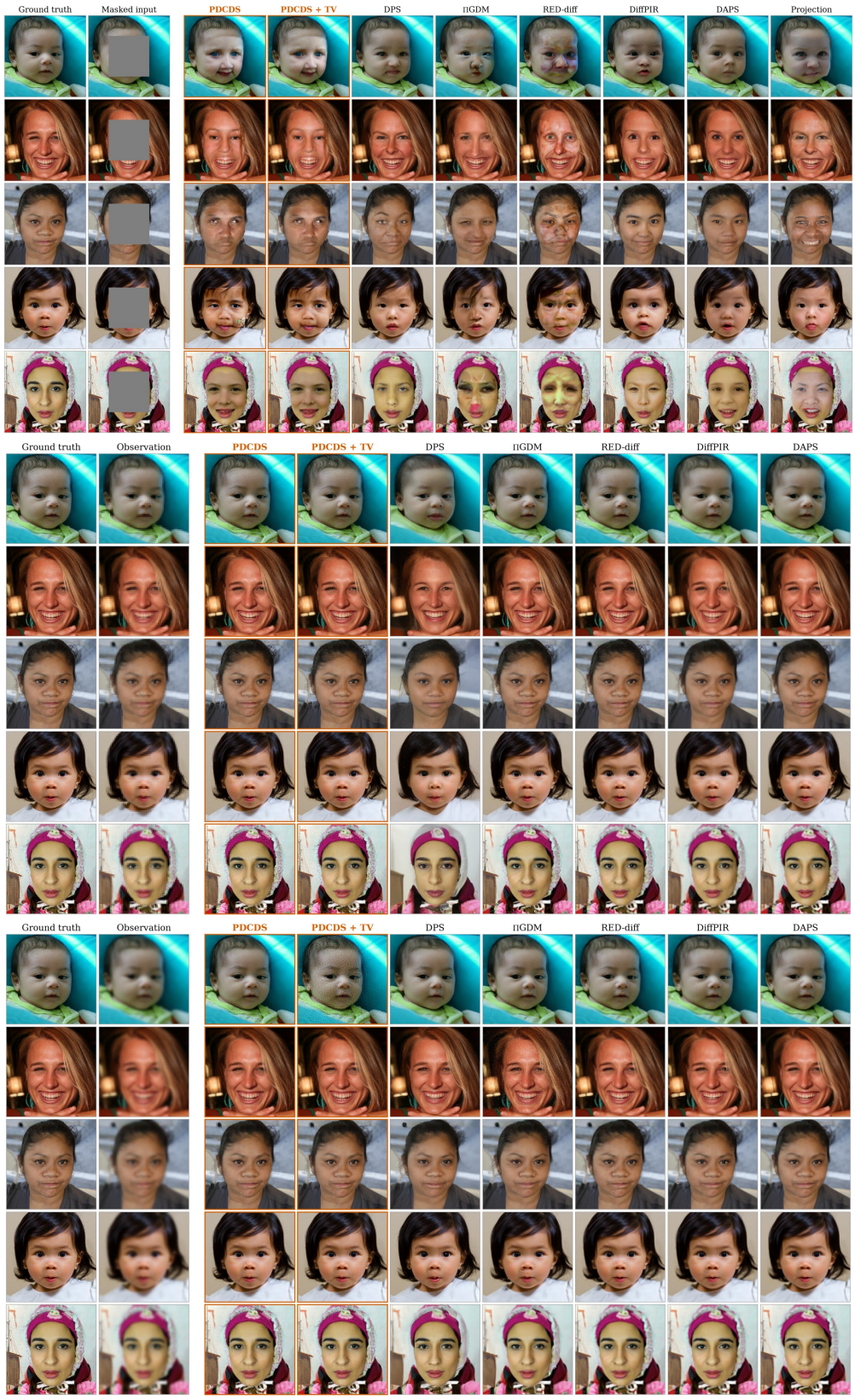


Figure 10: FFHQ reconstructions. Top: box inpainting; middle: super-resolution; bottom: Gaussian deblurring.

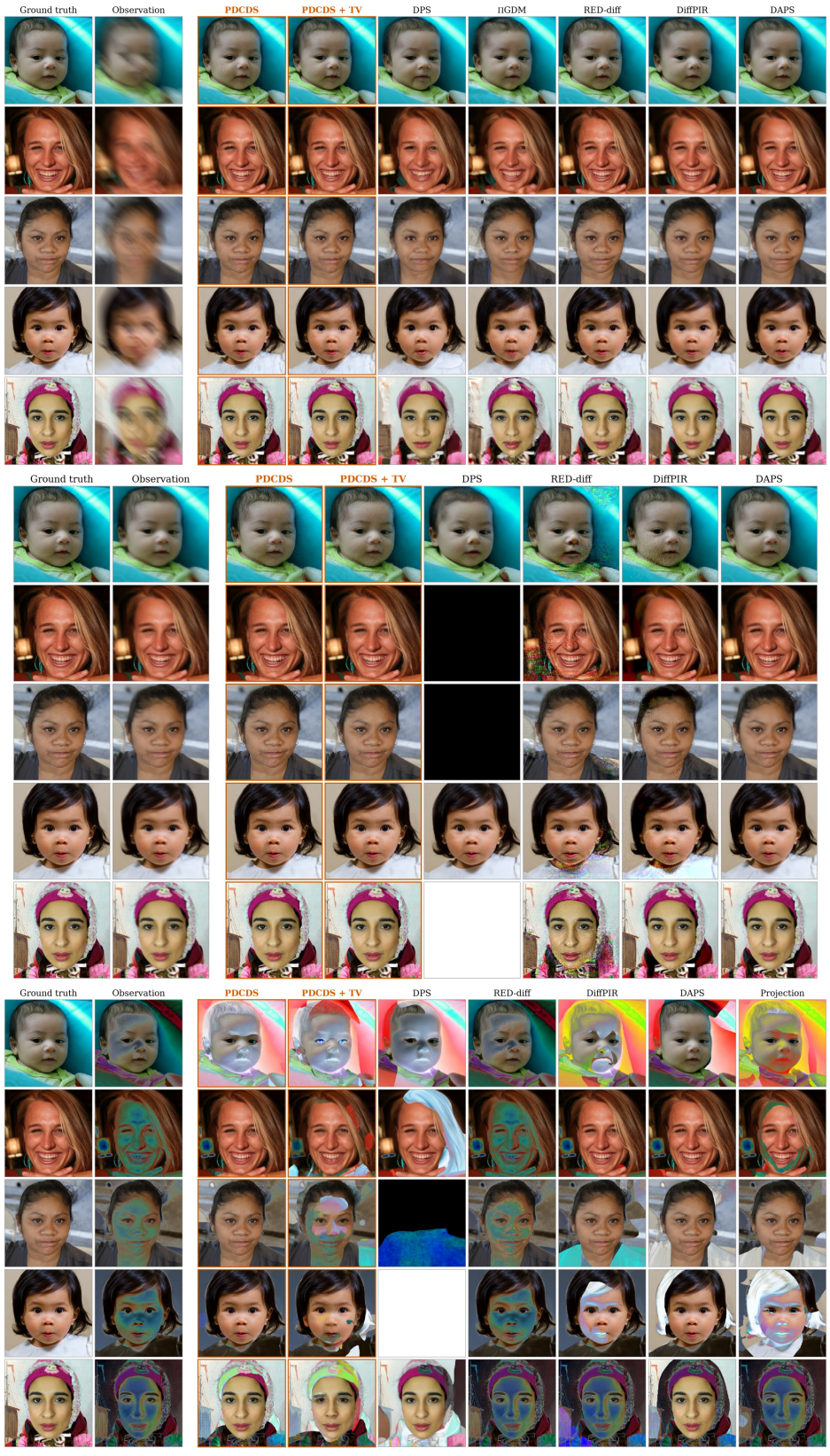


Figure 11: FFHQ reconstructions. Top: motion deblurring; middle: nonlinear deblurring; bottom: desolarization.



Figure 12: Class-conditional CIFAR-10 samples. Rows: methods; columns: target classes.