SELF-SUPERVISED VARIATIONAL DIFFUSION WITH TEMPORAL TUCKER MODELING FOR HYPERSPECTRAL IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperspectral image (HSI) restoration aims to recover spatial structures and spectrally faithful signatures from measurements affected by noise, missing observations, and spatial degradation. Existing supervised methods often generalize poorly across sensors and corruption distributions, while model-based approaches rely heavily on manually designed priors with limited expressive ability. Diffusion-based HSI restoration provides a promising alternative, but existing methods remain limited in modeling complex hyperspectral corruption and in fully exploiting the regularization imposed by diffusion sampling and HSI-specific priors. To address these limitations, we propose VITTA, a self-supervised variational diffusion framework with temporal Tucker modeling for HSI restoration. Specifically, VITTA couples a likelihood step and a prior step within the reverse diffusion process. In the likelihood step, a two-factor probabilistic model characterizes complex HSI corruption, and the inferred non-Gaussian uncertainty is projected into Gaussian observation channels to provide reliability-aware measurement conditioning. Complementarily, the prior step employs a Temporal Tucker Prior Module to estimate the clean endpoint under joint spatial, spectral, and reverse-step regularization. The resulting likelihood and prior information are then integrated through reliability-guided conditional diffusion sampling to progressively restore the HSI. Theoretically, we establish approximation guarantees and recursive error bounds, verifying the feasibility of the proposed progressive recovery process. Extensive experiments demonstrate that our proposed method outperforms state-of-the-art methods on various tasks including HSI denoising, band completion and pansharpening.

1 INTRODUCTION

Hyperspectral imaging records a scene over tens or hundreds of contiguous spectral bands, providing material-discriminative information that is unavailable to conventional RGB imagery (Bioucas-Dias et al., 2013; 2012). This capability is central to remote sensing, environmental monitoring, precision agriculture, mineral exploration, and target analysis (Govender et al., 2007; Dale et al., 2013; Bioucas-Dias et al., 2013). However, due to the inherent trade-off between acquisition efficiency and imaging quality, as well as hardware limitations and atmospheric interference, acquired HSIs often exhibit degraded quality, including noise contamination and limited spatial resolution (Xiong et al., 2019; Sidorov & Hardeberg, 2019; Yokoya et al., 2017). Hyperspectral restoration is therefore an essential step for preserving both spatial structures and spectral signatures (Wang et al., 2018).

Existing restoration paradigms exhibit distinct limitations. Supervised convolutional and transformer models can learn expressive degradation-to-clean mappings. However, their performance depends heavily on whether the training corpus adequately covers the scene statistics and degradation distributions encountered at test time, which limits cross-sensor and cross-corruption generalization (Miao et al., 2023; Li et al., 2023a;b). By contrast, model-based and self-supervised methods reduce the dependence on paired training data, but their performance is still strongly shaped by hand-crafted or architecture-induced priors, such as low-rankness, total variation, and deep image priors (Zhang et al., 2014; Wang et al., 2018; Peng et al., 2020; Luo et al., 2022; Ulyanov et al., 2018). Although these priors are interpretable, they can be insufficiently expressive for complex HSI structures, leading to oversmoothing or artifacts (Ulyanov et al., 2018; Miao et al., 2022).

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

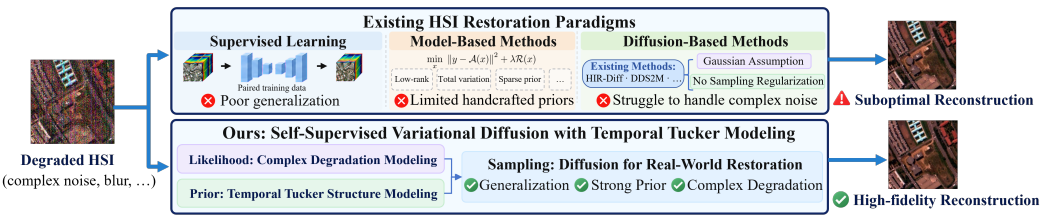


Figure 1: Comparison between existing HSI restoration paradigms and our proposed approach.

Recently, diffusion models have emerged as a promising alternative for image restoration, owing to the strong image priors encoded in their iterative denoising process (Ho et al., 2020; Song et al., 2021a;b). Methods such as SNIPS, DDRM, and DDNM (Kawar et al., 2021; 2022; Wang et al., 2023) enable diffusion priors to be conditioned on inverse-problem measurements at inference time without task-specific retraining. More recent studies extend diffusion-based restoration to general or complex observation models through likelihood guidance, variational inference, or learned noise distributions (Chung et al., 2023; Tang et al., 2024; Cheng et al., 2023; Cheng & Tan, 2024; Legin et al., 2023; Stevens et al., 2025). *Nevertheless, directly applying these approaches to HSI restoration remains challenging for three reasons. First, existing models often rely on oversimplified likelihood assumptions, such as Gaussian or Poisson distributions, which fail to characterize the complex corruptions encountered in HSIs. Secondly, learned noise models require representative noise samples or additional training distributions, which are difficult to obtain in hyperspectral imaging, where noise characteristics vary significantly across sensors and scenes, and matched training data are often scarce (Zhang et al., 2014; Xiong et al., 2019; Tan et al., 2025). Finally, general diffusion inverse solvers rely on pretrained conventional-image priors with fixed low-dimensional channel representations. However, HSIs exhibit sensor-dependent spatial and spectral dimensions, with substantially more spectral channels than natural images, making pretrained generative priors difficult to apply directly. While recent studies attempt to adapt diffusion models to HSI restoration (Pang et al., 2024; Miao et al., 2023), they are still limited to Gaussian noise assumptions and cannot be adapted to complex noise easily.*

To alleviate these issues, we propose a self-supervised variational diffusion framework namely VITTA for HSI restoration with complex degradations. The core idea is to make diffusion restoration applicable to complex HSI corruption through probabilistic noise modeling and Gaussian projection, while providing a stronger scene prior through joint spatial–spectral–temporal regularization. Specifically, we introduce a two-factor variational model to characterize band-dependent dense noise and sparse heavy-tailed corruption. The inferred non-Gaussian uncertainty is then projected into Gaussian observation channels, enabling conditional diffusion sampling under complex noise. Meanwhile, we develop a Temporal Tucker Prior Module to exploit spatial, spectral, and reverse-step dependencies for clean-endpoint estimation. The resulting observation statistics and structural prior are integrated into the reverse transition for progressive restoration along the entire reverse trajectory. Theoretically, we show that Gaussian projection preserves key restoration statistics with residual errors further attenuated by diffusion, while recursive endpoint-error analysis characterizes the stability of progressive HSI recovery. The framework comparison is shown in Figure 1.

The main contributions are summarized as follows:

- We propose VITTA, a self-supervised variational diffusion framework that bridges complex degradations with diffusion sampling, achieving state-of-the-art performance across multiple HSI restoration tasks. Theoretically, we derive recursive endpoint error bounds that formally verify the stability and progressive convergence of our recovery dynamics.
- We develop a variational likelihood and Gaussian projection scheme to enable principled diffusion conditioning under complex HSI noise. Additionally, we prove that this projection exactly preserves the quadratic restoration risk while diffusion contracts higher-order discrepancies, thereby guaranteeing faithful spectral-spatial recovery.
- We develop a Temporal Tucker Prior Module (TTPM) that jointly models scene structure and reverse-step evolution, providing trajectory-aware regularization for clean-endpoint estimation throughout the reverse process.

2 BACKGROUND

Let $X_0 \in [0,1]^{N_h \times N_w \times B}$ be the unknown hyperspectral image (HSI), $x_0 = \mathrm{vec}(X_0) \in \mathbb{R}^n$, and $n = N_h N_w B$. We consider a single degraded observation

$$y = Hx_0 + z, \qquad H \in \mathbb{R}^{m \times n}, \tag{1}$$

where the known operator H describes the structural degradation and z denotes additive measurement corruption. The goal is to recover x_0 from y.

DDS2M (Miao et al., 2023) addresses such inverse problems through measurement-conditioned diffusion sampling. Let $u_0 = 2x_0 - \mathbf{1}_n$ denote the clean image rescaled to the diffusion domain. Starting from u_0, each forward diffusion step applies a Gaussian perturbation

$$q(u_t \mid u_{t-1}) = \mathcal{N}\Big(\sqrt{1-\beta_t}\, u_{t-1}, \beta_t I_n\Big), \qquad 0 < \beta_t < 1, \tag{2}$$

which yields the closed-form marginal

$$q^{\mathrm{fwd}}(u_t \mid u_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, u_0, (1-\bar{\alpha}_t) I_n\big), \qquad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s). \tag{3}$$

Restoration is then formulated as a conditional reverse process. For a fixed observation model, let $q(u_{1:T} \mid u_0, y, H)$ denote an ideal reference path with initial law q_T and reverse kernels $q_t = q(u_{t-1} \mid u_t, u_0, y, H)$ for $t = 2, \ldots, T$. The corresponding implementable process, parameterized by Θ, consists of an initial distribution p_T and transition kernels $p_{\Theta,t}$. If the clean endpoint u_0 were available, a natural objective would be to maximize the conditional log-likelihood $\log p_\Theta(u_0 \mid y, H)$, thereby encouraging the reverse process to assign high probability to the desired clean reconstruction. However, marginalizing the latent intermediate states is intractable. Introducing q as an auxiliary distribution and applying Jensen's inequality gives

$$\begin{aligned} \log p_\Theta(u_0 \mid y, H) \geq\, & \mathbb{E}_q \log p_{\Theta,0}(u_0 \mid u_1, y, H) \\ & - D_{\mathrm{KL}}(q_T \| p_T) - \sum_{t=2}^{T} \mathbb{E}_q D_{\mathrm{KL}}(q_t \| p_{\Theta,t}). \end{aligned} \tag{4}$$

With q and p_T fixed, the initial KL term is constant with respect to Θ. The reconstruction term encourages recovery of the clean endpoint u_0, while the transition KL terms encourage $p_{\Theta,t}$ to approximate the ideal reverse kernels q_t. Since these kernels depend on the unknown u_0, DDS2M, inspired by deep image prior (Ulyanov et al., 2018) and self-supervised learning, constructs $p_{\Theta,t}$ within the same Gaussian transition family as q_t but replaces u_0 with its timestep-dependent estimate $\widehat{u}_t(\Theta)$. Accordingly, motivated by the Gaussian transition KL, we use the following observable timestep-wise endpoint-fitting objective:

$$\mathcal{L}_{\mathrm{diff},t}(\Theta) = \frac{\|u_t - \sqrt{\bar{\alpha}_t}\widehat{u}_t(\Theta)\|^2}{n(1-\bar{\alpha}_t)}. \tag{5}$$

After fitting $\widehat{u}_t$ with this observable surrogate at timestep t, it is substituted into $p_{\Theta,t}$, which is combined with the observation model to sample u_{t-1}. The sampled state then serves as the input for endpoint estimation at the next reverse step. DDS2M thus alternates between scene-specific endpoint estimation and measurement-conditioned reverse sampling, progressively evolving from a noisy diffusion state toward the final clean reconstruction. Appendix C provides the complete conditional-path factorization. However, this framework remains limited in handling complex real-world HSI observations, as its simplified Gaussian assumption cannot adequately characterize heterogeneous noise patterns. Moreover, its scene-specific endpoint estimation does not explicitly exploit the joint spatial–spectral structure of HSIs and the temporal dependencies across reverse diffusion steps.

3 METHOD

3.1 OVERALL FRAMEWORK

To address limitations of existing approaches, we propose a self-supervised variational diffusion framework for HSI restoration under complex degradations. The proposed framework comprises

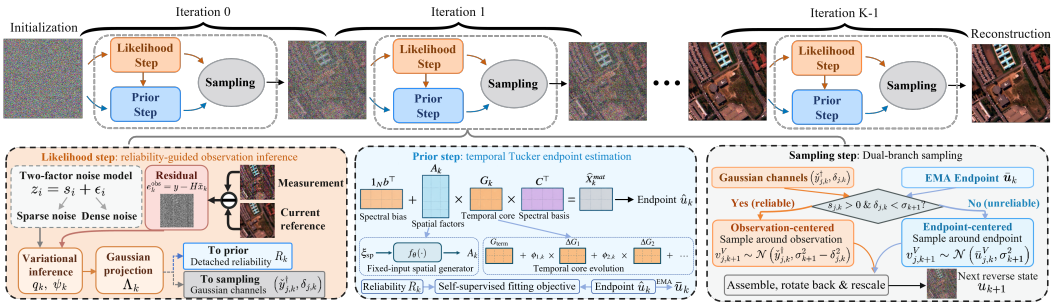


Figure 2: Overview of the proposed framework. The likelihood step infers complex corruption and Gaussian observation statistics with associated reliability, while the prior step estimates the clean endpoint using the temporal Tucker model. These two sources of information are then integrated through reliability-guided conditional sampling to generate the next reverse state.

three coordinated stages, referred to as the likelihood, prior, and sampling steps. At each reverse timestep t, the *likelihood step* infers the statistics of complex corruption from the current residual and projects them into a Gaussian representation compatible with diffusion conditioning. In parallel, the *prior step* employs a temporal Tucker prior to estimate the clean endpoint $\widehat{u}_t$ from the current diffusion state u_t by exploiting the spatial, spectral, and reverse-time structure of the HSI. The *sampling step* then combines the projected observation information and the estimated endpoint to construct the conditional transition from u_t to u_{t-1}. Repeating these three steps along the reverse trajectory progressively refines both the observation characterization and the underlying clean-scene estimate. An overview of the proposed framework is shown in Figure 2. The complete algorithmic flow is provided in Appendix B. The following subsections describe each component in detail.

3.2 Likelihood Step: Variational Corruption Modeling

Corruption inference. To model the complex corruption in HSI observations beyond the conventional Gaussian assumption, we decompose the observation residual into band-dependent dense noise and sparse heavy-tailed corruption. Specifically, the noise of acquired measurement i is modeled as

$$z_i = s_i + \epsilon_i, \qquad \epsilon_i \mid \tau_{b(i)} \sim \mathcal{N}(0, \tau_{b(i)}^{-1}), \qquad c_i \sim \text{Categorical}(\pi_n, \pi_s). \tag{6}$$

Here $b(i)$ denotes the spectral band, τ_b is the corresponding dense-noise precision, and c_i indicates whether the measurement belongs to the ordinary or sparse state. In the ordinary state, $s_i = 0$. In contrast, the sparse state is modeled as

$$\omega_i \sim \text{Gamma}(\nu/2, \nu/2), \qquad s_i \mid \omega_i, c_i = s \sim \mathcal{N}(0, v_{b(i)}/\omega_i). \tag{7}$$

This scale-mixture construction yields a Student-t distribution for the sparse amplitude (Andrews & Mallows, 1974), where $v_b > 0$ controls its scale. We use shape–rate Gamma distributions with $\nu > 1$. Let $\widetilde{x}_t$ be the detached radiometric reference, let $e_{i,t}^{\text{obs}} = y_i - h_i^\top \widetilde{x}_t$, and define $\psi = \{\pi, \tau_{1:B}, v_{1:B}\}$. The inference target is the latent posterior $p_\psi(\zeta \mid y, H\widetilde{x}_t)$, where ζ collects the state variables together with the active sparse amplitudes and scales. For fixed ψ, the mean-field inference objective is

$$\min_{q \in \mathcal{Q}} D_{\text{KL}}(q(\zeta) \| p_\psi(\zeta \mid y, H\widetilde{x}_t)). \tag{8}$$

Using the mean-field factors $q(c_i = s) = r_{is}$, $q(s_i \mid c_i = s) = \mathcal{N}(m_i, V_i)$, and $q(\omega_i \mid c_i = s) = \text{Gamma}(a_\omega, b_{\omega i})$, coordinate optimization of Eq. (8) yields

$$\begin{aligned} V_i &= (\tau_b + \bar{\omega}_i / v_b)^{-1}, \qquad m_i = V_i \tau_b e_{i,t}^{\text{obs}}, \\ a_\omega &= (\nu + 1)/2, \qquad b_{\omega i} = \tfrac{1}{2}\{\nu + (m_i^2 + V_i)/v_b\}, \\ r_{ij} &= \text{softmax}_j(\log \pi_j + A_{ij}), \qquad \bar{\omega}_i = a_\omega / b_{\omega i}. \end{aligned} \tag{9}$$

Here $b = b(i)$, and A_{ij} denotes the state evidence. Iterating these updates yields the variational corruption distribution q_t^{obs}, which characterizes the sparse-state probability and noise-scale statistics required by the subsequent Gaussian projection. The derivation and conjugate nuisance-parameter updates are provided in Appendix D.

Gaussian projection. To adapt the non-Gaussian observation residual for Gaussian-formulated conditional diffusion, we project its inferred distribution onto a Gaussian distribution. At reverse timestep t, for acquired measurement i, averaging the zero-centered conditional noise model over the inferred corruption state and scale defines the auxiliary residual density

$$f_{i,t}^{q}(e) = (1 - r_{is})\mathcal{N}(e \mid 0, s_{n,i}) + r_{is}\mathbb{E}_{q_i(\omega|c_i=s)}\mathcal{N}(e \mid 0, s_i(\omega)), \tag{10}$$

where

$$s_{n,i} = \tau_{b(i)}^{-1}, \qquad s_i(\omega) = s_{n,i} + v_{b(i)}/\omega.$$

Thus, $f_{i,t}^{q}$ summarizes the residual noise distribution implied by the variational inference at the current reverse timestep. We obtain its Gaussian representation by minimizing the forward KL divergence

$$(\mu_{i,t}^{\mathrm{G}}, S_{i,t}^{\mathrm{G}}) = \arg\min_{\mu\in\mathbb{R},\ S>0} D_{\mathrm{KL}}\left(f_{i,t}^{q} \| \mathcal{N}(\mu, S)\right). \tag{11}$$

Under the Gaussian projection in Eq. (11), the forward-KL optimum is the moment-matched Gaussian. Since $f_{i,t}^{q}$ is zero centered, the optimal projected mean, variance, and precision are

$$\mu_{i,t}^{\mathrm{G}} = 0, \qquad S_{i,t}^{\mathrm{G}} = \tau_{b(i)}^{-1} + r_{is}v_{b(i)}\mathbb{E}_q[\omega_i^{-1}], \qquad \Lambda_{i,t} = (S_{i,t}^{\mathrm{G}})^{-1}. \tag{12}$$

Incorporating the degradation operator H, we normalize the measurement as $y^u = 2y - H\mathbf{1}_n = Hu_0 + 2z$ and define

$$\mathsf{W}_t = \mathrm{diag}(\Lambda_t)/4, \qquad \mathsf{W}_t^{1/2}H = U_t\Sigma_t V_t^{\top}, \qquad \breve{y}_t = U_t^{\top}\mathsf{W}_t^{1/2}y^u. \tag{13}$$

For each positive singular value $s_{j,t}$, this yields the independent Gaussian observation channel

$$\breve{y}_{j,t}^{\dagger} = u_{0,j}^{V} + \delta_{j,t}\epsilon_j, \qquad \breve{y}_{j,t}^{\dagger} = \breve{y}_{j,t}/s_{j,t}, \qquad \delta_{j,t} = s_{j,t}^{-1},$$

where $\epsilon_j \sim \mathcal{N}(0,1)$. Thus, the likelihood step converts the inferred complex corruption into Gaussian observation channels directly usable by the conditional sampler. While higher-order distributional shape is not retained, its effect is limited within the proposed diffusion framework and the projection does more than select a Gaussian approximation, as will be detailed in Section 3.5.

3.3 Prior Step: Temporal Tucker Endpoint Modeling

The prior step estimates the clean endpoint required by conditional reverse diffusion, particularly in directions that are insufficiently constrained by the observation. To regularize this estimation in the self-supervised setting, we introduce the Temporal Tucker Prior Module (TTPM), which exploits both the spatial–spectral structure of HSIs and the dependence among endpoint estimates across reverse diffusion steps. Let $N = N_h N_w$, and let $\mathcal{V}$ reshape an HSI into an $N \times B$ matrix. At reverse timestep t, the estimated clean HSI is represented as

$$\widehat{X}_t^{\mathrm{mat}} = \mathbf{1}_N\boldsymbol{b}^{\top} + A_t G_t C^{\top}, \qquad \widehat{u}_t = 2\,\mathrm{vec}\Big(\mathcal{V}^{-1}(\widehat{X}_t^{\mathrm{mat}})\Big) - \mathbf{1}_n. \tag{14}$$

Here $A_t \in \mathbb{R}^{N\times R_s}$ represents spatial patterns, $C \in \mathbb{R}^{B\times R_\lambda}$ provides shared spectral coordinates, $G_t \in \mathbb{R}^{R_s\times R_\lambda}$ couples the spatial and spectral factors, and $\boldsymbol{b} \in \mathbb{R}^B$ is a spectral bias. To further impose an implicit spatial prior, the columns of A_t are generated from fixed random inputs as

$$A_t = [\mathrm{vec}(f_{\theta_1}(\xi_{\mathrm{sp}})), \ldots, \mathrm{vec}(f_{\theta_{R_s}}(\xi_{\mathrm{sp}}))], \tag{15}$$

where ξ_{sp} is sampled once and kept fixed throughout restoration. The generator parameters, C, and $\boldsymbol{b}$ are jointly optimized across reverse timesteps, providing a scene-specific deep image prior (Ulyanov et al., 2018; Miao et al., 2023). Since all reverse timesteps estimate the same underlying scene at different diffusion noise levels, we constrain the Tucker cores to evolve along a shared low-dimensional cosine trajectory. For $t = T, \ldots, 1$ and $1 \le R_{\mathrm{temp}} \le T$, we define

$$\begin{aligned} T_{t\ell} &= \sqrt{2}\cos\left(\frac{(T - t + 1/2)\ell\pi}{T}\right), \qquad \phi_{\ell,t} = T_{t\ell} - T_{1,\ell}, \\ G_t &= G_{\mathrm{term}} + \sum_{\ell=1}^{R_{\mathrm{temp}}-1} \phi_{\ell,t}\Delta G_\ell. \end{aligned} \tag{16}$$

This parameterization allows timestep-dependent adaptation while encouraging smooth evolution along the reverse trajectory, with the final stage anchored at $G_1 = G_{\text{term}}$.

With the inferred observation statistics fixed, TTPM estimates the clean endpoint by minimizing

$$\mathcal{J}_t = \mathcal{L}_{\text{diff},t} + \lambda_{\text{spec}}\mathcal{L}_{\text{spec}} + \lambda_{\text{rel},t}\mathcal{L}_{\text{rel},t} + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{phys},t}. \tag{17}$$

The auxiliary terms constrain spectral smoothness, reliability-aware fitting, temporal core variation, and physical consistency, respectively. Their definitions and optimization details are provided in Appendix E. The resulting endpoint estimate is then passed to the sampling step.

3.4 Sampling Step: Reliability-Guided Conditional Diffusion

The sampling step combines the Gaussian observation channels from Section 3.2 with the clean-endpoint estimate from Section 3.3. Following the measurement-conditioned sampling of DDRM (Kawar et al., 2022) and DDS2M (Miao et al., 2023), we construct the reverse transition in the singular-vector space of the observation model. In contrast to the predefined, globally shared Gaussian noise level used in DDRM and DDS2M, our transition uses the channel-wise uncertainty inferred from complex corruption to determine whether each direction should rely on the observation or the estimated endpoint.

Let $\sigma_t^2 = (1-\bar{\alpha}_t)/\bar{\alpha}_t$ denote the diffusion variance on the clean-image scale, let $\bar{u}_t$ denote the running average of the fitted endpoints $\hat{u}_t$, and let $\overline{u}_t^V = V_t^\top \overline{u}_t$. For each singular direction, the next reverse state is sampled as

$$\frac{(V_t^\top u_{t-1})_j}{\sqrt{\bar{\alpha}_{t-1}}} = \begin{cases} \check{y}_{j,t}^\dagger + \sqrt{\sigma_{t-1}^2 - \delta_{j,t}^2}\,\xi_{j,t}, & s_{j,t} > 0,\ \delta_{j,t} < \sigma_{t-1}, \\ \overline{u}_{j,t}^V + \sigma_{t-1}\xi_{j,t}, & \text{otherwise}, \end{cases} \qquad \xi_t \sim \mathcal{N}(0, I_n). \tag{18}$$

Reliable observation channels are centered at the projected measurements, whereas unreliable or unobserved directions rely on the TTPM endpoint estimate. The variance assigned to the two branches follows from matching the target diffusion distribution, as formalized below.

Proposition 1 (KL-optimal variance completion). *Let $Y = u + \delta Z$ with $Z \sim \mathcal{N}(0,1)$ be a Gaussian observation channel, and let $\xi \sim \mathcal{N}(0,1)$ be independent. For $T = Y + \sqrt{c}\,\xi$, $c \geq 0$, and a target scale $\sigma > 0$,*

$$\arg\min_{c\geq 0} D_{\text{KL}}\big(\mathcal{N}(u, \delta^2 + c) \| \mathcal{N}(u, \sigma^2)\big) = (\sigma^2 - \delta^2)_+. \tag{19}$$

Moreover, for a fixed endpoint mean m, the variance minimizing $D_{\text{KL}}(\mathcal{N}(m,c)\|\mathcal{N}(u,\sigma^2))$ over $c > 0$ is $c = \sigma^2$.

Proposition 1 directly justifies the two branches in Eq. (18). When $\delta_{j,t} < \sigma_{t-1}$, the observation channel already contains variance $\delta_{j,t}^2$, and adding $\sigma_{t-1}^2 - \delta_{j,t}^2$ exactly matches the target diffusion variance and achieves zero KL mismatch. When $\delta_{j,t} > \sigma_{t-1}$, this mismatch cannot be corrected by adding nonnegative Gaussian noise, motivating the endpoint-centered branch with target variance σ_{t-1}^2. Unobserved directions follow the same branch since no measurement channel is available. The proof is provided in Appendix F. The sampled coordinates are finally rotated back and rescaled to obtain u_{t-1}.

3.5 Analysis: Approximation and Restoration Guarantees

Gaussian-projection approximation. We characterize the influence of Gaussian projection through moment preservation, bounded higher-order discrepancy, and diffusion-induced attenuation.

At a fixed inference snapshot, Eq. (11) gives the moment-matched Gaussian $G_t^q = \prod_i \mathcal{N}(0, S_{i,t}^{\text{G}})$ for the inferred residual distribution $F_t^q = \prod_i f_{i,t}^q$. The following result establishes that the second-order information relevant to Gaussian conditional sampling is preserved exactly.

Proposition 2 (Second-order invariance under Gaussian projection). *Assume finite positive scales and $\nu > 1$. Let $E \sim F_t^q$, $E_{\text{G}} \sim G_t^q$, and let ξ be independent standard Gaussian noise. For any fixed affine transition $\mathcal{A}(e,\xi) = a + Le + C\xi$, $\mathcal{A}(E,\xi)$ and $\mathcal{A}(E_{\text{G}},\xi)$ have identical means and covariances. Moreover, for any fixed target u,*

$$\mathbb{E}\|\mathcal{A}(E,\xi) - u\|^2 = \mathbb{E}\|\mathcal{A}(E_{\text{G}},\xi) - u\|^2. \tag{20}$$

With the inferred statistics and channel construction fixed, the conditional transition is affine in the observation residual with additive Gaussian innovation. Proposition 2 therefore implies that replacing the non-Gaussian residual by its projection does not alter the transition mean, covariance, or quadratic restoration risk, which are the dominant quantities governing the conditional transition. In addition, the discarded higher-order information can also be quantified explicitly and admits a stagewise upper bound, as summarized below.

Proposition 3 (Bounded projection discrepancy)**.** *Let* $\varepsilon_{i,t} := D_{\mathrm{KL}}\big(f^q_{i,t} \| \mathcal{N}(0, S^{\mathrm{G}}_{i,t})\big)$. *Then*

$$0 \le \varepsilon_{i,t} \le \mathcal{B}_{i,t} := \tfrac{1}{2}\big[\log S^{\mathrm{G}}_{i,t} - (1 - r_{is}) \log s_{n,i} - r_{is}\mathbb{E}_q \log s_i(\omega)\big], \quad D_{\mathrm{KL}}(F^q_t \| G^q_t) \le \sum_i \mathcal{B}_{i,t}. \tag{21}$$

Finally, the remaining discrepancy is further reduced once the projected observation enters the diffusion transition.

Proposition 4 (Diffusion attenuation of projection discrepancy)**.** *Let* E *be a centered scalar residual with* $\mathrm{Var}(E) = \delta^2 < \sigma^2$ *and*

$$\varepsilon_{\mathrm{in}} = D_{\mathrm{KL}}\big(\mathcal{L}(E) \| \mathcal{N}(0, \delta^2)\big).$$

For independent $\xi \sim \mathcal{N}(0, 1)$,

$$\varepsilon_{\mathrm{out}} := D_{\mathrm{KL}}\Big(\mathcal{L}\Big(u + E + \sqrt{\sigma^2 - \delta^2}\,\xi\Big) \| \mathcal{N}(u, \sigma^2)\Big) \le \frac{\delta^2}{\sigma^2}\varepsilon_{\mathrm{in}}. \tag{22}$$

Since $\delta^2/\sigma^2 < 1$, Proposition 4 shows that variance-completing Gaussian noise contracts the residual shape discrepancy in every observation-centered channel. This attenuation is generally stronger at early reverse timesteps, where the target diffusion variance is relatively large. As the reverse trajectory approaches the clean endpoint, σ_{t-1} decreases and the condition $\delta_{j,t} < \sigma_{t-1}$ becomes increasingly restrictive, so only sufficiently reliable channels continue to inject the projected observation directly. Together, these results show that Gaussian projection has limited influence on restoration because the second-order information governing conditional sampling is preserved exactly, the remaining higher-order discrepancy is bounded, and the diffusion transition further attenuates this discrepancy. The corresponding proofs and details are provided in Appendices G, F, H, and I.

Restoration mechanism. We next examine how the reverse transition generates the fitting signal for clean-endpoint estimation at the subsequent timestep. Let $\mathcal{A}_t = \{j : s_{j,t} > 0, \delta_{j,t} < \sigma_{t-1}\}$ denote the observation-centered directions, $F = \widehat{u}_{t-1}(\Theta)$, $F^V = V_t^\top F$, and $\mathcal{P}_{t-1} = \mathcal{J}_{t-1} - \mathcal{L}_{\mathrm{diff},t-1}$.

Proposition 5 (Sampled fitting-objective decomposition)**.** *Up to a positive scaling that preserves the minimizer, the expected fitting objective induced by Eq. (18) satisfies*

$$\begin{aligned}\mathbb{E}_{\xi_t}\left[\tfrac{1}{2}\left\|F - \frac{u_{t-1}}{\sqrt{\bar\alpha_{t-1}}}\right\|^2 + \lambda^{\mathrm{fit}}_{t-1}\mathcal{P}_{t-1}(\Theta)\right] &= \underbrace{\tfrac{1}{2}\sum_{j\in\mathcal{A}_t}\left(F^V_j - \check{y}^\dagger_{j,t}\right)^2}_{\textit{measurement consistency}} + \underbrace{\tfrac{1}{2}\sum_{j\notin\mathcal{A}_t}\left(F^V_j - \overline{u}^V_{j,t}\right)^2}_{\textit{endpoint continuity}} \\ &+ \underbrace{\lambda^{\mathrm{fit}}_{t-1}\mathcal{P}_{t-1}(\Theta)}_{\textit{structural and physical regularization}} + \underbrace{\tfrac{1}{2}\sum_{j\in\mathcal{A}_t}\left(\sigma^2_{t-1} - \delta^2_{j,t}\right) + \tfrac{1}{2}\sum_{j\notin\mathcal{A}_t}\sigma^2_{t-1}}_{\textit{sampling variance}},\end{aligned} \tag{23}$$

where $\lambda^{\mathrm{fit}}_{t-1} = n\sigma^2_{t-1}/2$.

Proposition 5 shows that reliable directions are corrected by the observations, while unreliable or unobserved directions inherit the preceding endpoint and are further regularized by TTPM. Under the observation–regularization complementarity and finite-fitting conditions, the endpoint error further satisfies

$$\mathcal{E}_{t-1} \le \rho_t \mathcal{E}_t + \eta_t, \qquad 0 \le \rho_t < 1, \tag{24}$$

where $\mathcal{E}_t = (\mathbb{E}\|\overline{u}_t - u_0\|^2)^{1/2}$ and η_t collects the accumulated perturbation terms. Thus, repeated sampling and fitting suppress inherited endpoint errors while keeping newly introduced errors explicit. The corresponding derivations and finite-horizon bound are provided in Appendices H and I.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate complex-noise denoising, missing-band completion, and noisy pansharpening on Chikusei, Pavia University, and Houston 2018 (Yokoya & Iwasaki, 2016; Huang & Zhang, 2009; Xu et al., 2019) using PSNR, SSIM, and SAM (Wang et al., 2004; Kruse et al., 1993). Preprocessing, degradations, and optimization settings are detailed in Appendices J and K.

4.2 MAIN RESULTS

Tables 1–3 show that VITTA achieves consistently strong performance across all three restoration tasks, validating the effectiveness of the proposed unified framework for diverse HSI degradations. Figures 3–4 illustrate restoration on synthetic and real data, which demonstrates that VITTA preserves spatial structures and spectral details while effectively suppressing complex degradations. These consistent improvements can be attributed to the coordinated interaction between reliability-aware corruption modeling and Temporal Tucker-regularized diffusion sampling, which enables VITTA to adapt to complex degradations while preserving image fidelity. More results such as spectra visualization and complexity comparison are provided in Appendix L.

Table 1: Quantitative comparison for complex-noise HSI denoising.

Method	Chikusei			Pavia University			Houston 2018		
	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓
Supervised methods									
SST (2023a)	29.288	0.8655	0.1533	28.906	0.8631	0.2111	29.501	0.8156	0.2652
SERT (2023b)	29.537	0.8642	0.1485	28.934	0.8581	0.2077	29.730	0.8278	0.2544
MP-HSIR (2025)	29.493	0.8405	0.1358	29.127	0.8537	0.1595	30.174	0.8320	0.2199
Unsupervised methods									
NGMeet (2019)	27.073	0.7725	0.1763	25.189	0.7622	0.2368	25.541	0.6185	0.3767
HLRTF (2022)	32.111	0.8463	0.0700	30.620	0.8303	0.1143	33.373	0.8347	0.1483
DDS2M (2023)	30.013	0.8408	0.1404	28.204	0.8255	0.2206	29.292	0.7635	0.3668
HIR-Diff (2024)	27.407	0.8004	0.1429	25.287	0.7749	0.2177	25.959	0.6433	0.3750
ATNN (2025)	30.554	0.8016	0.1045	29.800	0.7883	0.1338	31.898	0.7985	0.1650
GNRSR (2025)	32.446	0.8664	0.0731	31.338	0.8574	0.1089	32.972	0.8420	0.1552
LMELTV (2026)	31.073	0.8222	0.1463	30.475	0.8179	0.2901	32.778	0.8294	0.3424
Ours	**33.690**	**0.9084**	**0.0614**	**31.517**	**0.8871**	**0.1062**	**34.208**	**0.8779**	**0.1194**

Table 2: Quantitative comparison for missing-band completion.

Method	Chikusei			Pavia University			Houston 2018		
	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓
DIP2D (2018)	23.382	0.6311	0.5431	22.495	0.6315	0.5575	22.979	0.5715	0.7259
DIP3D (2018)	26.317	0.6844	0.1712	25.175	0.6790	0.2363	25.810	0.5870	0.3807
HLRTF (2022)	31.053	0.8253	0.0784	29.925	0.8088	0.1188	32.440	0.8052	0.1662
DDS2M (2023)	27.628	0.7624	0.1969	25.982	0.7307	0.2642	27.784	0.7123	0.4251
ATNN (2025)	29.645	0.7813	0.0879	27.891	0.7179	0.1694	30.731	0.7406	0.1971
GNRSR (2025)	31.311	0.8338	0.0789	29.979	0.8132	0.1248	31.818	0.8017	0.1697
ScoreTR (2026c)	27.774	0.7492	0.2092	27.448	0.7589	0.2003	28.211	0.7001	0.2643
3DeepRep (2026)	24.127	0.5386	0.4487	23.432	0.5614	0.5154	24.920	0.5147	0.5949
LMELTV (2026)	30.650	0.8136	0.2206	29.108	0.7724	0.1250	30.926	0.7684	0.3524
SHARE (2026)	24.878	0.5658	0.4779	24.680	0.6197	0.5389	23.730	0.4359	0.8533
Ours	**33.148**	**0.8981**	**0.0648**	**31.045**	**0.8742**	**0.0964**	**33.558**	**0.8587**	**0.1375**

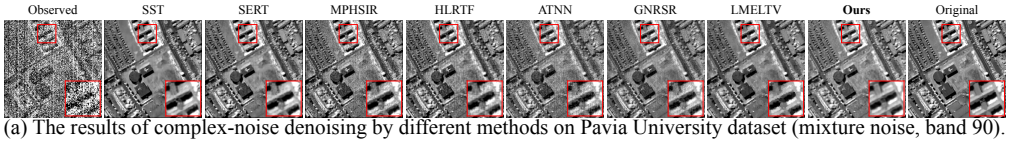
Table 3: Quantitative comparison for noisy pansharpening.

Method	Chikusei			Pavia University			Houston 2018		
	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓
SERT→PLRD (2023b; 2024)	26.559	0.6814	0.1588	27.384	0.7928	0.1901	29.803	0.7921	0.2622
SERT→RPNN (2023b; 2024)	23.644	0.6321	0.1836	23.994	0.6913	0.2245	27.124	0.7499	**0.2570**
MP-HSIR→PLRD (2025; 2024)	25.636	0.6391	0.1584	26.252	0.7288	0.2007	29.428	0.7796	0.2689
MP-HSIR→RPNN (2025; 2024)	23.604	0.6267	0.1788	24.176	0.7050	0.2078	27.216	0.7523	0.2663
PLRD→SERT (2024; 2023b)	22.363	0.6568	0.2488	23.121	0.7375	0.2990	25.862	0.7652	0.3533
PLRD→MP-HSIR (2024; 2025)	22.265	0.6555	0.2592	22.919	0.7349	0.3173	25.732	0.7598	0.3549
RPNN→SERT (2024; 2023b)	22.660	0.4641	0.2998	23.665	0.6245	0.2968	26.824	0.7260	0.3321
RPNN→MP-HSIR (2024; 2025)	21.705	0.4029	0.3228	22.664	0.5712	0.3316	26.489	0.6968	0.3516
Hipandas (2025)	26.934	0.7403	0.1780	27.437	0.8037	0.2206	29.721	0.7935	0.3180
UCL (2026)	25.352	0.7273	0.2171	24.663	0.7638	0.2930	27.248	0.7831	0.3726
Ours	**29.139**	**0.8080**	**0.1193**	**28.398**	**0.8074**	**0.1711**	**31.393**	**0.8161**	0.2602

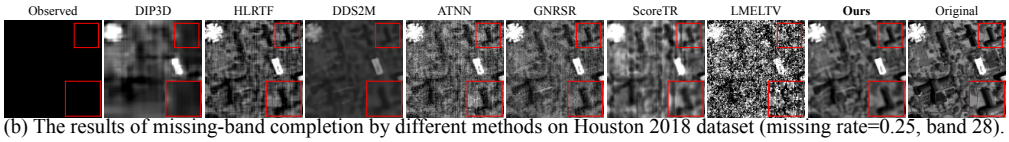
4.3 ABLATION RESULTS

Table 4 shows that each proposed module contributes to the overall restoration capability, while their combination leads to consistently improved performance across different tasks. Specifically, the two-factor noise model provides a more faithful characterization of heterogeneous corruption, thereby enabling reliability-aware observation conditioning and reducing the impact of unreliable

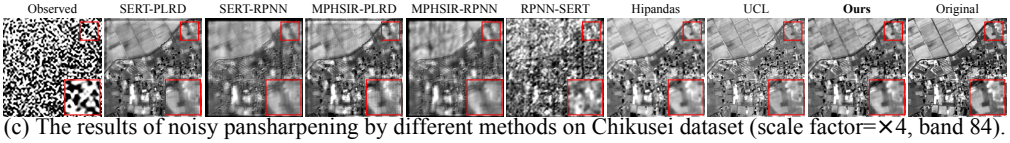
measurements during reverse diffusion. Building on this, the Tucker representation and temporal coupling exploit the intrinsic spatial–spectral structure of HSIs and the dependencies across reverse steps, which further stabilizes clean-endpoint estimation. Consistently, Table 5 shows that the introduced loss terms provide complementary regularization for endpoint estimation.



(a) The results of complex-noise denoising by different methods on Pavia University dataset (mixture noise, band 90).

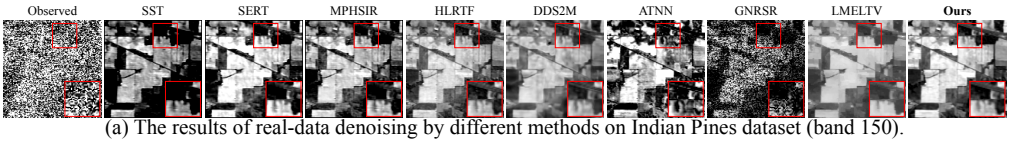


(b) The results of missing-band completion by different methods on Houston 2018 dataset (missing rate=0.25, band 28).

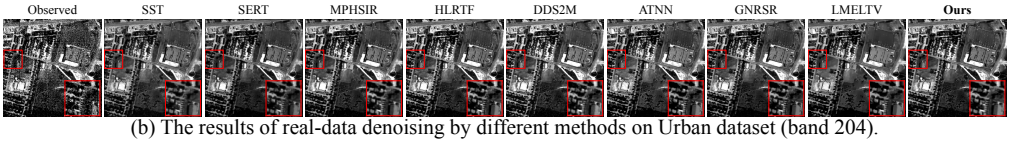


(c) The results of noisy pansharpening by different methods on Chikusei dataset (scale factor=×4, band 84).

Figure 3: Visual comparisons on three representative HSI restoration tasks.



(a) The results of real-data denoising by different methods on Indian Pines dataset (band 150).



(b) The results of real-data denoising by different methods on Urban dataset (band 204).

Figure 4: Visual comparison for real HSI restoration.

Table 4: Ablation study of key components, with results averaged over three datasets. "Ours w/ DIP" retains the full model but replaces the diffusion-based recovery process with Deep Image Prior (DIP).

	Model configuration				Denoising			Band missing			Pansharpening		
Variant	Two-factor noise	Tucker	Temporal	Recovery	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓
Baseline (DDS2M)	–	–	–	Diffusion	29.170	0.8099	0.2426	27.131	0.7351	0.2954	24.567	0.6056	0.4111
Ours w/o Tucker	✓	–	–	Diffusion	32.340	0.8735	0.1053	31.610	0.8550	0.1067	28.773	0.7839	0.2002
Ours w/o Temporal Coupling	✓	✓	–	Diffusion	32.784	0.8835	0.0966	32.426	0.8754	0.1027	29.034	0.7971	0.1994
Ours w/ DIP	✓	✓	✓	DIP	32.438	0.8824	0.1095	28.528	0.6863	0.2305	28.394	0.7627	0.2230
Ours	✓	✓	✓	Diffusion	**33.138**	**0.8911**	**0.0957**	**32.584**	**0.8770**	**0.0996**	**29.643**	**0.8105**	**0.1836**

Table 5: Ablation study of loss terms, with results averaged over three datasets.

Loss terms				Denoising			Band missing			Pansharpening		
$\mathcal{L}_{\text{diff}}$	$\mathcal{L}_{\text{spec}}$	$\mathcal{L}_{\text{rel}}$	$\mathcal{L}_{\text{traj}}$	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓
✓				31.939	0.8740	0.1431	29.386	0.7731	0.2165	28.174	0.7779	0.2434
✓	✓			32.068	0.8779	0.1385	29.878	0.8058	0.1929	28.487	0.7800	0.2315
✓	✓	✓		32.571	0.8824	0.1010	32.426	0.8753	0.1030	28.489	0.8069	0.1897
✓	✓	✓	✓	**33.138**	**0.8911**	**0.0957**	**32.584**	**0.8770**	**0.0996**	**29.643**	**0.8105**	**0.1836**

5 Conclusion

In this work, we proposed VITTA, a self-supervised variational diffusion framework for HSI restoration under complex degradations. VITTA models band-dependent dense noise and sparse heavy-tailed corruption through variational inference, and converts the inferred uncertainty into reliability-aware Gaussian observation channels for conditional diffusion sampling. Meanwhile, the proposed Temporal Tucker Prior Module exploits spatial, spectral, and reverse-time dependencies to provide scene-specific clean-endpoint estimation throughout the restoration trajectory. Our analysis further shows that Gaussian projection preserves the dominant second-order information used by diffusion sampling while controlling the remaining approximation error, and explains how reliability-guided sampling and endpoint fitting jointly drive progressive restoration. Extensive experiments on multiple restoration tasks demonstrate the effectiveness of the proposed framework.

AI USE STATEMENT

The authors conducted the literature review, developed the research ideas, algorithmic design and mathematical reasoning, prepared the manuscript outline and portions of the text, and created the figures. Generative AI tools were used to implement the proposed method and to assist with mathematical reasoning and the verification of derivations. Additionally, AI tools supported manuscript refinement. The authors reviewed the AI-assisted implementation, mathematical arguments, and text for correctness and consistency with the proposed framework. The authors take full responsibility for the final content of this work, including its text, claims, results, figures, and accompanying code.

ETHICS STATEMENT

This work focuses on the methodological and empirical study of hyperspectral image restoration using established research datasets. We have identified no specific ethical concerns arising from the research described in this paper.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide the methodological and experimental details needed to reproduce our results. Specifically, Section 3 presents the proposed algorithm and theoretical results, with detailed derivations, assumptions, and proofs provided in Appendices C–I. Section 4 describes the evaluation tasks, datasets, and metrics, while Appendices J and K document data preprocessing, degradation synthesis, hyperparameters, optimization schedules, and random seeds. In addition, we have submitted the implementation code as supplementary material to support independent reproduction and verification of the reported results.

REFERENCES

David F. Andrews and Colin L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B*, 36(1):99–102, 1974.

José M. Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, 2012.

José M. Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser M. Nasrabadi, and Jocelyn Chanussot. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and Remote Sensing Magazine*, 1(2):6–36, 2013. doi: 10.1109/MGRS.2013.2244672.

Ching-Heng Cheng, Chia-Ming Lee, Wo-Ting Liao, Yu-Fan Lin, Yi-Ching Cheng, Fu-En Yang, Yu-Chiang Frank Wang, and Chih-Chung Hsu. PromptHSI: Universal hyperspectral image restoration with vision–language modulated frequency adaptation. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–16, 2026a. doi: 10.1109/TGRS.2026.3660506. Art. no. 5503516.

Jun Cheng and Shan Tan. Diffusion priors for variational likelihood estimation and image denoising. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1954.

Jun Cheng, Tao Liu, and Shan Tan. Score priors guided deep variational inference for unsupervised real-world single image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12937–12948, 2023.

Libo Cheng, Wenlin Du, Zhe Li, and Xiaoning Jia. A self-supervised conditional diffusion network with multihead wavelet attention for remote sensing image denoising. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19:7498–7518, 2026b. doi: 10.1109/JSTARS.2026.3664428.

Zhengyun Cheng, Changhao Wang, Guanwen Zhang, Yi Xu, Wei Zhou, and Xiangyang Ji. Score-based model for low-rank tensor recovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20527–20535, 2026c. doi: 10.1609/aaai.v40i25.39188.

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

Laura M. Dale, André Thewis, Christelle Boudry, Ioan Rotar, Pierre Dardenne, Vincent Baeten, and Juan A. Fernández Pierna. Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review. *Applied Spectroscopy Reviews*, 48(2):142–159, 2013. doi: 10.1080/05704928.2012.705800.

Haiyan Fan, Yunjin Chen, Yulan Guo, Hongyan Zhang, and Gangyao Kuang. Hyperspectral image restoration using low-rank tensor recovery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(10):4589–4604, 2017.

Megandhren Govender, K. Chetty, and H. Bulcock. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water SA*, 33(2):145–151, 2007. doi: 10.4314/wsa.v33i2.49049.

Giuseppe Guarino, Matteo Ciotola, Gemine Vivone, and Giuseppe Scarpa. Band-wise hyperspectral image pansharpening using CNN model propagation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5500518, 2024.

Wei He, Quanming Yao, Chao Li, Naoto Yokoya, and Qibin Zhao. Non-local meets global: An integrated paradigm for hyperspectral denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6868–6877, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Dong Hu, Yusen Tan, Yong Wang, Tao Jia, and Zhi Wang. Hyperspectral image denoising via enhanced laplacian total variation regularizer. *Expert Systems with Applications*, 322:132332, 2026. doi: 10.1016/j.eswa.2026.132332.

Xin Huang and Liangpei Zhang. A comparative study of spatial approaches for urban mapping using hyperspectral ROSIS images over pavia city, northern italy. *International Journal of Remote Sensing*, 30(12):3205–3221, 2009. doi: 10.1080/01431160802559046.

Bahjat Kawar, Gregory Vaksman, and Michael Elad. SNIPS: Solving noisy inverse problems stochastically. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21757–21769, 2021.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23593–23606, 2022.

Fred A. Kruse, A. B. Lefkoff, Joseph W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and Alexander F. H. Goetz. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, 44(2–3):145–163, 1993. doi: 10.1016/0034-4257(93)90013-N.

Ronan Legin, Alexandre Adam, Yashar Hezaveh, and Laurence Perreault-Levasseur. Beyond gaussian noise: A generalized approach to likelihood analysis with non-gaussian noise. *The Astrophysical Journal Letters*, 949(2):L41, 2023. doi: 10.3847/2041-8213/acd645.

Miaoyu Li, Ying Fu, and Yulun Zhang. Spatial-spectral transformer for hyperspectral image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1368–1376, 2023a.

Miaoyu Li, Ji Liu, Ying Fu, Yulun Zhang, and Dejing Dou. Spectral enhanced rectangle transformer for hyperspectral image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5805–5814, 2023b.

Yunshan Li, Wenwu Gong, Qianqian Wang, Chao Wang, and Lili Yang. 3deeprep: 3d deep low-rank tensor representation for hyperspectral image inpainting. *Neurocomputing*, 678:133160, 2026. doi: 10.1016/j.neucom.2026.133160.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Yisi Luo, Xi-Le Zhao, Deyu Meng, and Tai-Xiang Jiang. Hlrtf: Hierarchical low-rank tensor factorization for inverse problems in multi-dimensional imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19303–19312, June 2022.

Yu-Chun Miao, Xi-Le Zhao, Xiao Fu, Jian-Li Wang, and Yu-Bang Zheng. Hyperspectral denoising using unsupervised disentangled spatiospectral deep priors. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

Yuchun Miao, Lefei Zhang, Liangpei Zhang, and Dacheng Tao. Dds2m: Self-supervised denoising diffusion spatio-spectral model for hyperspectral image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12086–12096, October 2023.

Li Pang, Xiangyu Rui, Long Cui, Hongzhong Wang, Deyu Meng, and Xiangyong Cao. HIR-Diff: Unsupervised hyperspectral image restoration via improved diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3005–3014, 2024.

Jiangjun Peng, Qi Xie, Qian Zhao, Yao Wang, Yee Leung, and Deyu Meng. Enhanced 3dtv regularization and its applications on hsi denoising and compressed sensing. *IEEE Transactions on Image Processing*, 29:7889–7903, 2020.

Jiangjun Peng, Hailin Wang, Xiangyong Cao, and Shuang Xu. Fast guaranteed tensor recovery with adaptive tensor nuclear norm. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6057–6065, 2025. doi: 10.24963/ijcai.2025/674.

Xiangyu Rui, Xiangyong Cao, Li Pang, Zeyu Zhu, Zongsheng Yue, and Deyu Meng. Unsupervised hyperspectral pansharpening via low-rank diffusion model. *Information Fusion*, 107:102325, 2024.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. doi: 10.1109/TPAMI.2022.3204461.

Oleksii Sidorov and Jon Yngve Hardeberg. Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

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

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

Tristan S. W. Stevens, Hans van Gorp, Faik C. Meral, Junseob Shin, Jason Yu, Jean-Luc Robert, and Ruud J. G. van Sloun. Removing structured noise using diffusion models. *Transactions on Machine Learning Research*, 2025(3), 2025.

Yusen Tan, Dong Hu, Yao Fu, Chunming Yang, and Zhi Wang. Generalized nonconvex low-rank approximation and sparse regularizer for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–17, 2025. doi: 10.1109/TGRS.2025.3643597. Art. no. 5536617.

Haoyue Tang, Tian Xie, Aosong Feng, Hanyu Wang, Chenyang Zhang, and Yang Bai. Solving general noisy inverse problem via posterior sampling: A policy gradient viewpoint. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2116–2124. PMLR, 2024.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

Lucien Wald, Thierry Ranchin, and Michel Marc. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63(6):691–699, 1997.

Yao Wang, Jiangjun Peng, Qian Zhao, Yee Leung, Xi-Le Zhao, and Deyu Meng. Hyperspectral image restoration via total variation regularized low-rank tensor decomposition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(4):1227–1243, 2018.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023.

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

Zhehui Wu, Yong Chen, Naoto Yokoya, and Wei He. MP-HSIR: A multi-prompt framework for universal hyperspectral image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13009–13020, 2025.

Jin-Liang Xiao, Ting-Zhu Huang, Liang-Jian Deng, Huidong Jiang, Qibin Zhao, and Gemine Vivone. Unsupervised coefficient learning framework for variational pansharpening. *Information Fusion*, 127:103790, 2026. ISSN 1566-2535. doi: 10.1016/j.inffus.2025.103790.

Jiangwei Xie, Zhang Wen, Mike Davies, and Dongdong Chen. Share: A fully unsupervised framework for single hyperspectral image restoration. *arXiv preprint arXiv:2601.13987*, 2026.

Fengchao Xiong, Jun Zhou, and Yuntao Qian. Hyperspectral restoration via ℓ_0 gradient regularized low-rank tensor factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12): 10410–10425, 2019.

Shuang Xu, Zixiang Zhao, Haowen Bai, Chang Yu, Jiangjun Peng, Xiangyong Cao, and Deyu Meng. Hipandas: Hyperspectral image joint denoising and super-resolution by image fusion with the panchromatic image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12002–12011, 2025.

Yonghao Xu, Bo Du, Liangpei Zhang, Daniele Cerra, Miguel Pato, Emiliano Carmona, Saurabh Prasad, Naoto Yokoya, Ronny Hänsch, and Bertrand Le Saux. Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(6):1709–1724, 2019. doi: 10.1109/JSTARS.2019.2911113.

Naoto Yokoya and Akira Iwasaki. Airborne hyperspectral data over chikusei. *Space Application Laboratory, University of Tokyo, Technical Report SAL-2016-05-27*, 2016.

Naoto Yokoya, Claas Grohnfeldt, and Jocelyn Chanussot. Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, 5(2):29–56, 2017. doi: 10.1109/MGRS.2016.2637824.

Haijin Zeng, Jiezhang Cao, Kai Zhang, Yongyong Chen, Hiep Luong, and Wilfried Philips. Unmixing diffusion for self-supervised hyperspectral image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27820–27830, 2024.

Hongyan Zhang, Wei He, Liangpei Zhang, Huanfeng Shen, and Qiangqiang Yuan. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4729–4743, 2014.

A Related Work

A.1 Diffusion Models for Image Restoration

Diffusion models support image restoration through task-specific conditional generation and measurement-conditioned sampling with generative priors (Ho et al., 2020; Song et al., 2021a;b; Saharia et al., 2023; Lugmayr et al., 2022). For linear inverse problems, SNIPS, DDRM, and DDNM incorporate observations through analytical transitions or subspace-based conditioning (Kawar et al., 2021; 2022; Wang et al., 2023). Building on this connection between generative modeling and measurement consistency, DPS and DPG develop posterior guidance for more general likelihoods (Chung et al., 2023; Tang et al., 2024).

A complementary line of work adapts the observation model to complex corruption. ScoreDVI combines score priors with variational inference and a non-i.i.d. Gaussian mixture noise model (Cheng et al., 2023), while DiffusionVI estimates noise precision during reverse diffusion (Cheng & Tan, 2024). Other methods learn non-Gaussian or structured likelihoods from noise data (Legin et al., 2023; Stevens et al., 2025). Together, these approaches establish observation modeling as an important component of diffusion-based restoration. VITTA develops this component for hyperspectral measurements: its two-factor variational model estimates band-dependent dense noise and sparse heavy-tailed corruption from the observed scene, and its moment-matched Gaussian projection supplies channel-wise reliability to the conditional reverse transition.

A.2 Hyperspectral Image Restoration

HSI restoration combines spatial reconstruction with preservation of spectral signatures. Model-based approaches encode this structure through low-rank, total-variation, and related regularizers (Zhang et al., 2014; Fan et al., 2017; Wang et al., 2018; Xiong et al., 2019; Peng et al., 2020; Luo et al., 2022). Supervised CNN and transformer models learn restoration mappings from training data (Li et al., 2023a;b; Wu et al., 2025; Cheng et al., 2026a). In parallel, self-supervised and untrained methods fit scene-specific representations directly to degraded observations, using network structure as an image prior (Ulyanov et al., 2018; Sidorov & Hardeberg, 2019; Miao et al., 2022).

Diffusion-based HSI methods further combine iterative sampling with spatial–spectral modeling. DDS2M performs scene-specific restoration with spatial–spectral networks (Miao et al., 2023); HIR-Diff transfers a pretrained diffusion prior through a low-rank spectral representation (Pang et al., 2024); and Diff-Unmix combines spectral unmixing with self-supervised diffusion for denoising (Zeng et al., 2024). For complex-noise HSI denoising, DMWAN constructs pseudo-supervised pairs from degraded observations to support self-supervised conditional diffusion (Cheng et al., 2026b).

VITTA brings scene-specific corruption inference and endpoint modeling into the same reverse process. On the observation side, variational state and scale estimates determine the reliability of each Gaussian channel. On the prior side, TTPM couples spatial generators, shared spectral factors, and a low-dimensional temporal core trajectory. The conditional sampler then combines these two sources of information, yielding a shared restoration mechanism for denoising, missing-band completion, and pansharpening.

B Complete Algorithmic Flow

Following Section 3, Algorithm 1 summarizes the full restoration procedure.

C Conditional Diffusion ELBO and Endpoint Estimation

Following Section 2, this section derives the conditional-path ELBO and connects Gaussian endpoint substitution to the fitting loss in Eq. (5). We first establish the path factorization and then describe how the current diffusion state supplies the endpoint-fitting signal.

Algorithm 1 VITTA for self-supervised HSI restoration

Require: Degraded observation y, degradation operator H, optional PAN observation, selected reverse grid $1 \le \tau_1 < \cdots < \tau_K \le T$ with $K \ge 3$, TTPM ranks, and prescribed inference, optimization, and averaging settings.

Ensure: Restored HSI $\widehat{X}$.

1: Initialize the TTPM parameters, fixed spatial inputs, and the temporal dictionary for K fitting stages; initialize any task-specific parameters.
2: Initialize the running endpoint average from an observation-based reconstruction and initialize u_{τ_K} using the task-specific observation-conditioned initialization.
3: Initialize the corruption statistics from the acquired measurements, and construct the initial projected precisions, confidence weights, and Gaussian observation channels.
4: **for** $j = K, K-1, \ldots, 1$ **do**
5: Set the TTPM temporal coordinate and scheduled loss weights for stage j.
6: **if** $j < K$ and stage j is scheduled for observation refresh **then**
7: Obtain a detached radiometric reference $\widetilde{x}_{\tau_j}$ from a provisional TTPM update.
8: Update the variational corruption distribution and nuisance parameters from $y - H\widetilde{x}_{\tau_j}$ using Eq. (9) and Appendix D.
9: Compute the projected variances $S^{\mathrm{G}}_{i,\tau_j}$ and precisions $\Lambda_{i,\tau_j} = (S^{\mathrm{G}}_{i,\tau_j})^{-1}$ using Eq. (12).
10: Compute the detached confidence $R_{i,\tau_j} = \Lambda_{i,\tau_j}/\tau_{b(i)}$ and its image-grid weights as described in Appendix E.
11: Refresh the Gaussian observation channels from y, H, and Λ_{τ_j} using Eq. (13).
12: **else**
13: Reuse the most recent corruption statistics, confidence weights, and Gaussian observation channels.
14: **end if**
15: **for** each prescribed endpoint-fitting update at stage j **do**
16: Take an optimization step for Eq. (17), keeping u_{τ_j} and the inferred corruption statistics fixed, and evaluate the fitted endpoint $\widehat{u}_{\tau_j}$.
17: Update the running endpoint average $\overline{u}_{\tau_j}$ using $\widehat{u}_{\tau_j}$ as described in Appendix E.
18: **end for**
19: **if** $j > 1$ **then**
20: Sample $u_{\tau_{j-1}}$ using $\overline{u}_{\tau_j}$ and the current Gaussian observation channels according to Eq. (18), with target variance $\sigma^2_{\tau_{j-1}}$.
21: **end if**
22: **end for**
23: Return $\widehat{X}$ by clipping $(\overline{u}_{\tau_1} + \mathbf{1}_n)/2$ to $[0,1]$ and reshaping it to the HSI tensor.

C.1 Forward Coordinates and Conditional Path Densities

Throughout the appendix, $x_0 = \mathrm{vec}(X_0)$ denotes the radiometric image, $u_0 = 2x_0 - \mathbf{1}_n$ its normalized clean endpoint, and u_t the diffusion state at timestep t. Thus Eq. (2) gives

$$u_t = \sqrt{\bar{\alpha}_t}\, u_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_n). \tag{25}$$

It is convenient to express the same state on the clean-image scale:

$$v_t := \frac{u_t}{\sqrt{\bar{\alpha}_t}}, \qquad \sigma_t^2 := \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}, \qquad v_t \mid u_0 \sim \mathcal{N}(u_0, \sigma_t^2 I_n) \quad \text{under } q^{\mathrm{fwd}}. \tag{26}$$

The reverse process proceeds from $t = T$ to $t = 1$. In these coordinates, the endpoint loss becomes $\mathcal{L}_{\mathrm{diff},t} = \|v_t - \widehat{u}_t(\Theta)\|^2/(n\sigma_t^2)$.

For fixed y, H and an observation-model snapshot, the reference and generative path densities factor as

$$q(u_{1:T} \mid u_0, y, H) = q_T(u_T \mid u_0, y, H) \prod_{t=2}^{T} q_t(u_{t-1} \mid u_t, u_0, y, H), \tag{27}$$

$$p_\Theta(u_0, u_{1:T} \mid y, H) = p_T(u_T \mid y, H) \prod_{t=2}^{T} p_{\Theta,t}(u_{t-1} \mid u_t, y, H) p_{\Theta,0}(u_0 \mid u_1, y, H). \tag{28}$$

Here $p_{\Theta,0}$ is a normalized reconstruction density. For example, a Gaussian decoder with mean $F_{\Theta,0}(u_1, y)$ and covariance $\epsilon_{\mathrm{dec}}^2 I_n$, $\epsilon_{\mathrm{dec}} > 0$, provides such a density. The restoration output is the final averaged endpoint, as detailed in Appendix F.4.

Assuming common support for the change of measure and existence of the displayed expectations, Jensen's inequality yields

$$\begin{aligned}\log p_\Theta(u_0 \mid y, H) &= \log \mathbb{E}_q \frac{p_\Theta(u_0, u_{1:T} \mid y, H)}{q(u_{1:T} \mid u_0, y, H)} \\ &\geq \mathbb{E}_q \log \frac{p_\Theta(u_0, u_{1:T} \mid y, H)}{q(u_{1:T} \mid u_0, y, H)} = \mathcal{E}_{\mathrm{diff}}(\Theta, q). \end{aligned} \tag{29}$$

Substituting Eq. (28) gives

$$\begin{aligned}\mathcal{E}_{\mathrm{diff}}(\Theta, q) = \mathbb{E}_q \log p_{\Theta,0}(u_0 \mid u_1, y, H) - D_{\mathrm{KL}}(q_T \| p_T) \\ - \sum_{t=2}^{T} \mathbb{E}_q D_{\mathrm{KL}}(q_t \| p_{\Theta,t}),\end{aligned} \tag{30}$$

which recovers Eq. (4). The expectations in the sum are taken over the reference conditioning states. Moreover, the exact gap is

$$\log p_\Theta(u_0 \mid y, H) - \mathcal{E}_{\mathrm{diff}}(\Theta, q) = D_{\mathrm{KL}}(q(u_{1:T} \mid u_0, y, H) \| p_\Theta(u_{1:T} \mid u_0, y, H)) \,.$$

Consequently, with q and p_T fixed, the parameter-dependent terms comprise the reconstruction likelihood and the transition discrepancies. The same factorization extends to a selected time grid. For adaptive updates, the conditioning variable additionally contains the preceding reverse history, as formalized in Appendix I.5.

C.2 Endpoint Substitution and the Observable Fitting Loss

DDS2M places a scene-specific clean-endpoint estimate in its conditional transitions (Miao et al., 2023). At a fixed conditioning state, consider the Gaussian pair

$$q_t = \mathcal{N}(\mathsf{B}_t u_0 + \boldsymbol{d}_t, \mathsf{C}_t), \qquad p_{\Theta,t} = \mathcal{N}(\mathsf{B}_t \widehat{u}_t(\Theta) + \boldsymbol{d}_t, \mathsf{C}_t), \qquad \mathsf{C}_t \succ 0.$$

The symbols B_t, $\boldsymbol{d}_t$, and C_t denote the endpoint map, common offset, and transition covariance, respectively. In particular, C_t is distinct from the singular-value matrix Σ_t in Eq. (13). The Gaussian KL formula gives

$$D_{\mathrm{KL}}(q_t \| p_{\Theta,t}) = \tfrac{1}{2}(\widehat{u}_t - u_0)^\top \mathsf{B}_t^\top \mathsf{C}_t^{-1} \mathsf{B}_t (\widehat{u}_t - u_0). \tag{31}$$

Thus the transition objective measures endpoint error in the directions weighted by the conditional kernel.

For scene-specific fitting, the current diffusion state supplies an observable target. For a candidate endpoint F, the forward Gaussian density satisfies

$$-\frac{2}{n} \log q^{\mathrm{fwd}}(u_t \mid F) = \frac{\|u_t - \sqrt{\bar{\alpha}_t} F\|^2}{n(1 - \bar{\alpha}_t)} + \log(2\pi(1 - \bar{\alpha}_t)) \,. \tag{32}$$

Since the final term is constant in F, this scaled negative log density yields the observable surrogate $\mathcal{L}_{\mathrm{diff},t}$ in Eq. (5). Adding the structural and physical regularizers gives the endpoint objective $\mathcal{J}_t$ in Eq. (17).

The relation to clean-endpoint accuracy can also be expressed directly. For a fixed candidate F and a reference draw $v = u_0 + \sigma\epsilon$, $\epsilon \sim \mathcal{N}(0, I_n)$,

$$\mathbb{E} \frac{\|v - F\|^2}{n\sigma^2} = 1 + \frac{\|F - u_0\|^2}{n\sigma^2}.$$

For a fitted map $F(v)$, expansion of the same square gives

$$\mathbb{E} \frac{\|v - F(v)\|^2}{n\sigma^2} = 1 + \frac{\mathbb{E}\|F(v) - u_0\|^2}{n\sigma^2} - \frac{2\mathbb{E}[\epsilon^\top (F(v) - u_0)]}{n\sigma}. \tag{33}$$

The final term quantifies the interaction between the sampled state and its fitted endpoint. Accordingly, Appendix H derives the objective induced by the conditional sampler, and Appendix I tracks the endpoint produced by finite, adaptive fitting.

D LIKELIHOOD STEP: VARIATIONAL UPDATES AND GAUSSIAN PROJECTION

This section expands the likelihood step in Section 3.2. We first derive the latent coordinate updates, then give the conjugate parameter means and the moments used by Gaussian projection. Each inference block conditions on the detached radiometric reference $\widetilde{x}_t$. Gamma distributions use shape–rate parameters, and $\mathrm{InvGamma}(a, b)$ has density proportional to $v^{-a-1}\exp(-b/v)$.

D.1 LATENT POSTERIOR AND VARIATIONAL OBJECTIVE

Let $e_{i,t}^{\mathrm{obs}} = y_i - h_i^\top \widetilde{x}_t$ and $b = b(i)$, as in the main text. Conditional on the sparse state,

$$p(y_i, s_i, \omega_i, c_i = s \mid H\widetilde{x}_t, \psi) = \pi_s \mathcal{N}(e_{i,t}^{\mathrm{obs}} \mid s_i, \tau_b^{-1})\mathcal{N}(s_i \mid 0, v_b/\omega_i)\mathrm{Gamma}(\omega_i \mid \nu/2, \nu/2).$$

The ordinary state contributes $\pi_n \mathcal{N}(e_{i,t}^{\mathrm{obs}} \mid 0, \tau_b^{-1})$ with $s_i = 0$. Integrating out the prior scale gives a Student-t sparse amplitude with degrees of freedom ν and scale $\sqrt{v_b}$. The variational family factorizes across measurements and, within the sparse state, as

$$q(s_i, \omega_i \mid c_i = s) = q(s_i \mid c_i = s)q(\omega_i \mid c_i = s).$$

The ordinary-state amplitude is represented by its point mass at zero.

Bayes' rule yields

$$\begin{aligned}\log p_\psi(y \mid H\widetilde{x}_t) + \log p(\psi) = \mathcal{L}_{\mathrm{obs}}(q, \psi; \widetilde{x}_t) \\ + D_{\mathrm{KL}}(q(\zeta) \| p_\psi(\zeta \mid y, H\widetilde{x}_t))\,.\end{aligned} \tag{34}$$

Therefore, for fixed ψ and $\widetilde{x}_t$, the posterior-KL objective in Eq. (8) is equivalent to maximizing

$$\mathcal{L}_{\mathrm{obs}} = \sum_i \sum_{j \in \{n,s\}} r_{ij}(\log \pi_j + A_{ij} - \log r_{ij}) + \log p(\psi), \tag{35}$$

where $r_{in} = 1 - r_{is}$ and the state evidences A_{ij} are specified below. A prescribed sequence of coordinate updates produces the stagewise approximation q_t^{obs}.

D.2 AMPLITUDE, SCALE, AND STATE UPDATES

With the complementary factors fixed, the amplitude-dependent expected log density is

$$-\tfrac{\tau_b}{2}(e_{i,t}^{\mathrm{obs}} - s_i)^2 - \tfrac{\bar{\omega}_i}{2v_b}s_i^2 + \mathrm{const.}$$

Completing the square gives

$$q(s_i \mid c_i = s) = \mathcal{N}(m_i, V_i), \qquad V_i = (\tau_b + \bar{\omega}_i/v_b)^{-1}, \qquad m_i = V_i \tau_b e_{i,t}^{\mathrm{obs}}.$$

Similarly, the scale-dependent terms are

$$\left(\frac{\nu+1}{2} - 1\right)\log\omega_i - \frac{1}{2}\left(\nu + \frac{m_i^2 + V_i}{v_b}\right)\omega_i,$$

which yield

$$q(\omega_i \mid c_i = s) = \mathrm{Gamma}(a_\omega, b_{\omega i}), \qquad a_\omega = (\nu+1)/2, \qquad b_{\omega i} = \tfrac{1}{2}\{\nu + (m_i^2 + V_i)/v_b\}.$$

The corresponding scale mean is $\bar{\omega}_i = a_\omega / b_{\omega i}$.

For the state update, define

$$A_{in} = \tfrac{1}{2}(\log\tau_b - \log 2\pi) - \tfrac{\tau_b}{2}(e_{i,t}^{\mathrm{obs}})^2, \tag{36}$$

$$A_{is} = \mathbb{E}\log p(y_i \mid s_i) + \mathbb{E}\log p(s_i \mid \omega_i) + \mathbb{E}\log p(\omega_i) + \mathcal{H}(q(s_i)) + \mathcal{H}(q(\omega_i)). \tag{37}$$

All sparse-state expectations and entropies use the conditional variational factors. Writing $\ell_i^\omega = \mathrm{digamma}(a_\omega) - \log b_{\omega i}$, the terms are

$$\begin{aligned}\mathbb{E}\log p(y_i \mid s_i) &= \tfrac{1}{2}(\log\tau_b - \log 2\pi) - \tfrac{\tau_b}{2}[(e_{i,t}^{\mathrm{obs}} - m_i)^2 + V_i], \\ \mathbb{E}\log p(s_i \mid \omega_i) &= \tfrac{1}{2}[\ell_i^\omega - \log(2\pi v_b)] - \frac{\bar{\omega}_i(m_i^2 + V_i)}{2v_b}, \\ \mathbb{E}\log p(\omega_i) &= \tfrac{\nu}{2}\log(\nu/2) - \log\Gamma(\nu/2) + (\nu/2 - 1)\ell_i^\omega - \tfrac{\nu}{2}\bar{\omega}_i, \\ \mathcal{H}(q(s_i)) &= \tfrac{1}{2}\log(2\pi e V_i), \\ \mathcal{H}(q(\omega_i)) &= a_\omega - \log b_{\omega i} + \log\Gamma(a_\omega) + (1 - a_\omega)\,\mathrm{digamma}(a_\omega).\end{aligned}$$

Maximization over the probability simplex gives $r_{ij} = \mathrm{softmax}_j(\log \pi_j + A_{ij})$, completing the derivation of Eq. (9). Each analytical latent update maximizes its own block; hence their sequential application increases or preserves the fixed-reference, fixed-parameter observation ELBO.

D.3 Conjugate Parameter Means

The parameter priors are

$$\pi \sim \mathrm{Dirichlet}(\kappa_0 \pi_0), \qquad \tau_b \sim \mathrm{Gamma}(a_0, b_0), \qquad v_b \sim \mathrm{InvGamma}(a_0, b_0),$$

where $a_0 > 1$, $b_0 > 0$, $\kappa_0 > 0$, and π_0 has positive entries. Exponentiating the expected complete log density together with each log prior gives an auxiliary conjugate density. Its mean supplies the point parameter used in subsequent latent updates. With

$$D_i = (1 - r_{is})(e_{i,t}^{\mathrm{obs}})^2 + r_{is}[(e_{i,t}^{\mathrm{obs}} - m_i)^2 + V_i],$$

the updates are

$$\pi_j \leftarrow \frac{\kappa_0 \pi_{0j} + \sum_i r_{ij}}{\kappa_0 + m}, \tag{38}$$

$$\tau_b \leftarrow \frac{a_0 + n_b/2}{b_0 + \frac{1}{2}\sum_{i \in \mathcal{I}_b} D_i}, \tag{39}$$

$$a_{v_b} = a_0 + \tfrac{1}{2} \sum_{i \in \mathcal{I}_b} r_{is}, \tag{40}$$

$$b_{v_b} = b_0 + \tfrac{1}{2} \sum_{i \in \mathcal{I}_b} r_{is} \bar{\omega}_i (m_i^2 + V_i), \tag{41}$$

$$v_b \leftarrow b_{v_b}/(a_{v_b} - 1). \tag{42}$$

Here $\mathcal{I}_b = \{i : b(i) = b\}$, $n_b = |\mathcal{I}_b|$, and m is the number of acquired scalar measurements. These conjugate means combine prior information with the inferred sufficient statistics. Accordingly, the inference block alternates latent coordinate optimization with prior-regularized parameter estimation.

D.4 Moments Used by Projection and Endpoint Fitting

The inverse-scale moment needed by the Gaussian projection is

$$\mathbb{E}_q[\omega_i^{-1}] = \frac{b_{\omega i}}{a_\omega - 1}, \qquad a_\omega > 1. \tag{43}$$

Since $a_\omega = (\nu + 1)/2$, the condition $\nu > 1$ ensures a finite auxiliary residual variance. For comparison, the original Student-t prior has finite variance for $\nu > 2$.

The sparse-amplitude posterior also provides

$$\mu_i^s = r_{is} m_i, \qquad (\sigma_i^s)^2 = r_{is}(V_i + m_i^2) - (r_{is} m_i)^2. \tag{44}$$

These moments summarize the amplitude inferred for the acquired observation. In particular, μ_i^s supplies the sparse correction in the auxiliary LR fitting loss of Appendix E.4. Complementarily, the Gaussian observation channel uses the zero-centered residual model defined next.

D.5 Auxiliary Residual Distribution and Gaussian Projection

At timestep t, the density $f_{i,t}^q$ in Eq. (10) averages the zero-centered conditional noise model over the inferred state and scale. Equivalently, its analytical construction draws a state from $q(c_i)$, a scale from $q(\omega_i \mid c_i = s)$ in the sparse state, and a residual from the corresponding Gaussian component. The implementation evaluates the required moments in closed form. For $E_i \sim f_{i,t}^q$,

$$\mathbb{E}E_i = 0, \qquad \mathrm{Var}(E_i) = (1 - r_{is}) s_{n,i} + r_{is} \mathbb{E}_q s_i(\omega) = S_{i,t}^{\mathrm{G}}, \tag{45}$$

where $s_{n,i} = \tau_{b(i)}^{-1}$ and $s_i(\omega) = s_{n,i} + v_{b(i)}/\omega$. Thus the inferred state and scale adapt the observation uncertainty while preserving the zero-centered corruption model.

Under finite positive scales and $\nu > 1$, the density is bounded above because $s_i(\omega) \geq s_{n,i} > 0$. Its entropy is bounded above by the entropy of the variance-matched Gaussian and below by the density bound. Consequently, its entropy and its KL divergence to that Gaussian are finite.

Derivation of the Gaussian projection in Eq. (12). For $S > 0$ and $\mu \in \mathbb{R}$,

$$-\mathbb{E} \log \mathcal{N}(E_i \mid \mu, S) = \tfrac{1}{2} \log(2\pi S) + \frac{S_{i,t}^{\mathrm{G}} + \mu^2}{2S}.$$

Subtracting the value at $(\mu, S) = (0, S_{i,t}^{\mathrm{G}})$ yields

$$D_{\mathrm{KL}}(f_{i,t}^q \| \mathcal{N}(\mu, S)) = \varepsilon_{i,t} + \tfrac{1}{2} \left[\log \frac{S}{S_{i,t}^{\mathrm{G}}} + \frac{S_{i,t}^{\mathrm{G}} + \mu^2}{S} - 1 \right], \qquad \varepsilon_{i,t} := D_{\mathrm{KL}}\big(f_{i,t}^q \| \mathcal{N}(0, S_{i,t}^{\mathrm{G}})\big). \tag{46}$$

The second term equals $D_{\mathrm{KL}}(\mathcal{N}(0, S_{i,t}^{\mathrm{G}}) \| \mathcal{N}(\mu, S))$. It vanishes uniquely at $\mu = 0$ and $S = S_{i,t}^{\mathrm{G}}$. Therefore the unique optimum has precision $\Lambda_{i,t} = (S_{i,t}^{\mathrm{G}})^{-1}$, as stated in Eq. (12). □

For the factorized residual law $F_t^q = \prod_i f_{i,t}^q$, define $\mathsf{D}_t = \mathrm{diag}(S_{i,t}^{\mathrm{G}})$. Its optimal multivariate Gaussian is $G_t^q = \mathcal{N}(0, \mathsf{D}_t) = \prod_i \mathcal{N}(0, S_{i,t}^{\mathrm{G}})$, and

$$D_{\mathrm{KL}}(F_t^q \| G_t^q) = \sum_i \varepsilon_{i,t}, \qquad \varepsilon_{i,t} = \tfrac{1}{2} \log(2\pi e S_{i,t}^{\mathrm{G}}) - \mathcal{H}(f_{i,t}^q).$$

This projection preserves the full mean and covariance of the factorized residual model. After radiometric normalization, it supplies the whitening weight $\mathsf{W}_t = \mathsf{D}_t^{-1}/4$ in Eq. (13). Appendix G establishes the resulting moment, discrepancy, and diffusion-attenuation guarantees in the order used by Section 3.5.

E Prior Step: Temporal Tucker Endpoint Modeling

Following Section 3.3, this section develops the temporal core representation, defines the structural and physical penalties, and describes endpoint averaging. The same TTPM representation supports all three restoration tasks; the observation geometry determines the confidence maps and physical constraints.

E.1 Temporal Coordinates and the Cosine Dictionary

At a fixed parameter snapshot, the endpoint matrices follow Eq. (14). Define the constant cosine column by $T_{t0} = 1$ and use Eq. (16) for the remaining columns. For $1 \le R_{\mathrm{temp}} \le T$, finite cosine orthogonality gives

$$\sum_{t=1}^{T} T_{t\ell} T_{tj} = T\, \mathbb{I}\{\ell = j\}, \qquad 0 \le \ell, j < R_{\mathrm{temp}}.$$

This identity follows from the product-to-sum formula and equally spaced finite cosine sums. Hence the retained columns are linearly independent.

To relate this dictionary to the terminal-centered parameterization, write an ordinary truncated expansion as $G_t = C_0^{\mathrm{core}} + \sum_{\ell=1}^{R_{\mathrm{temp}}-1} T_{t\ell} C_\ell^{\mathrm{core}}$. Its terminal coordinates are

$$G_{\mathrm{term}} = C_0^{\mathrm{core}} + \sum_{\ell=1}^{R_{\mathrm{temp}}-1} T_{1,\ell} C_\ell^{\mathrm{core}}, \qquad \Delta G_\ell = C_\ell^{\mathrm{core}}.$$

Conversely, $C_0^{\mathrm{core}} = G_{\mathrm{term}} - \sum_{\ell=1}^{R_{\mathrm{temp}}-1} T_{1,\ell} \Delta G_\ell$. Terminal re-centering is therefore an invertible coordinate change that preserves the temporal subspace. Moreover, $\phi_{\ell,1} = 0$ gives $G_1 = G_{\mathrm{term}}$ at every parameter snapshot.

For $2 \le t \le T-1$, the interior second difference satisfies

$$T_{t+1,\ell} - 2T_{t\ell} + T_{t-1,\ell} = -4 \sin^2\left(\frac{\ell\pi}{2T}\right) T_{t\ell}.$$

The terminal shift leaves this second difference unchanged. Consequently, low-frequency truncation controls the available temporal frequencies, while the curvature penalty below controls their amplitude. The complete core sequence uses $R_{\text{temp}}R_sR_\lambda$ parameters, and setting all $\Delta G_\ell = 0$ recovers the static-core model. During restoration, warm-started optimization adapts the shared spatial, spectral, and core parameters; Appendix I incorporates this parameter movement into the endpoint-fitting analysis.

E.2 Structural Penalties and Confidence Weights

For $T \geq 3$ and $B \geq 3$, let D_2 denote the spectral second-difference operator. For each acquired measurement, define the detached dimensionless confidence

$$R_{i,t} = \frac{\Lambda_{i,t}}{\tau_{b(i)}} = \frac{\tau_{b(i)}^{-1}}{S_{i,t}^{\mathrm{G}}} = \left(1 + \tau_{b(i)} r_{is} v_{b(i)} \mathbb{E}_q[\omega_i^{-1}]\right)^{-1}.$$

Since $S_{i,t}^{\mathrm{G}} \geq \tau_{b(i)}^{-1} > 0$, this confidence satisfies $0 < R_{i,t} \leq 1$. Let $\overline{R}_{p,b,t}$ be its image-grid representation, with spatial index p. Define

$$w_{p,b,t} = 1 - \tfrac{1}{3}\left(\overline{R}_{p,b,t} + \overline{R}_{p,b+1,t} + \overline{R}_{p,b+2,t}\right), \qquad 1 \leq b \leq B-2.$$

Then the structural penalties in Eq. (17) are

$$\mathcal{L}_{\text{spec}} = \frac{\|D_2 C\|_F^2}{(B-2)R_\lambda} + \chi_{\text{b}} \frac{\|D_2 \boldsymbol{b}\|^2}{B-2}, \tag{47}$$

$$\mathcal{L}_{\text{rel},t} = \frac{\sum_{p,b} w_{p,b,t} [\Delta_\lambda^2 \widehat{X}_t^{\text{mat}}(p,b)]^2}{\max(\sum_{p,b} w_{p,b,t}, 1)}, \tag{48}$$

$$\mathcal{L}_{\text{traj}} = \frac{1}{(T-2)R_sR_\lambda} \sum_{j=2}^{T-1} \|(T-1)^2 (G_{j+1} - 2G_j + G_{j-1})\|_F^2. \tag{49}$$

Here $\Delta_\lambda^2 \widehat{X}_t^{\text{mat}}(p,b) = \widehat{X}_t^{\text{mat}}(p,b+2) - 2\widehat{X}_t^{\text{mat}}(p,b+1) + \widehat{X}_t^{\text{mat}}(p,b)$. The reference normalization uses $\chi_{\text{b}} = 1$; the configured multiplier changes the bias-smoothness summand. All cores in $\mathcal{L}_{\text{traj}}$ are evaluated with the same current parameters. For fewer than three spectral bands, the spectral second-difference penalties are set to zero. The temporal implementation requires at least three fitting stages; its curvature penalty vanishes for temporal rank one.

Let $p_t = (T-t)/(T-1)$ denote reverse progress for $T > 1$. The scheduled weight is $\lambda_{\text{rel},t} = \lambda_{\text{rel},0} p_t^{q_{\text{rel}}}$, with nonnegative coefficient and the configured exponent. Thus the aggregate regularizer used in the analysis is

$$\mathcal{P}_t(\Theta) := \mathcal{J}_t - \mathcal{L}_{\text{diff},t} = \lambda_{\text{spec}}\mathcal{L}_{\text{spec}} + \lambda_{\text{rel},t}\mathcal{L}_{\text{rel},t} + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{phys},t}.$$

For denoising, measurement confidence is reshaped directly to the image grid. Missing-band restoration inserts zero confidence in the missing bands, while pansharpening upsamples LR confidence to the HR grid by nearest-neighbor interpolation. These operations preserve $[0,1]$, so $0 \leq w_{p,b,t} \leq 1$. As a result, the reliability-weighted penalty places stronger spectral regularization on less reliable regions. The dimensionless confidence controls regularization strength, whereas $\Lambda_{i,t}$ controls the observation covariance.

E.3 Curvature Normalization and Efficient Evaluation

For a matrix path $G \in C^4([0,1])$ and $h = 1/(T-1)$, Taylor expansion yields

$$\left\| \frac{G(p+h) - 2G(p) + G(p-h)}{h^2} - G''(p) \right\|_F \leq \frac{h^2}{12} \sup_{s \in [p-h, p+h]} \|G^{(4)}(s)\|_F. \tag{50}$$

The discrete cosine sequence admits the interpolant

$$G^{(T)}(p) = G_{\text{term}} + \sqrt{2} \sum_{\ell=1}^{R_{\text{temp}}-1} \left[\cos\left(\frac{((T-1)p + 1/2)\ell\pi}{T} \right) - \cos\left(\frac{(T-1/2)\ell\pi}{T} \right) \right] \Delta G_\ell,$$

which equals G_t at $p = p_t$. For a fixed smooth path, or a family with uniformly bounded derivatives, finite-difference consistency and a Riemann-sum approximation give

$$\mathcal{L}_{\text{traj}} = \frac{1}{R_s R_\lambda} \int_0^1 \|G''(p)\|_F^2 \, dp + \mathcal{O}(h).$$

Thus the $(T-1)^2$ normalization expresses curvature on the common progress interval $[0, 1]$.

At a fixed parameter snapshot, the shared spatial factor A and spectral factor C further give

$$\|\Delta_t^2 \widehat{X}_t^{\text{mat}}\|_F \le \|A\|_2 \|C\|_2 \|\Delta_t^2 G_t\|_F.$$

For efficient evaluation, define the fixed curvature dictionary and its Gram matrix by

$$\begin{aligned} B_{j\ell}^{\text{curv}} &= -4(T-1)^2 \sin^2\left(\frac{\ell\pi}{2T}\right) T_{j\ell}, \quad 2 \le j \le T-1, \quad 1 \le \ell < R_{\text{temp}}, \\ Q^{\text{curv}} &= \frac{(B^{\text{curv}})^\top B^{\text{curv}}}{T-2}, \qquad \mathcal{L}_{\text{traj}} = \frac{1}{R_s R_\lambda} \sum_{\ell,j=1}^{R_{\text{temp}}-1} Q_{\ell j}^{\text{curv}} \langle \Delta G_\ell, \Delta G_j \rangle_F. \end{aligned} \tag{51}$$

This expression is algebraically equivalent to Eq. (49) and retains all cross-mode terms. The dictionary and Gram matrix are precomputed in double precision and stored at the floating-point precision of the model parameters. Analytical cosine differences provide stable curvature evaluation on long schedules. For $R_{\text{temp}} = 1$, the empty sum gives zero temporal curvature. Measurements and the remaining structural terms complement this regularizer by constraining constant-core directions, as formalized in Appendix I.

E.4 Task-Specific Physical Constraints

Denoising and missing-band restoration use $\mathcal{L}_{\text{phys},t} = 0$. For pansharpening, the LR HSI and PAN image provide complementary physical constraints. The conditional sampler uses the acquired LR observation as its center, while the auxiliary LR fitting term uses the sparse-corrected target

$$\widetilde{y}_{\text{LR},t} = y_{\text{LR}} - \mu_t^s, \qquad \mu_{i,t}^s = r_{is} m_i.$$

Let W_t^{LR} be detached nonnegative weights derived from the LR measurement precision Λ_t and let $\widehat{x}_t = (\widehat{u}_t + \mathbf{1}_n)/2$. The LR term is

$$\mathcal{L}_{\text{LR},t} = \frac{\sum_i W_{i,t}^{\text{LR}} [H\widehat{x}_t - \widetilde{y}_{\text{LR},t}]_i^2}{\max(\sum_i W_{i,t}^{\text{LR}}, 1)}. \tag{52}$$

When more than sixteen positive precision entries are available, the weighting routine caps them at a configured quantile, divides by the median positive capped value, and clips the weights to $[0, 8]$. Smaller sets use median normalization and clipping. These dimensionless weights balance the LR fitting residuals. The conditional sampler uses $\mathsf{W}_t = \text{diag}(\Lambda_t)/4$ in Eq. (13).

For the PAN image $\boldsymbol{p}$, let $R_{\boldsymbol{\rho}}$ combine spectral bands with nonnegative response weights ρ_b satisfying $\sum_b \rho_b = 1$. The prediction is $\widehat{\boldsymbol{p}}_t = g R_{\boldsymbol{\rho}} \widehat{x}_t + c\mathbf{1}$, with learned gain $g > 0$ and offset c. A softmax parameterization enforces the response constraint. Let $\varphi_\epsilon(r) = \sqrt{r^2 + \epsilon^2}$ with $\epsilon = 10^{-3}$. Using positive detached normalization scales S_I, S_G, define

$$\mathcal{L}_{I,t} = \text{mean}\, \varphi_\epsilon((\widehat{\boldsymbol{p}}_t - \boldsymbol{p})/S_I), \tag{53}$$

$$\begin{aligned} \mathcal{L}_{G,t} = {} & \tfrac{1}{2} \text{mean}\, \varphi_\epsilon((\nabla_x \widehat{\boldsymbol{p}}_t - \nabla_x \boldsymbol{p})/S_G) \\ & + \tfrac{1}{2} \text{mean}\, \varphi_\epsilon((\nabla_y \widehat{\boldsymbol{p}}_t - \nabla_y \boldsymbol{p})/S_G), \end{aligned} \tag{54}$$

$$\mathcal{L}_{\text{phys},t} = \lambda_{\text{LR}} \mathcal{L}_{\text{LR},t} + p_t^{q_{\text{PAN}}} (\lambda_I \mathcal{L}_{I,t} + \lambda_G \mathcal{L}_{G,t} + \lambda_\rho \mathcal{L}_{\text{resp},t}). \tag{55}$$

For $B \ge 3$, the response regularizer is

$$\mathcal{L}_{\text{resp},t} = \frac{\|D_2 \boldsymbol{\rho}\|^2}{B-2} + 10^{-4} \left[\left(\frac{g - g_{\text{init}}}{\max(|g_{\text{init}}|, 10^{-6})} \right)^2 + \left(\frac{c - c_{\text{init}}}{S_I} \right)^2 \right].$$

For fewer than three bands, the response-curvature term is set to zero. Initialization estimates the response and calibration from the LR HSI and the downsampled PAN image. Accordingly, the PAN

intensity and gradient terms transfer spatial information, while response smoothness stabilizes the learned spectral combination.

The intensity scale is the larger of a Gaussian-scaled median absolute deviation and one quarter of the 95th-to-5th percentile range of the PAN values. The gradient scale uses the larger of a Gaussian-scaled median absolute deviation and one quarter of the 90th percentile of the absolute pooled gradients. Machine-precision-dependent positive floors keep both scales finite and positive. These normalization quantities remain fixed during endpoint fitting.

E.5 Endpoint Averaging

TTPM returns the fitted endpoint $\widehat{u}_t$, and the conditional sampler uses its running average $\overline{u}_t$. In radiometric coordinates, a single averaging call is

$$\overline{x} \leftarrow \beta_{\mathrm{EMA}}\overline{x} + (1-\beta_{\mathrm{EMA}})\widehat{x}, \qquad 0 \le \beta_{\mathrm{EMA}} < 1.$$

After clipping, the EMA reconstruction is mapped back by $\overline{u}_t = 2\overline{x}_t - \mathbf{1}_n$. For Pavia University pansharpening only, its matrix form Z is first replaced by $Z(I_B + 0.5p_j^{\mathrm{grid}} D_2^\top D_2)^{-1}$ before this mapping. Averaging carries information across neighboring reverse stages, and Appendix I accounts for multiple EMA calls, clipping, and task-specific spectral proximal processing through the stagewise averaging weight and aggregate output correction.

F Sampling Step: Reliability-Guided Conditional Diffusion

This section develops the sampler in Section 3.4. We first express the channel-wise transition in matrix form, then prove its variance-completion rule and instantiate the observation operator for each task. The projection analysis in Appendix G subsequently uses this same transition to quantify moment preservation and attenuation of distributional discrepancy.

F.1 Gaussian Channels and the Compact Transition

The normalized measurement is $y^u = 2y - H\mathbf{1}_n = Hu_0 + 2z$. For the projected covariance $\mathsf{D}_t = \mathrm{diag}(S_{i,t}^{\mathrm{G}})$, the weight $\mathsf{W}_t = \mathsf{D}_t^{-1}/4$ gives the weighted SVD in Eq. (13). Under the projected model, each positive singular value yields the independent Gaussian channel

$$\check{y}_{j,t}^{\dagger} = (V_t^\top u_0)_j + \delta_{j,t}\epsilon_j, \qquad \delta_{j,t} = s_{j,t}^{-1}.$$

Under the auxiliary law F_t^q, the same transform preserves the channel means and covariances and carries the joint residual dependence into the transformed law.

For $t \ge 2$, define the active directions exactly as in the main text: $\mathcal{A}_t = \{j : s_{j,t} > 0,\ \delta_{j,t} < \sigma_{t-1}\}$. Using the singular vectors indexed by $\mathcal{A}_t$, set

$$\begin{gathered}\mathsf{A}_t^{\mathrm{obs}} = V_{t,\mathcal{A}_t}\,\mathrm{diag}(\delta_{j,t})U_{t,\mathcal{A}_t}^\top \mathsf{W}_t^{1/2},\\ O_t = V_{t,\mathcal{A}_t}V_{t,\mathcal{A}_t}^\top, \qquad M_t = I_n - O_t, \qquad u_t^{\mathrm{data}} = \mathsf{A}_t^{\mathrm{obs}} y^u,\\ \mathsf{R}_t^{\mathrm{obs}} = V_{t,\mathcal{A}_t}\,\mathrm{diag}(\delta_{j,t}^2)V_{t,\mathcal{A}_t}^\top, \qquad \Gamma_t = \sigma_{t-1}^2 I_n - \mathsf{R}_t^{\mathrm{obs}}.\end{gathered} \tag{56}$$

An empty active set gives zero observation matrices. The SVD identities yield $\mathsf{A}_t^{\mathrm{obs}} H = O_t$, $u_t^{\mathrm{data}} \in \mathrm{range}(O_t)$, and $O_t M_t = 0$. Consequently, Eq. (18) is equivalent in conditional distribution to

$$v_{t-1} = M_t\overline{u}_t + u_t^{\mathrm{data}} + \Gamma_t^{1/2}\xi_t, \qquad \xi_t \sim \mathcal{N}(0, I_n). \tag{57}$$

The strict active criterion ensures $0 \prec \Gamma_t \preceq \sigma_{t-1}^2 I_n$ at every positive target scale. In the complementary directions, the innovation covariance is exactly $\sigma_{t-1}^2 I$.

The transition mean also equals $\overline{u}_t + \mathsf{A}_t^{\mathrm{obs}}(y^u - H\overline{u}_t)$, which makes the measurement correction explicit. Moreover, because each active inverse singular value is smaller than σ_{t-1},

$$\|\mathsf{A}_t^{\mathrm{obs}} r\| \le \sigma_{t-1}\|\mathsf{W}_t^{1/2} r\|.$$

Thus reliability selection controls the stagewise amplification of a whitened measurement perturbation.

F.2 Variance Completion and Conditional Sampling Risk

Proof of Proposition 1. For $T = Y + \sqrt{c}\,\xi$, averaging over the Gaussian observation $Y = u + \delta Z$ gives $T \sim \mathcal{N}(u, \delta^2 + c)$. Let $\varphi(a) = a - 1 - \log a$ for $a > 0$. Its KL divergence to $\mathcal{N}(u, \sigma^2)$ is $\varphi((\delta^2 + c)/\sigma^2)/2$. The function decreases up to one and increases thereafter, so the minimum over $c \geq 0$ is attained at $c = (\sigma^2 - \delta^2)_+$. It equals zero when $\delta \leq \sigma$. At a zero-variance boundary, the KL is interpreted by its infinite limit, which leaves the optimizer unchanged. For a fixed endpoint mean m,

$$D_{\mathrm{KL}}(\mathcal{N}(m, c) \| \mathcal{N}(u, \sigma^2)) = \tfrac{1}{2}\left[\frac{c + (m-u)^2}{\sigma^2} - 1 + \log\frac{\sigma^2}{c}\right].$$

Differentiating with respect to $c > 0$ gives the unique optimum $c = \sigma^2$, with minimum $(m - u)^2/(2\sigma^2)$. □

The proposition establishes the optimal added variance within each transition family. The implemented strict rule assigns equality cases $\delta_{j,t} = \sigma_{t-1}$ to the endpoint-centered branch, preserving a positive conditional covariance for the stochastic path analysis.

For a fixed endpoint candidate m and fixed model statistics, replace the normalized observation by $Hu_0 + 2E$, where $E \sim F_t^q$. The resulting auxiliary transition output V_f has

$$\mathbb{E}V_f = u_0 + M_t(m - u_0), \qquad \mathrm{Cov}(V_f) = \mathsf{R}_t^{\mathrm{obs}} + \Gamma_t = \sigma_{t-1}^2 I_n. \tag{58}$$

Its Gaussian-projection counterpart V_g is Gaussian with the same moments. Therefore, for a fixed affine reconstruction $D(v) = Sv + b$,

$$\begin{aligned}\mathbb{E}\|D(V_f) - u_0\|^2 &= \mathbb{E}\|D(V_g) - u_0\|^2 \\ &= \|S[u_0 + M_t(m - u_0)] + b - u_0\|^2 + \sigma_{t-1}^2 \operatorname{tr}(SS^\top).\end{aligned} \tag{59}$$

For the acquired observation, define $e_t = \overline{u}_t - u_0$ and $n_t = \mathsf{A}_t^{\mathrm{obs}}(y^u - Hu_0)$. Conditioning on all quantities available before the fresh innovation gives

$$v_{t-1} = u_0 + M_t e_t + n_t + \Gamma_t^{1/2}\xi_t.$$

The mean-error terms lie in orthogonal subspaces, so

$$\mathbb{E}_{\xi_t}\|v_{t-1} - u_0\|^2 = \|M_t e_t\|^2 + \|n_t\|^2 + \operatorname{tr}\Gamma_t. \tag{60}$$

Furthermore, $\|n_t\| \leq \sigma_{t-1}\|\operatorname{diag}(\Lambda_t)^{1/2} z\|$. This conditional identity retains the realized measurement error and remains valid when the observation statistics are functions of the preceding reverse history.

F.3 Task-Specific Observation Channels

All tasks use Eq. (18). Their forward operators specify the measured singular directions, while the physical losses in Appendix E.4 complement endpoint fitting.

F.3.1 Denoising and Missing-Band Restoration

For denoising, $H = I_n$, and the weighted SVD reduces to a coordinatewise construction. Each positive projected precision gives

$$s_{i,t} = \sqrt{\Lambda_{i,t}}/2, \qquad \delta_{i,t} = 2/\sqrt{\Lambda_{i,t}}, \qquad \check{y}_{i,t}^{\dagger} = 2y_i - 1.$$

For an acquired band set Ω_{obs}, band-major vectorization gives $H = P_{\Omega_{\mathrm{obs}}} \otimes I_N$. Inference and projection operate on the acquired rows. Missing bands lie in the image-space null space and are reconstructed through the endpoint-centered branch, with zero confidence on their padded image-grid entries.

F.3.2 Weighted Channels for Pansharpening

Let D denote the spatial blur and downsampling operator for one band. The LR observation is $y_{\mathrm{LR}} = Hx_0 + z_{\mathrm{LR}}$, with $H = I_B \otimes D$ in band-major coordinates. Band-dependent operators use the corresponding D_b in each block. The normalized measurement is

$$y_b^u = 2y_{\mathrm{LR},b} - D\mathbf{1}_N,$$

which incorporates the actual boundary response of the spatial operator.

For each band and fixed projected precision, define $Q_{b,t} = \mathrm{diag}(\sqrt{\Lambda_{b,t}})/2$ and $A_{b,t} = Q_{b,t}D$. Retaining the positive singular values, compute

$$A_{b,t}A_{b,t}^\top = Q_{b,t}DD^\top Q_{b,t} = U_{b,t}\Sigma_{b,t}^2 U_{b,t}^\top, \qquad C_{b,t} = Q_{b,t}U_{b,t}\Sigma_{b,t}^{-1}. \tag{61}$$

The corresponding right singular vectors have the implicit actions

$$\begin{aligned} V_{b,t}^\top x = C_{b,t}^\top Dx, \qquad & V_{b,t}z = D^\top C_{b,t}z, \\ \check{y}_{b,t}^\dagger = C_{b,t}^\top y_b^u, \qquad & \delta_{j,b,t} = s_{j,b,t}^{-1}. \end{aligned} \tag{62}$$

Indeed, $V_{b,t} = D^\top Q_{b,t}U_{b,t}\Sigma_{b,t}^{-1}$ and $V_{b,t}^\top V_{b,t} = I$. Under the projected Gaussian covariance, the channel noise covariance is exactly $\Sigma_{b,t}^{-2}$. Hence the weighted eigendecomposition retains the full operator coupling, including spatially varying precision, while the implicit representation applies the HR basis through D and its adjoint.

Zero-weighted rows are omitted before decomposition, and their associated image directions are handled through the null space of the weighted operator. The numerical rank convention treats Gram eigenvalues below 10^{-12} times the largest eigenvalue as zero. Together with the declared floating-point tolerances, this specifies the implemented weighted Gaussian channels.

The spatial Gram matrix $DD^\top$ is constructed once from the forward operator and its exact adjoint. The weighted eigensystem is refreshed with the projected precision and reused between updates. For m_b LR pixels, a dense per-band eigendecomposition has cubic arithmetic cost and quadratic LR storage. Spectral chunking and host-memory caching control storage of these factors while preserving the defining weighted operator.

F.4 Terminal Output

The stochastic transition in Eq. (18) is applied between consecutive positive-noise fitting stages. At the final fitting stage, we retain the term "sampling" for consistency with the unified likelihood–prior–sampling workflow, but this terminal sampling is deterministic: no additional Gaussian innovation is introduced, and the averaged endpoint is emitted directly.

For the full-grid formulation, after fitting at $t = 1$, the terminal sampling step gives

$$u_{\mathrm{out}} = \overline{u}_1,$$

and the restored image is

$$\widehat{x}_{\mathrm{out}} = \mathrm{clip}\left(\frac{u_{\mathrm{out}} + \mathbf{1}_n}{2},\, 0,\, 1\right).$$

A selected reverse grid follows the same convention at its final fitting stage τ_1, with $u_{\mathrm{out}} = \overline{u}_{\tau_1}$. Thus, the final sampling block should be understood as deterministic terminal emission.

G Analysis of Gaussian Projection and Diffusion Attenuation

This section follows the approximation analysis in Section 3.5: it first proves second-order invariance, then bounds the projection discrepancy, and finally establishes its attenuation by diffusion. Each comparison conditions on a fixed scene reference and fixed inferred statistics. This stagewise analysis is complemented by the history-dependent recovery results in Appendix I.

G.1 Propagation of the Matched Moments

Proof of Proposition 2. By Eq. (45), the residuals $E \sim F_t^q$ and $E_{\mathrm{G}} \sim G_t^q$ have mean zero and covariance $\mathsf{D}_t = \mathrm{diag}(S_{i,t}^{\mathrm{G}})$. Let ξ be an independent standard Gaussian. For the fixed affine map in the proposition, define $Z_f = a + LE + C\xi$ and $Z_g = a + LE_{\mathrm{G}} + C\xi$. Both outputs have mean a and covariance $L\mathsf{D}_t L^\top + CC^\top$. Therefore, for any fixed target u and symmetric $Q \succeq 0$,

$$\begin{aligned}\mathbb{E}[(Z_f - u)^\top Q(Z_f - u)] &= (a-u)^\top Q(a-u) + \mathrm{tr}\{Q(L\mathsf{D}_t L^\top + CC^\top)\} \\ &= \mathbb{E}[(Z_g - u)^\top Q(Z_g - u)].\end{aligned} \tag{63}$$

Setting $Q = I_n$ proves Eq. (20). Composition with another fixed affine reconstruction map preserves the identity. □

Since the two residual models share their covariance, they produce the same whitening matrix, singular values, active directions, and innovation covariance. Thus the conditional transition preserves their first- and second-order information at the fixed inference snapshot.

The same reasoning extends to repeated affine updates that reuse a measurement residual. For deterministic coefficients, a deterministic initial state a_T, and independent Gaussian innovations, consider

$$a_{t-1} = L_t a_t + B_t E + c_t + C_t \xi_t.$$

Repeated substitution expresses every subsequent state as an affine function of the shared E and the innovations. Replacing E by its moment-matched Gaussian therefore preserves means, covariances, and quadratic risks, including the correlations induced by measurement reuse.

For a nonlinear reconstruction map $D(z) = Sz + b + r(z)$, the affine contribution remains explicit. Define $R_{\mathrm{aff}} = \mathbb{E}\|SZ_f + b - u_0\|^2$, $\zeta_f = (\mathbb{E}\|r(Z_f)\|^2)^{1/2}$, and $\zeta_g = (\mathbb{E}\|r(Z_g)\|^2)^{1/2}$. Expanding the squared errors and applying Cauchy–Schwarz gives

$$\left|\mathbb{E}\|D(Z_f) - u_0\|^2 - \mathbb{E}\|D(Z_g) - u_0\|^2\right| \le 2\sqrt{R_{\mathrm{aff}}}(\zeta_f + \zeta_g) + \zeta_f^2 + \zeta_g^2. \tag{64}$$

Hence controlled nonlinear remainders translate the affine invariance into a quantitative reconstruction-risk comparison.

G.2 Bound on the Projection Discrepancy

Proof of Proposition 3. Let the random conditional variance S_i equal $s_{n,i}$ in the ordinary state and $s_i(\omega)$ in the sparse state. Since $\mathbb{E}\mathsf{S}_i = S_{i,t}^{\mathrm{G}}$, convexity of KL under mixing gives

$$\begin{aligned}0 \le \varepsilon_{i,t} &\le \mathbb{E} D_{\mathrm{KL}}\big(\mathcal{N}(0,\mathsf{S}_i)\|\mathcal{N}(0,S_{i,t}^{\mathrm{G}})\big) \\ &= \tfrac{1}{2}[\log S_{i,t}^{\mathrm{G}} - \mathbb{E}\log \mathsf{S}_i] = \mathcal{B}_{i,t}, \\ \mathcal{B}_{i,t} &= \tfrac{1}{2}\left[\log S_{i,t}^{\mathrm{G}} - (1-r_{is})\log s_{n,i} - r_{is}\mathbb{E}_q \log s_i(\omega)\right].\end{aligned} \tag{65}$$

The Gaussian variance terms cancel after averaging over the mixture. Additivity of KL for product laws then gives $D_{\mathrm{KL}}(F_t^q\|G_t^q) = \sum_i \varepsilon_{i,t} \le \sum_i \mathcal{B}_{i,t}$, proving Eq. (21). □

The remaining one-dimensional log-scale expectation can be evaluated numerically under the inferred Gamma law. To interpret the bound, set $\bar{s}_{s,i} = \mathbb{E}_q s_i(\omega)$ and $\kappa_i = \bar{s}_{s,i}/s_{n,i}$. Then

$$\begin{aligned}2\mathcal{B}_{i,t} = \ &\log(1 - r_{is} + r_{is}\kappa_i) - r_{is}\log\kappa_i \\ &+ r_{is}\{\log \bar{s}_{s,i} - \mathbb{E}_q \log s_i(\omega)\}.\end{aligned} \tag{66}$$

The first line measures between-state variance contrast. It approaches zero when the variances coincide or when a state assignment becomes certain with bounded variance contrast. The second line measures scale dispersion within the sparse state. Together, the terms identify how state uncertainty and conditional scale variability contribute to the projection discrepancy.

For $q(\omega) = \mathrm{Gamma}(a,b)$ with $a > 1$, the within-state gap obeys

$$0 \le \log \bar{s}_s - \mathbb{E}_q \log s(\omega) \le \mathrm{digamma}(a) - \log(a-1). \tag{67}$$

To establish the upper bound, let $X = 1/\omega$ and differentiate $\log(c + v\mathbb{E}X) - \mathbb{E}\log(c + vX)$ with respect to $c \geq 0$. Convexity of the reciprocal makes the derivative nonpositive. At $c = 0$, the gap equals $\log \mathbb{E}X - \mathbb{E}\log X = \mathrm{digamma}(a) - \log(a-1)$. Thus a larger dense-noise variance relative to the sparse contribution reduces the scale gap. For the inferred Gamma law, a_ω specifies the shape component of this bound, while $b_{\omega i}$ adapts the scale. Small state-variance contrast and controlled scale dispersion provide explicit regimes of small projection discrepancy.

The optimality of full-mixture moment matching can also be quantified relative to pooling statewise precisions. Define

$$\Lambda_{i,t}^{\mathrm{pool}} = (1 - r_{is})/s_{n,i} + r_{is}/\bar{s}_{s,i}.$$

Equation (46) yields the excess KL

$$\begin{aligned} D_{\mathrm{KL}}(f_{i,t}^q \| \mathcal{N}(0, 1/\Lambda_{i,t}^{\mathrm{pool}})) - \varepsilon_{i,t} &= \tfrac{1}{2}(C_{i,t}^{\mathrm{pool}} - 1 - \log C_{i,t}^{\mathrm{pool}}), \\ C_{i,t}^{\mathrm{pool}} = \Lambda_{i,t}^{\mathrm{pool}} S_{i,t}^{\mathrm{G}} &= 1 + r_{is}(1 - r_{is})\frac{(\kappa_i - 1)^2}{\kappa_i} \geq 1. \end{aligned} \tag{68}$$

Accordingly, the full-mixture projection minimizes this Gaussian approximation error while retaining the aggregate variance.

G.3 Diffusion Attenuation of Residual Shape

Proof of Proposition 4. For $0 < \delta^2 < \sigma^2$ and finite $\varepsilon_{\mathrm{in}} = D_{\mathrm{KL}}(\mathcal{L}(E) \| \mathcal{N}(0, \delta^2))$, the scalar entropy power is

$$N(E) = (2\pi e)^{-1} \exp(2\mathcal{H}(E)) = \delta^2 \exp(-2\varepsilon_{\mathrm{in}}).$$

For an independent standard Gaussian ξ, the entropy-power inequality[1] gives

$$N(E + \sqrt{\sigma^2 - \delta^2}\,\xi) \geq \delta^2 e^{-2\varepsilon_{\mathrm{in}}} + \sigma^2 - \delta^2.$$

The output variance is σ^2. Writing $a = \delta^2/\sigma^2 \in (0, 1)$, its Gaussian entropy deficit satisfies

$$\begin{aligned} \varepsilon_{\mathrm{out}} &\leq -\tfrac{1}{2}\log(1 - a + ae^{-2\varepsilon_{\mathrm{in}}}) \\ &\leq a\varepsilon_{\mathrm{in}} = \frac{\delta^2}{\sigma^2}\varepsilon_{\mathrm{in}}. \end{aligned} \tag{69}$$

The second inequality follows from $1 - a + ae^{-2\varepsilon_{\mathrm{in}}} \geq e^{-2a\varepsilon_{\mathrm{in}}}$. This proves Eq. (22). The zero-variance case reduces to an exactly Gaussian output. □

Consequently, each active scalar channel contracts its residual shape discrepancy by a factor bounded by $\delta_{j,t}^2/\sigma_{t-1}^2$. The reduction strengthens as the observation variance becomes smaller relative to the target diffusion variance. Endpoint-centered directions propagate the estimated endpoint with Gaussian innovation, while the history-dependent analysis in Appendix I tracks their inherited error across refreshed inference snapshots.

A joint comparison additionally retains dependence among transformed residual coordinates. Let $B_{m,t}^q$ and $B_{m,t}^{\mathrm{G}}$ be the auxiliary transition laws formed from F_t^q and G_t^q, respectively, with the same fixed endpoint m and the moment-matched channel construction. Data processing through the shared affine-plus-Gaussian kernel gives

$$\varepsilon_t^{\mathrm{bridge}} := D_{\mathrm{KL}}(B_{m,t}^q \| B_{m,t}^{\mathrm{G}}) \leq \sum_i \varepsilon_{i,t} \leq \sum_i \mathcal{B}_{i,t}. \tag{70}$$

Their shared mean and covariance from Eq. (58) also imply

$$D_{\mathrm{KL}}(B_{m,t}^q \| \mathcal{N}(u_0, \sigma_{t-1}^2 I_n)) = \varepsilon_t^{\mathrm{bridge}} + \frac{\|M_t(m - u_0)\|^2}{2\sigma_{t-1}^2}. \tag{71}$$

This identity follows by subtracting the Gaussian log densities and using the matched moments. It separates the residual-shape contribution from the endpoint-mean contribution.

[1] S. Verdú and D. Guo, "A Simple Proof of the Entropy-Power Inequality," *IEEE Transactions on Information Theory*, 52(5), pp. 2165–2166, 2006. DOI https://doi.org/10.1109/TIT.2006.872978.

For a common fixed reconstruction map D with output in $[-1,1]^n$, define $\ell(v) = n^{-1}\|D(v)-u_0\|^2 \in [0,4]$. Pinsker's inequality gives

$$\left|\mathbb{E}_{B^q_{m,t}}\ell - \mathbb{E}_{B^{\mathrm{G}}_{m,t}}\ell\right| \le 4\min\left\{1, \sqrt{\varepsilon_t^{\mathrm{bridge}}/2}\right\}. \tag{72}$$

Thus the joint discrepancy controls normalized reconstruction-risk differences for both affine and nonlinear maps in this fixed-snapshot comparison. In radiometric coordinates $[0,1]^n$, the factor four becomes one.

G.4 Latent Inference and Fixed-Observation Likelihood Comparisons

Let $\varepsilon_t^{\mathrm{VI}}$ denote the latent KL in Eq. (34). Marginalize the exact latent posterior and its variational approximation to their state and scale variables, then apply the same residual-generation kernel. Denote the resulting laws by F_t^{post} and F_t^q. Data processing, the triangle inequality for total variation, and Pinsker's inequality yield

$$\begin{aligned} D_{\mathrm{KL}}(F_t^q \| F_t^{\mathrm{post}}) &\le \varepsilon_t^{\mathrm{VI}}, \\ \mathrm{TV}(F_t^{\mathrm{post}}, G_t^q) &\le \sqrt{\varepsilon_t^{\mathrm{VI}}/2} + \sqrt{\tfrac{1}{2}\sum_i \varepsilon_{i,t}}. \end{aligned} \tag{73}$$

All laws use the same reference and parameter values. Hence this bound separates the contribution of latent approximation from that of Gaussian projection.

The original observation model has residual density

$$f_i^{\mathrm{obs}}(e) = (1-\pi_s)\mathcal{N}(e \mid 0, s_{n,i}) + \pi_s \mathbb{E}_{\omega\sim\mathrm{Gamma}(\nu/2,\nu/2)}\mathcal{N}(e \mid 0, s_i(\omega)).$$

With $g_{i,t} = \mathcal{N}(0, S^{\mathrm{G}}_{i,t})$ and $\nu > 2$, mixture convexity gives

$$D_{\mathrm{KL}}(f_i^{\mathrm{obs}} \| g_{i,t}) \le \frac{1-\pi_s}{2}\varphi(\Lambda_{i,t}s_{n,i}) + \frac{\pi_s}{2}\mathbb{E}_{p(\omega)}\varphi(\Lambda_{i,t}s_i(\omega)), \qquad \varphi(a) = a - 1 - \log a.$$

The condition $\nu > 2$ ensures the prior residual variance used in this comparison is finite.

For a fixed observed image, a likelihood-ratio certificate provides a complementary characterization. On a compact feasible set $\mathcal{U}$, use a common proper scene prior and fix either residual likelihood f above together with the projected likelihood g. Define

$$L_y(u) = \sum_i \log \frac{f_i(y_i - h_i^\top(u + \mathbf{1}_n)/2)}{g_i(y_i - h_i^\top(u + \mathbf{1}_n)/2)}, \qquad \epsilon_y = \sup_{u\in\mathcal{U}} L_y(u) - \inf_{u\in\mathcal{U}} L_y(u).$$

When ϵ_y is finite, the corresponding normalized scene distributions Π_f and Π_g satisfy

$$e^{-\epsilon_y} \le \frac{d\Pi_f}{d\Pi_g} \le e^{\epsilon_y}, \qquad \max\{D_{\mathrm{KL}}(\Pi_f\|\Pi_g), D_{\mathrm{KL}}(\Pi_g\|\Pi_f)\} \le \epsilon_y. \tag{74}$$

Indeed, $d\Pi_f/d\Pi_g = e^{L_y}/\mathbb{E}_{\Pi_g}e^{L_y}$, and the denominator lies between the minimum and maximum of e^{L_y}. This certificate quantifies the effect of the frozen likelihood choice at the acquired observation. The implemented sampler is characterized directly by the conditional identities in Appendix F and the adaptive-path identity in Appendix I.5.

H Analysis of the Sampled Endpoint-Fitting Objective

We next follow the restoration-mechanism analysis in Section 3.5. The conditional transition supplies a fitting target with two complementary components: measurement consistency in active directions and endpoint continuity in the remaining directions. Structural and physical penalties then regularize the fitted scene.

H.1 Expected Objective Induced by the Conditional Transition

Multiplying Eq. (17) at timestep $t-1$ by $\lambda_{t-1}^{\mathrm{fit}} = n\sigma_{t-1}^2/2$ gives the equivalent minimization objective

$$\mathcal{F}_{t-1}(\Theta, v) = \tfrac{1}{2}\|\widehat{u}_{t-1}(\Theta) - v\|^2 + \lambda_{t-1}^{\mathrm{fit}}\mathcal{P}_{t-1}(\Theta). \tag{75}$$

This positive rescaling exposes the data-fitting geometry. Numerical optimization uses the original scale of $\mathcal{J}_{t-1}$.

Proof of Proposition 5. Fix the candidate $F = \widehat{u}_{t-1}(\Theta)$ and the penalty coefficients before drawing ξ_t. In the basis V_t, the active coordinates of v_{t-1} have mean $\breve{y}_{j,t}^{\dagger}$ and variance $\sigma_{t-1}^2 - \delta_{j,t}^2$. The complementary coordinates have mean $\overline{u}_{j,t}^{V}$ and variance σ_{t-1}^2. Applying $\mathbb{E}(a-Z)^2 = (a-\mathbb{E}Z)^2 + \mathrm{Var}(Z)$ coordinatewise gives Eq. (23). Equivalently, Eq. (57) yields

$$\begin{aligned}\mathbb{E}_{\xi_t}\mathcal{F}_{t-1}(\Theta, v_{t-1}) = \tfrac{1}{2}\|O_tF - u_t^{\mathrm{data}}\|^2 + \tfrac{1}{2}\|M_t(F-\overline{u}_t)\|^2 \\ + \lambda_{t-1}^{\mathrm{fit}}\mathcal{P}_{t-1}(\Theta) + \tfrac{1}{2}\operatorname{tr}\Gamma_t.\end{aligned} \tag{76}$$

The orthogonal decomposition $O_t + M_t = I_n$ separates measurement consistency from endpoint continuity, and the trace term collects the sampling variance. □

The expected objective therefore identifies how each source of information enters fitting. For a realized sampled loss, the innovation remains part of the fitting target. Appendix I carries that innovation through the local response and includes subsequent confidence refreshes and parameter updates in the finite-fitting discrepancy.

H.2 Stochastic Proximal Interpretation on an Affine Model

Consider a fixed affine representation $F(a) = c + Za$ with $Z^\top Z = I_d$, and suppose $\overline{u}_t = c + Za_t$. Let the penalty restricted to these coordinates, $\mathcal{P}_{t-1}(a)$, be proper, closed, and convex. Define

$$\ell_t(a) = \tfrac{1}{2}\|O_t(c+Za) - u_t^{\mathrm{data}}\|^2, \qquad \zeta_t = Z^\top\Gamma_t^{1/2}\xi_t.$$

Completing the square shows that the exact minimizer of the sampled objective is

$$a_{t-1}^{\mathrm{prox}} = \mathrm{prox}_{\lambda_{t-1}^{\mathrm{fit}}\mathcal{P}_{t-1}}(a_t - \nabla\ell_t(a_t) + \zeta_t). \tag{77}$$

Indeed, $Z^\top(v_{t-1} - c) = a_t - \nabla\ell_t(a_t) + \zeta_t$. Moreover, $\nabla^2\ell_t = Z^\top O_t Z \preceq I_d$, so the zero-innovation specialization is a unit-step proximal-gradient update. This affine formulation exposes the stochastic correction underlying sampled fitting. The next section extends the analysis to the endpoint returned by finite nonlinear optimization through an explicit local-fit certificate.

I Recovery of the Computed Endpoints

This section derives the endpoint-error recurrence in Eq. (24) and its finite-horizon consequence. The argument proceeds from a local regularized response to a certificate for the computed nonlinear fit, then combines measurement–regularizer complementarity with endpoint averaging. Finally, an adaptive-path identity connects endpoint accuracy to conditional transition accuracy.

I.1 Local Regularized Response and Noise Filtering

At a regular fixed-rank snapshot, the tangent space of a centered rank-r matrix in $\mathbb{R}^{N\times B}$ has dimension $r(N+B-r)$. A free spectral bias adds at most B directions. Thus a local orthonormal image basis $Z_t \in \mathbb{R}^{n\times d_t}$ can be chosen with

$$d_t \le \min\{n, r(N+B-r)+B\}, \qquad r \le \min(R_s, R_\lambda).$$

A regular fixed-rank chart provides this dimension reduction. At a rank-changing point, the same response analysis applies with a larger comparison subspace, including the full image space.

Before drawing ξ_t, fix an affine comparison chart $c_t + Z_t a$, with $Z_t^\top Z_t = I_{d_t}$, and a quadratic comparison penalty $a^\top \Omega_t a/2 + \boldsymbol{\ell}_t^\top a$, where $\Omega_t \succeq 0$. For $\lambda = \lambda_{t-1}^{\mathrm{fit}} > 0$, define

$$\begin{aligned} \mathsf{K}_t &= I_{d_t} + \lambda \Omega_t, \qquad \mathsf{S}_t = Z_t \mathsf{K}_t^{-1} Z_t^\top, \\ D_t^{\mathrm{lin}}(v) &= c_t + Z_t \mathsf{K}_t^{-1}[Z_t^\top (v - c_t) - \lambda \boldsymbol{\ell}_t]. \end{aligned} \tag{78}$$

This is the unique fitted output for the chosen local quadratic model. Its effective noise dimension is $d_{\mathrm{eff},t} = \mathrm{tr}(\mathsf{S}_t^2) = \mathrm{tr}(\mathsf{K}_t^{-2}) \le d_t$.

For fixed h and independent centered noise with covariance Σ, direct expansion gives

$$\begin{aligned} \mathbb{E}\|D_t^{\mathrm{lin}}(u_0 + h + \epsilon) - u_0\|^2 &= \|\mathsf{S}_t h + b_t^{\mathrm{mod}}\|^2 + \mathrm{tr}(\mathsf{S}_t \Sigma \mathsf{S}_t), \\ b_t^{\mathrm{mod}} &= D_t^{\mathrm{lin}}(u_0) - u_0. \end{aligned} \tag{79}$$

When $\Sigma \preceq \sigma^2 I_n$, the fresh-noise contribution is bounded by $\sigma^2 d_{\mathrm{eff},t}$. In particular, a lower-dimensional regularized response filters the innovation energy. If $u_0 = c_t + Z_t a_t^{\mathrm{clean}} + r_t^0$ with $Z_t^\top r_t^0 = 0$, then

$$b_t^{\mathrm{mod}} = -r_t^0 - \lambda Z_t \mathsf{K}_t^{-1}(\Omega_t a_t^{\mathrm{clean}} + \boldsymbol{\ell}_t).$$

The two terms explicitly identify representation error and regularization bias.

I.2 A Certificate for Finite Nonlinear Fitting

For the endpoint $\widehat{u}_{t-1}$ returned by the actual optimizer, define

$$\begin{aligned} \widehat{a} &= Z_t^\top(\widehat{u}_{t-1} - c_t), \qquad r^\perp = (I_n - Z_t Z_t^\top)(\widehat{u}_{t-1} - c_t), \\ s^{\mathrm{fit}} &= \mathsf{K}_t \widehat{a} - Z_t^\top (v_{t-1} - c_t) + \lambda \boldsymbol{\ell}_t. \end{aligned}$$

Subtracting the comparison normal equation yields

$$\begin{aligned} \widehat{u}_{t-1} &= D_t^{\mathrm{lin}}(v_{t-1}) + \varepsilon_{t-1}^{\mathrm{fit}}, \qquad \varepsilon_{t-1}^{\mathrm{fit}} = r^\perp + Z_t \mathsf{K}_t^{-1} s^{\mathrm{fit}}, \\ \|\varepsilon_{t-1}^{\mathrm{fit}}\|^2 &= \|r^\perp\|^2 + \|\mathsf{K}_t^{-1} s^{\mathrm{fit}}\|^2. \end{aligned} \tag{80}$$

This identity characterizes every returned endpoint relative to the selected local response. The orthogonal component records movement beyond the comparison subspace, while the normal-equation residual records the discrepancy within it. Both quantities are evaluable from the output and the chosen comparison model.

Sufficient conditions for a small certificate can be stated through a smooth normal-graph chart $F(a) = c_t + Z_t a + r(a)$ with $Z_t^\top r(a) = 0$, $r(0) = Dr(0) = 0$, and $\|D^2 r\| \le L_F$. Then $\|r(a)\| \le L_F \|a\|^2/2$. Suppose the restricted penalty satisfies $\|\nabla \mathcal{P}(a) - (\Omega_t a + \boldsymbol{\ell}_t)\| \le L_R \|a\|^2/2$. If g is the attained gradient of the scaled fitting objective in these coordinates, differentiation gives

$$\|s^{\mathrm{fit}}\| \le \|g\| + L_F \|a\| \, \|F(a) - v_{t-1}\| + \tfrac{\lambda L_R}{2}\|a\|^2.$$

Thus controlled chart displacement, bounded curvature, and a small attained gradient provide a quantitative route to a small finite-fitting discrepancy.

I.3 Complementarity of Measurements and Regularization

Let $\mathsf{G}_t = Z_t^\top O_t Z_t$ and $q_t^{\mathrm{c}} = \|\mathsf{S}_t M_t\|_2$. Since M_t is an orthogonal projector,

$$(q_t^{\mathrm{c}})^2 = \lambda_{\max}\{\mathsf{K}_t^{-1}(I_{d_t} - \mathsf{G}_t)\mathsf{K}_t^{-1}\} \le 1. \tag{81}$$

For $\lambda > 0$, equality occurs precisely when a nonzero vector is preserved by both contractions K_t^{-1} and $(I_{d_t} - \mathsf{G}_t)^{1/2}$. Such a vector lies in $\ker(\Omega_t) \cap \ker(\mathsf{G}_t)$, and every vector in that common kernel attains norm one. Finite-dimensional compactness therefore gives

$$q_t^{\mathrm{c}} < 1 \quad \Longleftrightarrow \quad \ker(Z_t^\top O_t Z_t) \cap \ker(\Omega_t) = \{0\}. \tag{82}$$

For $\lambda = 0$, the corresponding condition is $\mathsf{G}_t \succ 0$. The complementarity condition states that each local error direction is constrained by active measurements, regularization, or both. In particular, measurements can constrain constant temporal modes that the curvature penalty preserves.

At a fixed snapshot, consider the regularized observed-data objective

$$\Phi_t(a) = \tfrac{1}{2}\|O_t(c_t + Z_t a) - u_t^{\mathrm{data}}\|^2 + \lambda(\tfrac{1}{2}a^\top \Omega_t a + \boldsymbol{\ell}_t^\top a).$$

Under complementarity, $\mathsf{G}_t + \lambda\Omega_t \succ 0$, and its unique minimizer is

$$a^\dagger = (\mathsf{G}_t + \lambda\Omega_t)^{-1}(b_t^{\mathrm{data}} - \lambda\boldsymbol{\ell}_t), \qquad b_t^{\mathrm{data}} = Z_t^\top(u_t^{\mathrm{data}} - O_t c_t). \tag{83}$$

For repeated exact affine fits with this fixed snapshot, fixed covariance Γ, and fixed averaging weight $\gamma \in (0,1]$, let $a^{(j)}$ denote the averaged coefficient after j fits. Then

$$\begin{aligned} a^{(j+1)} - a^\dagger = [(1-\gamma)I_{d_t} + \gamma\mathsf{K}_t^{-1}(I_{d_t} - \mathsf{G}_t)](a^{(j)} - a^\dagger) \\ + \gamma\mathsf{K}_t^{-1}Z_t^\top\Gamma^{1/2}\xi^{(j)}. \end{aligned}$$

The linear response has norm at most $\rho = 1 - \gamma + \gamma q_t^{\mathrm{c}} < 1$. Independence and zero mean of the fresh innovations therefore yield

$$\begin{aligned} \mathbb{E}\|a^{(L)} - a^\dagger\|^2 \le \rho^{2L}\mathbb{E}\|a^{(0)} - a^\dagger\|^2 \\ + \gamma^2 \operatorname{tr}(\mathsf{K}_t^{-1}Z_t^\top\Gamma Z_t\mathsf{K}_t^{-1})\frac{1-\rho^{2L}}{1-\rho^2}. \end{aligned}$$

If $u_0 = c_t + Z_t a_t^{\mathrm{clean}}$ and $u_t^{\mathrm{data}} = O_t u_0 + n_t$, the regularized target has bias

$$a^\dagger - a_t^{\mathrm{clean}} = (\mathsf{G}_t + \lambda\Omega_t)^{-1}[Z_t^\top n_t - \lambda(\Omega_t a_t^{\mathrm{clean}} + \boldsymbol{\ell}_t)]. \tag{84}$$

Together, these equations quantify contraction toward the regularized data target and the measurement and regularization terms governing its distance to the clean image.

I.4 Adaptive Endpoints and the Finite-Horizon Bound

Let h_t collect the reverse history, optimizer state, inferred statistics, and endpoint-averaging state available before the fresh draw ξ_t. Choose the comparison quantities above from this history. Let $J_{t-1} \ge 1$ be the scheduled number of EMA calls at stage $t-1$, known from h_t, and set $\gamma_t = 1 - \beta_{\mathrm{EMA}}^{J_{t-1}} \in (0,1]$. Write $\widehat{u}_{t-1}$ for the final fitted endpoint and define the aggregate output correction by

$$c_{t-1}^{\mathrm{out}} := \overline{u}_{t-1} - (1-\gamma_t)\overline{u}_t - \gamma_t\widehat{u}_{t-1}.$$

This correction includes the contribution of intermediate fitted endpoints in multiple EMA calls, all clipping operations, and any task-specific spectral proximal processing in Appendix E. It also accounts for any difference between the stored EMA state and the emitted endpoint. Define

$$\varepsilon_{t-1} := \varepsilon_{t-1}^{\mathrm{fit}} + c_{t-1}^{\mathrm{out}}/\gamma_t.$$

Combining this definition with Eqs. (57) and (80) gives the exact error identity

$$\begin{aligned} e_{t-1} = [(1-\gamma_t)I_n + \gamma_t\mathsf{S}_t M_t]e_t + \gamma_t(\mathsf{S}_t n_t + b_t^{\mathrm{mod}}) \\ + \gamma_t\mathsf{S}_t\Gamma_t^{1/2}\xi_t + \gamma_t\varepsilon_{t-1}, \qquad e_t = \overline{u}_t - u_0. \end{aligned} \tag{85}$$

The discrepancy ε_{t-1} incorporates the realized nonlinear fit, refreshed confidence weights, and the aggregate output correction above; it need not be independent of ξ_t. Define $\zeta_t = (\mathbb{E}[\|\varepsilon_{t-1}\|^2 \mid h_t])^{1/2}$. Conditional Minkowski and the covariance bound give

$$\begin{aligned} (\mathbb{E}[\|e_{t-1}\|^2 \mid h_t])^{1/2} \le (1 - \gamma_t + \gamma_t q_t^{\mathrm{c}})\|e_t\| \\ + \gamma_t\big(\|\mathsf{S}_t n_t\| + \|b_t^{\mathrm{mod}}\| + \sigma_{t-1}\sqrt{d_{\mathrm{eff},t}} + \zeta_t\big). \end{aligned} \tag{86}$$

This conditional bound retains the joint dependence of fitting and sampling through ζ_t.

Proposition 6 (Finite-horizon recovery of the computed endpoints)**.** *Assume that the comparison quantities are measurable from h_t and that the displayed second moments are finite. Let deterministic $\rho_t \in [0,1]$ satisfy $1 - \gamma_t + \gamma_t q_t^{\mathrm{c}} \le \rho_t$ almost surely, and choose deterministic $\eta_t \ge 0$ such that*

$$\eta_t \ge \Big\|\gamma_t\big(\|\mathsf{S}_t n_t\| + \|b_t^{\mathrm{mod}}\| + \sigma_{t-1}\sqrt{d_{\mathrm{eff},t}} + \zeta_t\big)\Big\|_{L^2(P)}. \tag{87}$$

Then the root-mean-square endpoint error from the main text, $\mathcal{E}_t = (\mathbb{E}\|\overline{u}_t - u_0\|^2)^{1/2}$, satisfies

$$\mathcal{E}_{t-1} \le \rho_t \mathcal{E}_t + \eta_t.$$

For a full-grid trajectory from T to r, where $1 \le r < T$,

$$\mathcal{E}_r \le \left(\prod_{j=r+1}^{T} \rho_j\right) \mathcal{E}_T + \sum_{i=r+1}^{T} \left(\prod_{j=r+1}^{i-1} \rho_j\right) \eta_i, \tag{88}$$

with empty products equal to one.

Proof. Taking total L^2 norms in Eq. (86), applying the deterministic coefficient bound, and using Minkowski yields the one-step recurrence in Eq. (24). Iterating it in reverse-time order gives Eq. (88). □

A uniform complementarity margin supplies a deterministic strict contraction factor. Specifically, $q_t^{\mathrm{c}} \le \overline{q}_t < 1$ and $\gamma_t \ge \underline{\gamma}_t > 0$ imply $\rho_t = 1 - \underline{\gamma}_t(1 - \overline{q}_t) < 1$. Under $\rho_t \le \rho < 1$ and $\eta_t \le \eta$, the finite-horizon bound becomes

$$\mathcal{E}_{T-L} \le \rho^L \mathcal{E}_T + \frac{1-\rho^L}{1-\rho}\eta, \qquad 1 \le L \le T-1. \tag{89}$$

Thus repeated sampling and fitting suppress inherited endpoint error geometrically, while the explicit perturbation budget controls the accumulated measurement, model, innovation, and finite-fitting contributions. Across an asymptotic sequence of fitting cycles with the same uniform contraction and perturbations tending to zero, the recurrence yields mean-square endpoint convergence.

The recurrence also covers a nonexpansive tail. If its stages satisfy $\rho_t \le 1$, then for $r < r_0$,

$$\mathcal{E}_r \le \mathcal{E}_{r_0} + \sum_{t=r+1}^{r_0} \eta_t.$$

This form controls the tail by its entering error and remaining perturbation budget, including stages governed primarily by the endpoint prior. Returning to radiometric coordinates scales squared endpoint error by $1/4$. Projection onto $[0,1]^n$ further preserves or reduces Euclidean error to x_0.

I.5 Endpoint Error Along the Adaptive Sampling Path

Fix the acquired observation y and use u_0 as the clean reference. At a common history h_t, let κ_t be the implemented Gaussian kernel in Eq. (57), and let $\kappa_t^\star$ replace its endpoint $\overline{u}_t$ by u_0. The paired kernels share the observation basis, active set, and innovation covariance. This comparison isolates endpoint substitution while retaining the observation-inference and channel-construction rules.

Let χ collect initialization and auxiliary optimizer randomness, and let μ be the common initial law of (χ, v_T). For a positive-noise prefix ending at $r \ge 1$, define

$$\begin{aligned} P_{r:T}(d\chi, dv_{r:T}) &= \mu(d\chi, dv_T) \prod_{t=r+1}^{T} \kappa_t(dv_{t-1} \mid h_t), \\ P^\star_{r:T}(d\chi, dv_{r:T}) &= \mu(d\chi, dv_T) \prod_{t=r+1}^{T} \kappa^\star_t(dv_{t-1} \mid h_t). \end{aligned} \tag{90}$$

The products are evaluated in reverse-time order, and each process applies the same update rules to its own history.

Proposition 7 (Endpoint error in the adaptive path)**.** *For measurable finite endpoints on a positive-noise prefix,*

$$\begin{aligned} D_{\mathrm{KL}}(\kappa_t \| \kappa_t^\star) &= \frac{\|M_t(\overline{u}_t - u_0)\|^2}{2\sigma_{t-1}^2}, \\ \mathcal{D}_{r:T} := D_{\mathrm{KL}}(P_{r:T} \| P^\star_{r:T}) &= \sum_{t=r+1}^{T} \mathbb{E}_P \frac{\|M_t(\overline{u}_t - u_0)\|^2}{2\sigma_{t-1}^2}. \end{aligned} \tag{91}$$

The identities hold in the extended nonnegative sense, and finite weighted moments give finite path KL.

Proof. The paired kernels coincide on the observation-centered subspace. On its orthogonal complement, their means differ by $M_t e_t$ and their shared covariance is $\sigma_{t-1}^2 I$. The equal-covariance Gaussian KL formula gives the first identity. Taking the log likelihood ratio of Eq. (90) and its expectation under P gives the second by the conditional KL chain rule. □

In particular, the endpoint-error bounds above give

$$\mathcal{D}_{r:T} \le \sum_{t=r+1}^{T} \frac{\mathcal{E}_t^2}{2\sigma_{t-1}^2}.$$

Thus the noise-scaled endpoint-error budget directly controls path accuracy. A vanishing sum gives vanishing path KL; at an individual shrinking-noise step, $\mathbb{E}\|M_t e_t\|^2 = o(\sigma_{t-1}^2)$ is a sufficient rate. The deterministic terminal comparison is characterized by endpoint distance: conditional on the history, $W_2^2(\delta_{\overline{u}_1}, \delta_{u_0}) = \|\overline{u}_1 - u_0\|^2$ in normalized coordinates, where δ_a denotes a point mass at a.

J Datasets and Degradation Protocol

Following the experimental setup in Section 4, this section specifies preprocessing, complex-noise synthesis, and task-specific observations. The same realized degradation is supplied to every competing method within each experiment.

J.1 Datasets and Preprocessing

The evaluation uses central crops of 256×256 for Chikusei and Pavia University and 128×128 for Houston 2018 (Yokoya & Iwasaki, 2016; Huang & Zhang, 2009; Xu et al., 2019). Each HSI is linearly normalized to $[0, 1]$. For pansharpening, bands in the 350–1035 nm range are retained to match the support of the PAN spectral response.

J.2 Complex-Noise Synthesis

Let $X_{0,b}$ denote clean band b and let $P_b = N^{-1}\|X_{0,b}\|_F^2$ be its signal power, where $N = N_h N_w$. The dense component is sampled as $\epsilon_b^{\mathrm{syn}} \sim \mathcal{N}(0, \sigma_b^2 I_N)$, with $\mathrm{SNR}_b \sim \mathcal{U}(1, 1.2)$ dB and $\sigma_b^2 = P_b / 10^{\mathrm{SNR}_b/10}$. A random permutation of the interior bands assigns approximately 30% to stripe corruption, 30% to dead lines, and 30% to impulse corruption. The remaining bands contain the dense component.

For a stripe-corrupted band, $10 + \lceil 20U \rceil$ sampled columns receive offsets drawn uniformly from $[-0.25, 0.25]$. For a dead-line band, $5 + \lceil 10U \rceil$ sampled columns are set to zero. Each count uses a draw $U \sim \mathcal{U}(0, 1)$. For impulse corruption, a fraction drawn from $\mathcal{U}(0.1, 0.3)$ is replaced by zeros and ones with equal probability. The degradation-generation seed is 2025. These perturbations combine dense noise with spatially structured and impulsive defects. The two-factor model in Section 3.2 adapts to their local amplitudes and scale through its inferred sparse-state probabilities and scale distributions.

Using this complex-noise process, we construct the task-specific observations as follows.

Denoising applies the complex-noise process to the complete HSI. Missing-band restoration removes a uniformly sampled subset of $\lceil 0.25B \rceil$ distinct bands and applies complex noise to the acquired bands. The resulting observation operators are given in Appendix F.3.

For blind noisy pansharpening, the PAN image is generated by interpolating the IKONOS response at the retained wavelengths, clipping negative response values, and normalizing the weights. The LR HSI is generated with a bandwise 41×41 MTF filter using $G_{\mathrm{Nyq}} = 0.3$, followed by bicubic downsampling by a factor of four. Complex noise is then added to the LR observation (Wald et al., 1997; Yokoya et al., 2017). The generating PAN response is reserved for observation synthesis. During restoration, the spatial blur and downsampling operator is known, while the PAN response

and its calibration are estimated from the observed LR–PAN pair through the physical constraints in Appendix E.4.

K Implementation Details and Reproducibility

This section specifies the numerical settings and parameters used in the restoration experiments. All tasks share the terminal-centered cosine model, variational observation inference, and Gaussian-channel transition. The ranks, loss coefficients, and optimization schedules used in each experiment are listed below.

K.1 Reference Settings and Reverse-Grid Indexing

The reference forward process uses $T = 3000$ steps with linearly spaced variances from 10^{-4} to 5×10^{-3}. The corruption model uses Student-t degrees of freedom $\nu = 4$, default mixture-prior center $\pi_0 = (0.94, 0.06)$, default concentration $\kappa_0 = 64$, and noise-prior parameters $a_0 = 2$, $b_0 = 10^{-4}$. The reference scene-optimization learning rate is 5×10^{-4}, and the reference ranks are $(R_s, R_\lambda, R_{\text{temp}}) = (16, 6, 3)$. Sampling follows the fully stochastic two-branch transition in Eq. (18). The released Pavia University denoising and Houston 2018 missing-band examples set $\kappa_0 = N_h N_w B$, giving $256 \times 256 \times 103$ and $128 \times 128 \times 48$, respectively, where B denotes the full spectral band count before band removal.

The derivations above use the full-grid order $t = T, \dots, 1$. For a selected reverse grid, write its diffusion timesteps as $1 \le \tau_1 < \cdots < \tau_K \le T$; execution proceeds from τ_K to τ_1. At stage $j \ge 2$, the transition maps u_{τ_j} to $u_{\tau_{j-1}}$ and uses target scale $\sigma_{\tau_{j-1}}$. The temporal dictionary uses the actual number of fitting stages K: Eq. (16) is evaluated with (T, t) replaced by (K, j). Similarly, the stage-based progress is $p_j^{\text{grid}} = (K - j)/(K - 1)$ for $K > 1$. This convention ties the diffusion variance to the selected diffusion time and the temporal prior to the executed fitting grid.

Uniform reverse grids are constructed either with the requested number of stages or with an integer stride. For integer-stride sampling, K denotes the number of selected timesteps. The implementation uses zero-based diffusion-array indices, so array index $\tau_j - 1$ corresponds to diffusion timestep τ_j.

K.2 Projected Precision and Measurement Confidence

At each observation-refresh stage, the variational state and scale moments determine the precision of acquired measurement i:

$$\Lambda_{i,t} = \left(\tau_{b(i)}^{-1} + r_{is} v_{b(i)} \frac{b_{\omega i}}{a_\omega - 1} \right)^{-1}, \qquad a_\omega = (\nu + 1)/2 > 1.$$

The conditional sampler uses $\mathsf{W}_t = \operatorname{diag}(\Lambda_t)/4$ to construct the Gaussian observation channels in Eq. (13). Endpoint fitting uses the dimensionless confidence $R_{i,t} = \Lambda_{i,t}/\tau_{b(i)}$ defined in Appendix E. The inferred moments and confidence are detached during endpoint optimization, and the observation statistics and channels are reused between scheduled refreshes. For missing-band restoration, inference is restricted to acquired measurements, and image-grid precision and confidence arrays contain zeros at missing bands.

Numerical evaluation bounds the mixture variance, dense-noise precision, and positive observation precisions to $[\epsilon, \epsilon^{-1}]$, with $\epsilon = 10^{-6}$. The sparse-scale variance has a lower bound of 10^{-5}, and the evaluated confidence ratios are restricted to $[0, 1]$.

K.3 Parameter Settings

Tables 6 and 7 list the main parameters used for the three restoration tasks. Here "stride" and "exact" refer to the two reverse-grid constructions defined above. For the loss ablation in Table 5, only the four displayed loss terms are varied. The task-specific physical constraint $\mathcal{L}_{\text{phys},t}$ is retained for all pansharpening variants, while it is identically zero for denoising and missing-band restoration, as defined in Appendix E.4.

All experiments use restoration seed 3407. The mixture-prior concentration is set to $\kappa_0 = N_h N_w B$ for denoising and band-missing restoration and to $\kappa_0 = 64$ for pansharpening. Endpoint fitting uses

Table 6: Main model and sampling parameters. The rank tuple denotes $(R_s, R_\lambda, R_{\text{temp}})$; m is the interval between variational-inference updates.

Task	Dataset	Grid	K	$(R_s, R_\lambda, R_{\text{temp}})$	β_{EMA}	m
Denoising	Chikusei	stride	1000	$(40, 16, 2)$	0.985	200
	Pavia University	exact	1000	$(32, 12, 4)$	0.985	210
	Houston 2018	exact	900	$(32, 12, 4)$	0.988	90
Band missing	Chikusei	stride	1000	$(32, 8, 2)$	0.985	200
	Pavia University	exact	980	$(24, 8, 3)$	0.98	200
	Houston 2018	exact	1000	$(24, 8, 4)$	0.99	150
Pansharpening	Chikusei	exact	581	$(40, 12, 4)$	0.96	20
	Pavia University	exact	969	$(32, 12, 2)$	0.95	20
	Houston 2018	exact	1000	$(32, 12, 4)$	0.97	20

Table 7: Main objective-function parameters. The LR consistency weight is used only for pansharpening.

Task	Dataset	λ_{spec}	$\lambda_{\text{rel},0}$	q_{rel}	χ_{b}	λ_{traj}	λ_{LR}
Denoising	Chikusei	0.022	33.95	4	4	0	–
	Pavia University	0.05	5.55	1.7	1	0.01	–
	Houston 2018	0.08	7.125	2	0	0.012	–
Band missing	Chikusei	0.05	60.6	4	4	0.005	–
	Pavia University	0.1075	89.925	2	1	0.003	–
	Houston 2018	0.1	10	2	0	0.01	–
Pansharpening	Chikusei	0.05	90	4	4	0.01	3.0
	Pavia University	0.05	1.51125	2	1	0.01	2.32
	Houston 2018	0.05	3.75	2	0	0.01	0.5

one update per stage and two updates during the final 20% of reverse progress. For Houston 2018, $\pi_0 = (0.995, 0.005)$ for denoising and $(0.96, 0.04)$ for pansharpening.

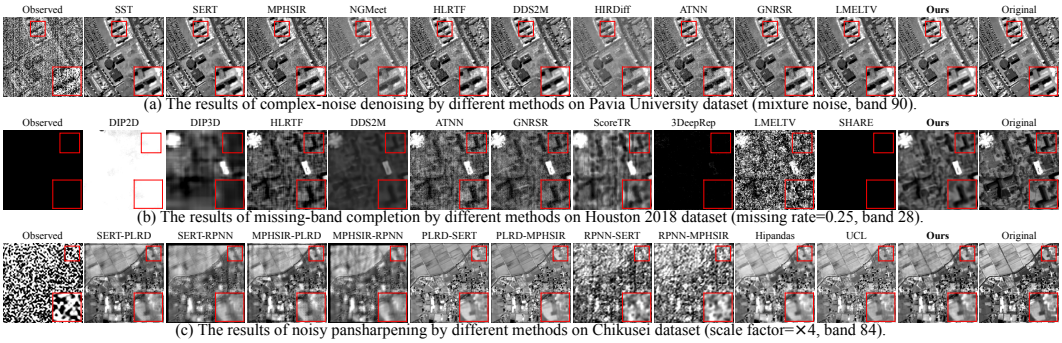
K.4 Implementation Sources of Competing Methods

We evaluate the competing methods using their author-released implementations. Code sources are provided in the footnotes below. For supervised denoisers and pretrained diffusion priors, we use publicly released weights without retraining or fine-tuning them on the evaluation scenes. Methods requiring scene-specific optimization are run separately on each observed cube.

Supervised denoisers. SST (Li et al., 2023a)[2] and SERT (Li et al., 2023b)[3] use the author-released complex-noise checkpoints (`model_complex.pth`). MP-HSIR (Wu et al., 2025)[4] uses `Natural_scene.ckpt` together with the CLIP ViT-B/32 encoder required by its implementation.[5]

Optimization-based and self-supervised methods. NGMeet (He et al., 2019)[6], HLRTF (Luo et al., 2022)[7], DDS2M (Miao et al., 2023)[8], ATNN (Peng et al., 2025)[9], and GNRSR (Tan et al., 2025)[10] are run using their official solvers or scene-specific optimization procedures, without external

[2] https://github.com/MyuLi/SST
[3] https://github.com/MyuLi/SERT
[4] https://github.com/ZhehuiWu/MP-HSIR
[5] https://github.com/openai/CLIP
[6] https://github.com/quanmingyao/NGMeet
[7] https://github.com/YisiLuo/HLRTF
[8] https://github.com/miaoyuchun/DDS2M
[9] https://github.com/andrew-pengjj/adaptive_tensor_nuclear_norm
[10] https://github.com/wangzhi-swu/GNRSR



(a) The results of complex-noise denoising by different methods on Pavia University dataset (mixture noise, band 90).

(b) The results of missing-band completion by different methods on Houston 2018 dataset (missing rate=0.25, band 28).

(c) The results of noisy pansharpening by different methods on Chikusei dataset (scale factor=×4, band 84).

Figure 5: Full visual comparisons for denoising, missing-band completion, and pansharpening.

restoration checkpoints. In particular, the DDS2M comparison results are obtained using the authors' official repository.

For missing-band restoration, DIP2D and DIP3D use the Deep Hyperspectral Prior implementation (Sidorov & Hardeberg, 2019) [11]. SHARE (Xie et al., 2026)[12], ScoreTR (Cheng et al., 2026c)[13], 3DeepRep (Li et al., 2026)[14], and LMELTV (Hu et al., 2026)[15] are likewise optimized on each observed cube. LMELTV is also evaluated for denoising.

Pretrained diffusion priors. HIR-Diff (Pang et al., 2024)[16] and PLRD (Rui et al., 2024)[17] use the pretrained diffusion checkpoint `I190000_E97_gen.pth` distributed through DDPM-CD,[18] following the respective implementation instructions. The pretrained diffusion weights remain fixed, while any required scene-specific guidance modules follow the corresponding method's optimization procedure.

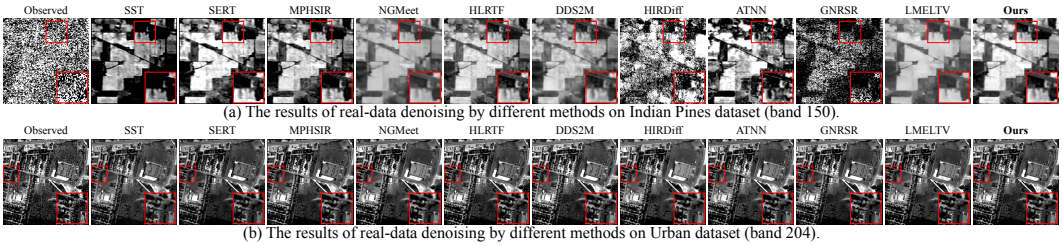
Pansharpening methods and pipelines. RPNN (Guarino et al., 2024)[19] uses the official implementation and author-released sensor weights. Hipandas (Xu et al., 2025)[20] uses its official zero-shot optimization procedure without an external HSI checkpoint. The two-stage baselines combine SERT or MP-HSIR with PLRD or RPNN in the order indicated in Table 3, using the same component implementations and pretrained weights. Hipandas and UCL (Xiao et al., 2026)[21] are evaluated as joint restoration methods.

L Additional Main Results

L.1 Full Visual Comparisons

Figures 5 and 6 present the full visual comparisons on synthetic and real HSI data.

[11] https://github.com/acecreamu/deep-hs-prior
[12] https://github.com/xuwayyy/SHARE
[13] https://github.com/CZY-Code/ScoreTR
[14] https://github.com/Kendradra/3DeepRep
[15] https://github.com/wangzhi-swu/LMELTV
[16] https://github.com/LiPang/HIRDiff
[17] https://github.com/earth-insights/PLRDiff
[18] https://github.com/wgcban/ddpm-cd
[19] https://github.com/giu-guarino/R-PNN
[20] https://github.com/shuangxu96/Hipandas
[21] https://github.com/Jin-liangXiao/UCL



(a) The results of real-data denoising by different methods on Indian Pines dataset (band 150).

(b) The results of real-data denoising by different methods on Urban dataset (band 204).

Figure 6: Full visual comparisons on Indian Pines and Urban.

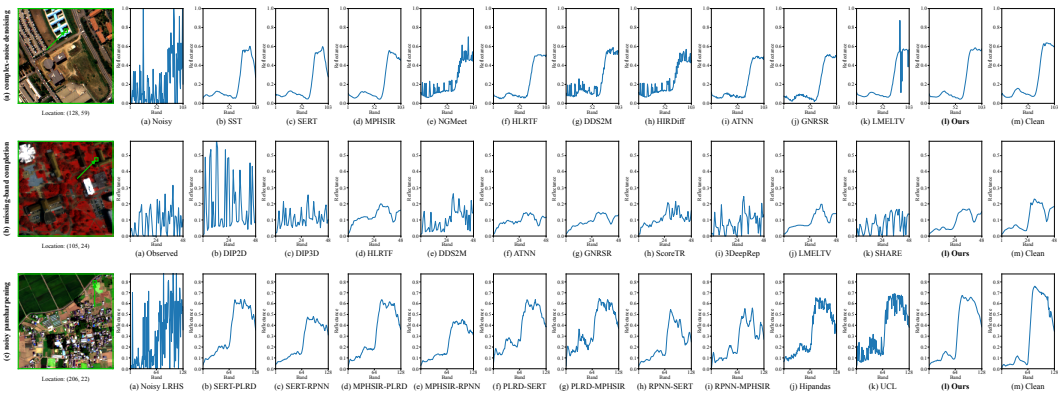


Figure 7: Spectral profile comparisons for denoising (top), missing-band completion (middle), and pansharpening (bottom).

L.2 Spectral Profile Comparisons

Figure 7 reports pixel-wise spectral profiles at the locations used in the main visual comparisons. From top to bottom, the examples correspond to Pavia University denoising, Houston 2018 missing-band completion, and Chikusei pansharpening. Each spectrum is extracted at the pixel marked as *Location* in the corresponding RGB image. The restored spectrum of VITTA follows the clean signature more closely than competing methods on all three tasks, which is consistent with the quantitative gains in the main tables.

L.3 Real No-Reference Restoration

Table 8 complements the real-data visual comparisons in Figure 4. VITTA achieves the lowest composite rank on both Indian Pines and Urban. The NIQE and spectral-roughness statistics provide complementary no-reference indicators for interpreting the spatial appearance and spectral regularity of the restored images. Rough. denotes spectral roughness. The composite rank averages the NIQE and roughness ranks within each dataset. Lower values are preferred for all three measures. The best and second-best values are shown in bold and underlined, respectively.

L.4 Computational Complexity and Runtime

Computational complexity. Let H and W denote the restored spatial dimensions, $N = HW$ the number of spatial pixels, and B the number of bands. The spatial, spectral, and temporal Tucker ranks are R_s, R_λ, and R_{temp}, respectively. Let K denote the actual number of reverse stages, Q the total number of endpoint-optimization calls, including provisional reference-fitting updates, and N_{VI} the number of variational-inference calls. The effective local variational-update budget is denoted by L_{vem}.

Table 8: No-reference restoration results on Indian Pines and Urban.

	Indian Pines			Urban		
Method	NIQE ↓	Rough. ↓	Rank ↓	NIQE ↓	Rough. ↓	Rank ↓
SST	9.4364	0.0193	6.0	5.7098	0.0189	8.5
SERT	9.9584	0.0251	8.0	5.8877	0.0190	9.5
MP-HSIR	8.4385	0.0176	3.5	5.2343	0.0186	6.0
NGMeet	9.3532	0.0265	8.5	6.7942	0.0089	7.0
HLRTF	8.1061	0.0221	4.75	5.5377	0.0092	5.0
DDS2M	8.6076	0.0277	8.5	4.9792	0.0095	4.0
HIRDiff	10.6082	0.0264	9.5	6.4521	0.0122	8.5
ATNN	7.7500	0.0221	3.75	5.4668	0.0091	4.0
GNRSR	7.8253	0.0252	5.5	5.5524	0.0104	6.5
LMELTV	12.1515	**0.0165**	6.0	6.4801	**0.0051**	5.5
Ours	**7.7490**	0.0179	**2.0**	**4.5037**	0.0056	**1.5**

Let $F_{\mathrm{skip}}(H, W)$ be the cost of one skip-network forward and backward pass. Ignoring fixed architectural constants, the principal arithmetic cost of an endpoint update is

$$\begin{aligned}\mathcal{F}_{\mathrm{upd}} = R_s F_{\mathrm{skip}}(H, W) + N R_s R_\lambda + N R_\lambda B \\ + R_s R_\lambda B + R_{\mathrm{temp}}^2 R_s R_\lambda + NB.\end{aligned}$$

The reconstruction is evaluated again after the optimizer step, adding a forward pass without changing this asymptotic order. The temporal dictionary and curvature Gram are precomputed once in $\mathcal{O}(KR_{\mathrm{temp}}^2)$ time. Consequently, curvature evaluation within an update depends on the temporal rank rather than on reconstruction of the complete K-stage core sequence.

For denoising and missing-band restoration, let B_{obs} denote the number of acquired bands, with $B_{\mathrm{obs}} = B$ for denoising. The principal computational cost is

$$\mathcal{C}_{\mathrm{direct}} = \mathcal{O}\big(KR_{\mathrm{temp}}^2 + Q\mathcal{F}_{\mathrm{upd}} + N_{\mathrm{VI}} L_{\mathrm{vem}} N B_{\mathrm{obs}} + KNB\big) + \mathcal{C}_{\mathrm{stat}},$$

where $\mathcal{C}_{\mathrm{stat}}$ collects robust initialization, median/MAD, quantile, and diagnostic-statistic costs that are not represented by the pointwise variational-update term. The reconstruction and sampling still operate on all B bands in the missing-band task.

For pansharpening, let $M = H_{\mathrm{LR}} W_{\mathrm{LR}}$ be the number of LR spatial pixels. Let $\mathcal{D}_1$ denote the cost of applying the spatial forward operator and its adjoint to one band, and let $\mathcal{D}_B$ denote the corresponding cost for a B-band cube. Let E be the number of refreshes requiring per-band weighted eigendecompositions. In the general dense case, these decompositions cost $\mathcal{O}(EBM^3)$, while the modal projections and lifts cost $\mathcal{O}(KBM^2)$. The initial spatial Gram construction and decomposition cost $\mathcal{O}(M\mathcal{D}_1 + M^3)$. With I_{pan} response-initialization iterations and M_{fit} sampled LR spatial points, the leading cost is

$$\begin{aligned}\mathcal{C}_{\mathrm{pan}} = \mathcal{O}\Big(KR_{\mathrm{temp}}^2 + M\mathcal{D}_1 + M^3 + I_{\mathrm{pan}} M_{\mathrm{fit}} B \\ + Q(\mathcal{F}_{\mathrm{upd}} + \mathcal{D}_B) + N_{\mathrm{VI}} L_{\mathrm{vem}} MB \\ + EBM^3 + K(\mathcal{D}_B + BM^2)\Big) + \mathcal{C}_{\mathrm{stat}}.\end{aligned}$$

For Pavia University pansharpening, the spectral proximal correction adds $\mathcal{O}(Q(B^3 + NB^2))$. Here latent inference is performed on the LR observation. The per-band factor cache requires $\mathcal{O}(BM^2)$ storage, which may reside in host memory. Shared-factor or orthogonal-row cases can reduce the weighted-channel construction cost.

A single feed-forward FLOP count does not characterize the complete restoration workload across these methods. We therefore report wall-clock measurements alongside the configuration-dependent stage and optimization budgets.

Measurement protocol and scope. Measurements use an NVIDIA RTX 4090 and an AMD EPYC 7763 CPU, with Python 3.11, PyTorch 2.5.1+cu121, and four OpenMP/MKL threads. Each

Table 9: VITTA full restoration time and peak memory. Times are medians of three timed runs; GPU and RSS values are maxima of the three per-run peaks. Q includes provisional reference-fitting updates. Chikusei and PaviaU use 256×256 output grids; Houston18 uses 128×128.

Task	Dataset	Bands	K	Q	Time (s)	GPU peak (GiB)	RSS peak (GiB)
Denoising	Chikusei	128	1000	1203	1268.9	12.76	1.69
Denoising	PaviaU	103	1000	1203	1118.2	11.93	1.64
Denoising	Houston18	48	900	1087	567.0	4.01	1.55
Band missing	Chikusei	128	1000	1203	1071.2	12.18	1.69
Band missing	PaviaU	103	980	1179	851.3	11.35	1.64
Band missing	Houston18	48	1000	1205	230.0	3.56	1.58
Pansharpening	Chikusei	128	581	721	1427.0	12.46	9.77
Pansharpening	PaviaU	103	969	1201	1680.8	11.76	9.74
Pansharpening	Houston18	48	1000	1239	470.1	4.15	1.66

configuration uses one warmup followed by three independent timed runs. The reported time is the median of the three totals. For multi-stage methods, stage times are summed within each run before taking the median. All timed runs are retained. Ground-truth evaluation and result serialization are excluded.

For VITTA, timing starts with the input resident in host memory and ends after restoration completes and CUDA execution is synchronized. The measured interval includes model initialization, observation setup, PAN-response initialization, spatial-Gram and bridge construction where applicable, and the complete reverse process.

For competing methods, the timing boundaries are method-specific. Pretrained denoisers report patch-based inference after data and checkpoint loading. HIR-Diff reports its sampling loop. DDS2M reports its reverse-diffusion segment, excluding model initialization and observation setup. HLRTF, DIP2D/3D, SHARE, and 3DeepRep report per-image optimization. ATNN, GNRSR, NGMeet, and LMELTV are timed on the CPU. For Houston18, LMELTV uses the linearly prefilled observation, consistent with its restoration-quality evaluation. ScoreTR reports fixed-epoch training after spectral prefilling; separate reconstruction and evaluation are excluded. Its epoch budgets are 16, 51, and 101 for Chikusei, Pavia University, and Houston 2018, respectively.

For pansharpening, PLRD pipeline times include preprocessing computed from the current observation. This preprocessing is recomputed for each timed run rather than reused from an existing cache. UCL is evaluated on the GPU. For the MP-HSIR-to-PLRD pipeline on Pavia University, the configuration uses MP-HSIR with the fixed Pavia University denoising settings, followed by PLRD with $\eta_1 = 2$, $\eta_2 = 2$, rank three, and 500 accelerated sampling steps. Both the reported restoration-quality result and runtime correspond to this fixed configuration. The different timing boundaries should be considered when comparing latencies; loop-only measurements do not represent complete restoration time.

For VITTA, peak GPU memory and peak RSS are each the maximum of the three timed runs' individual peaks, excluding warmup. GPU memory denotes PyTorch-allocated memory during restoration, before quality evaluation. RSS denotes process resident-set size and includes host-memory factor caches. Memory is reported in GiB, where one GiB equals 2^{30} bytes. Some configurations exhibit substantial variability: VITTA Houston18 denoising takes 278.8–1003.0 seconds across the three runs, and Houston18 pansharpening takes 430.9–1400.1 seconds. The smallest recorded median in each comparison-table column is shown in bold.

Discussion and Limitations. The current implementation introduces additional runtime due to iterative endpoint optimization and periodic corruption-statistics updates. This computation supports scene-specific adaptation and reliability-aware restoration across different degradations, while the present design mainly prioritizes restoration quality and modeling flexibility. We leave the development of a more computationally efficient and lightweight implementation to future work.

Table 10: Recorded median denoising times in seconds. VITTA includes initialization and restoration; HIR-Diff and DDS2M report sampling-loop times. Other method-specific scopes are described in the text.

Method	Type	Chikusei	PaviaU	Houston18
SST	Pretrained network	29.4	19.4	5.4
SERT	Pretrained network	10.8	9.0	1.9
MP-HSIR	Pretrained network	32.3	24.2	11.2
HLRTF	Per-image factorization	16.0	10.1	5.8
HIR-Diff	Pretrained diffusion prior	**8.9**	**7.0**	**1.6**
DDS2M	Self-supervised diffusion	276.9	606.6	107.9
ATNN	Tensor nuclear norm	199.8	285.2	22.1
GNRSR	Low-rank and sparse	774.4	669.8	35.4
NGMeet	Nonlocal and low-rank	411.9	411.6	188.5
LMELTV	Per-image optimization	379.9	210.5	48.3
VITTA (Ours)	Self-supervised diffusion	1268.9	1118.2	567.0

Table 11: Recorded median missing-band restoration times in seconds. ScoreTR uses fixed epoch budgets of 16, 51, and 101 for Chikusei, PaviaU, and Houston18. VITTA reports full restoration time; method-specific timing scopes are described in the text.

Method	Type	Chikusei	PaviaU	Houston18
3DeepRep	Per-image deep low-rank	**5.1**	**3.0**	**1.2**
DIP2D	Per-image DIP	11.0	10.8	6.6
HLRTF	Per-image factorization	12.2	13.7	5.9
GNRSR	Low-rank and sparse	930.5	630.3	22.3
DDS2M	Self-supervised diffusion	278.9	191.4	115.4
DIP3D	Per-image 3D DIP	430.8	451.7	50.1
SHARE	Equivariant self-supervision	418.1	435.4	94.2
ATNN	Tensor nuclear norm	102.1	85.1	17.6
LMELTV	Per-image optimization	292.8	225.5	57.0
ScoreTR	Per-image score training	4005.5	7875.1	1731.5
VITTA (Ours)	Self-supervised diffusion	1071.2	851.3	230.0

Table 12: Recorded median noisy-pansharpening times in seconds. PLRD pipeline times include per-image preprocessing, recomputed for each timed run. RPNN pipelines include per-band fine-tuning, and Hipandas includes per-image pretraining and training. UCL is evaluated on the GPU. VITTA includes all initialization and restoration stages.

Method	Processing order	Chikusei	PaviaU	Houston18
SERT → RPNN	Denoise then pansharpen	99.1	**79.0**	84.6
MP-HSIR → RPNN	Denoise then pansharpen	121.4	94.1	69.0
SERT → PLRD	Denoise then pansharpen	104.1	109.4	42.8
MP-HSIR → PLRD	Denoise then pansharpen	214.9	93.4	65.1
PLRD → SERT	Pansharpen then denoise	82.6	151.2	50.3
PLRD → MP-HSIR	Pansharpen then denoise	192.6	125.6	75.7
RPNN → SERT	Pansharpen then denoise	111.7	82.5	52.4
RPNN → MP-HSIR	Pansharpen then denoise	143.8	97.8	20.8
Hipandas	Joint per-image optimization	73.7	110.5	41.2
UCL	Joint restoration	**71.6**	80.6	**15.6**
VITTA (Ours)	Single-trajectory optimization	1427.0	1680.8	470.1