WHITE-BOX PROMPT TRANSFORMERS: VARIATIONALLY GROUNDED PROMPT–ATTENTION COUPLING FOR UNIFIED IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can soft prompts in vision Transformers be made explainable? Prompt-based models have achieved remarkable success in image restoration, yet they remain largely opaque: the underlying Transformer operations and the mechanism by which prompts modulate attention are poorly understood. This work revisits *guided image restoration*, where an auxiliary modality A assists in restoring a target modality B. We interpret A as a prompt and formulate a tailored structure-tensor total variation (*STV*) model, whose gradient suggests a white-box correspondence to prompt–attention interactions. This provides a principled bridge between prompts and attention. In scenarios where A is unavailable, we abstract its role into learnable soft prompts, enabling end-to-end training within standard Transformer pipelines. By unrolling the gradient flow of the *STV* variational problem, we derive the *White-Box Prompt Transformer (WBPT)*, a cascaded architecture that embeds interpretability directly into attention operations. Extensive experiments on multiple benchmarks demonstrate that WBPT achieves state-of-the-art restoration performance while offering interpretable, controllable, and robust prompt–attention dynamics.

1 INTRODUCTION

Prompt-based Transformers have recently reshaped unified image restoration, enabling a single model to tackle diverse degradations through learnable soft prompts (Potlapalli et al., 2023). These prompts condition the restoration process by modulating attention mechanisms and consistently deliver strong empirical results (e.g., Jia et al., 2022; Kong et al., 2025). However, despite their success, prompt-based designs remain fundamentally opaque: the inner workings of the Transformer and the interaction between prompts and attention lack interpretability, limiting both theoretical understanding and practical controllability (Chefer et al., 2021; Jain & Wallace, 2019). This opacity impedes reliable deployment in trust-sensitive applications (Rudin, 2019).

This motivates our central question:

Can prompt-driven attention be explained from first principles, providing a theoretically grounded interpretation of the black box?

We draw inspiration from *guided image restoration*, where an auxiliary modality A (e.g., T_1-weighted MRI) provides structural guidance for restoring a target modality B (He et al., 2012; Li et al., 2016; Ehrhardt & Betcke, 2016a). In this setting, A acts as a prior, naturally analogous to a prompt guiding the restoration of B (Jia et al., 2022; Potlapalli et al., 2023). Since explicit auxiliary data are often unavailable (Havaei et al., 2016), we abstract the role of A into learnable soft prompts—trainable tokens that emulate auxiliary guidance through end-to-end optimization. This perspective reinterprets prompts not as heuristic inputs but as principled surrogates for classical guidance (Li & Liang, 2021; Zhou et al., 2022).

Building on this analogy, we introduce a variational perspective on prompt-based restoration. Specifically, we cast guided restoration as a *structure-tensor total variation (STV)* problem (Chambolle & Pock, 2011; Lefkimmiatis et al., 2015). Through gradient analysis, we show that the optimization dynamics naturally align with a white-box attention mechanism, suggesting a formal link

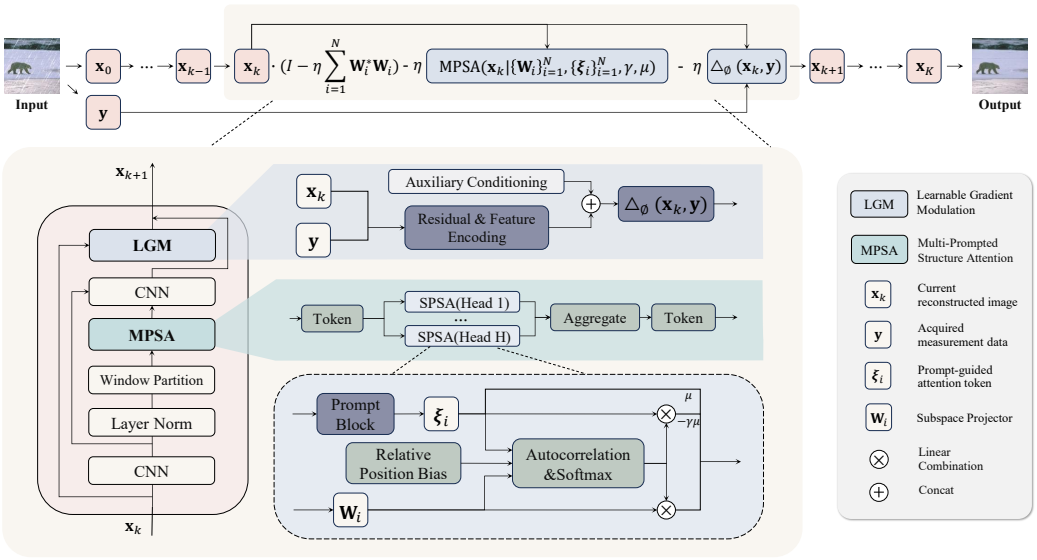

Figure 1: Overview of the White-box Prompt Transformer (WBPT). Image restoration is achieved by unrolling K gradient-flow steps. At iteration k (Eq. 8), the update combines a Multi-Prompted Structure Attention (MPSA) block with a learnable gradient-modulation (LGM) data-consistency term. MPSA consists of Single-Prompted Structure Attention (SPSA) heads: tokenized features interact with prompt tokens $\boldsymbol{\xi}_i$ and learnable projectors $\mathbf{W}_i$, are aggregated, and mapped back to image domain. Across stages, prompts interact with features at multiple levels, enriching structural context while preserving fidelity to the measurements.

between guidance priors (or their prompt surrogates) and Transformer attention (Yu et al., 2023; Wang et al., 2018; Meng et al., 2024). While the derivation involves approximations, it provides a principled foundation for understanding and designing interpretable prompt-driven restoration.

Crucially, this formulation not only offers interpretability but also suggests a concrete architectural design. By unrolling the gradient flow of the *STV* variational problem, we obtain the *White-Box Prompt Transformer (WBPT)* (Chen et al., 2015; Monga et al., 2021). Each Transformer layer corresponds to an optimization step, with every attention operation tied intuitively to terms in the underlying energy functional. WBPT thus unifies variational analysis with deep learning, embedding interpretability into the model without compromising performance.

Contributions.

- *Variational Perspective on Prompt-based Restoration:* Guided restoration is formulated as a tailored *STV* problem. Its optimization dynamics reveal a white-box attention mechanism, offering a principled explanation of how prompts influence Transformer attention.
- *White-Box Prompt Transformer:* A cascaded Transformer derived by unrolling the *STV* gradient flow, where each layer corresponds to an optimization step and attention operations align with terms in the underlying energy functional.
- *Bridging Classical and Modern AI:* The framework connects variational principles with deep prompt-based models, providing a foundation for interpretable image restoration and controllable attention mechanisms with clear theoretical grounding.
- *Empirical Validation:* WBPT achieves state-of-the-art results on multiple image restoration benchmarks while enabling transparent analysis of prompt–attention dynamics via rigorous, comprehensive visualization and controlled perturbation studies.

2 METHODS

In this section, we introduce WBPT, a variationally inspired framework for guided image restoration that interprets the guidance modality as *soft prompts* in a principled manner. WBPT integrates a tailored *STV* prior with learnable transformations $\mathbf{W}_i$ and soft prompt tokens $\boldsymbol{\xi}_i$, achieving restoration by unrolling K gradient-flow steps. The overall pipeline and information flow across the K cascaded stages are illustrated in Fig. 1, as depicted schematically.

2.1 Structured Modeling Framework for Guided Image Restoration

In guided image restoration, structural consistency across image modalities—where one modality (e.g., modality A) provides complementary information to enhance the restoration of another modality (e.g., modality B)—can be effectively exploited in practice. Rapidly acquired or higher-quality modality A images can serve as informative priors to guide the restoration of modality B (Ehrhardt & Betcke, 2016b). To systematically and rigorously model this guidance, we treat the modality A image as a *prompt*, explicitly encoding its anatomical information through dedicated operators to assist in restoring modality B (Potlapalli et al., 2023; Jia et al., 2022).

Formally, guided image restoration can be expressed as the following optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}(\mathbf{x};\boldsymbol{\xi}) := \mathcal{R}_1(\mathbf{x}) - \mu\mathcal{R}_2(\mathbf{x};\boldsymbol{\xi}), \tag{1}$$

where $\mathbf{x}:\Omega\to\mathbb{R}^d$ denotes the target image (modality B) to be restored, and $\boldsymbol{\xi}$ represents the feature embedding of the guidance image (modality A). The function space $\mathcal{X}$ is chosen appropriately (e.g., Sobolev space $H^1(\Omega;\mathbb{R}^d)$ or $L^2(\Omega;\mathbb{R}^d)$) to ensure well-definedness of the optimization and its variational derivatives (Evans, 2022). The functional $\mathcal{R}_1(\mathbf{x})$ encodes intrinsic priors of $\mathbf{x}$, while minimizing $-\mathcal{R}_2(\mathbf{x};\boldsymbol{\xi})$ enforces consistency between $\mathbf{x}$ and $\boldsymbol{\xi}$ in the transformed domain. The parameter $\mu>0$ balances this trade-off.

Specifically, we redesign an enhanced *STV* prior to represent $\mathcal{R}$, which not only characterizes the structural priors of the target image but also enforces consistency with guidance features $\boldsymbol{\xi}_i$ in the domain defined by $\mathbf{W}_i$. This motivates our proposed weighted prompt formulation:

$$\begin{aligned}\mathcal{R}\left(\mathbf{x};\,\{\mathbf{W}_i\}_{i=1}^N,\,\{\boldsymbol{\xi}_i\}_{i=1}^N\right) :=& \frac{1}{2}\sum_{i=1}^N\int_\Omega \mathrm{Tr}\big(\psi\big(\mathbf{W}_i\mathbf{x}(s)(\mathbf{W}_i\mathbf{x}(s))^\top\big)\big)\,ds \\ &- \mu\sum_{i=1}^N\int_\Omega \mathrm{Tr}\big(\psi\big(\boldsymbol{\xi}_i(s)(\mathbf{W}_i\mathbf{x}(s))^\top\big)\big)\,ds.\end{aligned} \tag{2}$$

Classical *STV* instantiates $\mathbf{W}_i$ as gradient operators capturing local structures, whereas nonlocal *STV* incorporates global interactions for superior performance. Motivated by this, we parameterize $\mathbf{W}_i$ as learnable global transformations via fully connected layers rather than local convolutional kernels (Wang et al., 2018). In parallel, $\{\boldsymbol{\xi}_i\}_{i=1}^N$ serve as prompts in the transformed domain. When explicit guidance images are unavailable, these prompts are relaxed into learnable soft tokens (Jia et al., 2022; Potlapalli et al., 2023). Finally, $\psi(\cdot)$ is a sparsity-inducing penalty, for which nonconvex forms such as $\psi(u)=\ln(1+u)$ are effective.

2.2 White-box Prompt Transformer via Variational Derivation

The gradient of the energy functional (2) is derived via variational calculus, resulting in an interpretable form:

$$\begin{aligned}\frac{\delta\mathcal{R}(\mathbf{x};\mathbf{W}_i,\boldsymbol{\xi}_i)}{\delta\mathbf{x}} \approx& \mathbf{W}_i^*\mathbf{W}_i\mathbf{x} + \gamma\mathbf{W}_i^*\mathbf{W}_i\mathbf{x}\cdot\mathrm{softmax}\big((\mathbf{W}_i\mathbf{x})^\top\mathbf{W}_i\mathbf{x}\big) \\ &- \mu\,\mathbf{W}_i^*\boldsymbol{\xi}_i - \gamma\mu\,\mathbf{W}_i^*\boldsymbol{\xi}_i\cdot\mathrm{softmax}\big((\mathbf{W}_i\mathbf{x})^\top\boldsymbol{\xi}_i\big).\end{aligned} \tag{3}$$

Here, $\mathcal{R}(\mathbf{x};\mathbf{W}_i,\boldsymbol{\xi}_i)$ denotes the i-th component of

$$\mathcal{R}\big(\mathbf{x};\,\{\mathbf{W}_i\}_{i=1}^N,\,\{\boldsymbol{\xi}_i\}_{i=1}^N\big) = \sum_{i=1}^N \mathcal{R}(\mathbf{x};\mathbf{W}_i,\boldsymbol{\xi}_i).$$

This gradient inspires the *Single-Prompted Structure Attention (SPSA)* module:

$$\mathrm{SPSA}(\mathbf{x}\mid\mathbf{W}_i,\boldsymbol{\xi}_i,\gamma,\mu) := \mathbf{W}_i\mathbf{x}\cdot\mathrm{softmax}\big((\mathbf{W}_i\mathbf{x})^\top\mathbf{W}_i\mathbf{x}\big) - \gamma\mu\,\boldsymbol{\xi}_i\cdot\mathrm{softmax}\big((\mathbf{W}_i\mathbf{x})^\top\boldsymbol{\xi}_i\big) + \mu\,\boldsymbol{\xi}_i. \tag{4}$$

For multiple prompts, we define the *Multi-Prompted Structure Attention (MPSA)* module:

$$\mathrm{MPSA}(\mathbf{x}\mid\{\mathbf{W}_i\},\{\boldsymbol{\xi}_i\},\gamma,\mu) := [\mathbf{W}_1^* \quad \cdots \quad \mathbf{W}_N^*]\begin{bmatrix}\mathrm{SPSA}(\mathbf{x}\mid\mathbf{W}_1,\boldsymbol{\xi}_1,\gamma,\mu)\\ \vdots \\ \mathrm{SPSA}(\mathbf{x}\mid\mathbf{W}_N,\boldsymbol{\xi}_N,\gamma,\mu)\end{bmatrix}. \tag{5}$$

This formulation provides an explicitly controllable, prompt-driven white-box attention mechanism with three functional components:

- *Self-Reconstruction Term:* $\mathbf{W}_i\mathbf{x} \cdot \mathrm{softmax}((\mathbf{W}_i\mathbf{x})^\top \mathbf{W}_i\mathbf{x})$, enhancing intrinsic feature coherence via self-expression.
- *Prompt-Alignment Term:* $-\lambda_1 \boldsymbol{\xi}_i \cdot \mathrm{softmax}((\mathbf{W}_i\mathbf{x})^\top \boldsymbol{\xi}_i)$, introducing a repulsive force to prevent trivial imitation while enabling structural adaptation.
- *Prompt-Bias Term:* $+\lambda_2 \boldsymbol{\xi}_i$, injecting prior knowledge as a static inductive bias to ensure faithful restoration.

2.3 Cascaded Transformer Architecture via Gradient Flow Unrolling

To optimize (2) while enforcing consistency with measurements $\mathbf{y}$, we consider the continuous-time gradient flow:

$$\frac{\partial \mathbf{x}(t)}{\partial t} = -\left(\frac{\delta \mathcal{R}}{\delta \mathbf{x}}(\mathbf{x}(t)) + \Delta(\mathbf{x}(t), \mathbf{y})\right), \tag{6}$$

where $\Delta(\mathbf{x}(t), \mathbf{y})$ denotes the gradient of the data fidelity term.

Discretizing via explicit Euler with step size η gives:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\left(\frac{\delta \mathcal{R}}{\delta \mathbf{x}}(\mathbf{x}_k) + \Delta(\mathbf{x}_k, \mathbf{y})\right), \tag{7}$$

where k indexes the iteration. Substituting the MPSA module (5), the update becomes:

$$\mathbf{x}_{k+1} = \left(\boldsymbol{I} - \eta \sum_{i=1}^{N} \mathbf{W}_i^* \mathbf{W}_i\right)\mathbf{x}_k - \eta \mathrm{MPSA}\left(\mathbf{x}_k \mid \{\mathbf{W}_i\}, \{\boldsymbol{\xi}_i\}, \gamma, \mu\right) - \eta \Delta_\phi(\mathbf{x}_k, \mathbf{y}). \tag{8}$$

In practice, the forward degradation model is often unknown, so Δ cannot be computed explicitly. We replace it with a learnable data-consistency term $\Delta_\phi(\mathbf{x}_k, \mathbf{y})$, parameterized by ϕ. This module captures the discrepancy between $\mathbf{x}_k$ and $\mathbf{y}$ while interacting with $\{\boldsymbol{\xi}_i\}$ and $\{\mathbf{W}_i\}$.

Unrolling K iterations of this process yields a deep cascaded network alternating between prompt-driven attention blocks and learnable data-consistency modules.

3 Experiments

We evaluate our WBPT against both general-purpose restoration methods and specialized All-in-One approaches (Table 1). Averaged across tasks, WBPT raises the mean PSNR from 30.16 to 31.02,dB, narrowing the gap to PromptIR while maintaining transparency and controllability.

To address the lack of multi-scale processing in WBPT—an element shown to be critical in models such as Restormer—we introduce WBPT†. This variant augments WBPT's interpretable attention blocks with a pyramid pathway that provides multi-scale feature aggregation. Although the aggregator is currently implemented as a learned, black-box component, the core attention mechanism remains fully white-box. A complete white-box formulation of multi-scale processing is under development. Empirically, WBPT† matches PromptIR on average and exceeds it on selected tasks.

Task-level results further highlight the advantages of the proposed framework. On Rain100L (deraining), WBPT† surpasses PromptIR, and Fig. 2 confirms removal of rain streaks with diverse orientations more thoroughly than WBPT, producing cleaner outputs. On SOTS (dehazing), while WBPT† falls short of PromptIR in PSNR, it demonstrates the benefit of multi-scale modeling; as shown in Fig. 3, haze removal is clearer and scene details are better preserved. On BSD68 (denoising), WBPT† achieves PSNR comparable to PromptIR and yields consistently higher SSIM.

Datasets. For denoising, training is conducted on BSD400 (Arbelaez et al., 2010) and WED (Ma et al., 2016) with Gaussian noise levels $\sigma \in \{15, 25, 50\}$, and evaluation is performed on BSD68 (Martin et al., 2001) and Urban100 (Huang et al., 2015). For deraining, we use Rain100L (200 training / 100 test images) (Yang et al., 2020). For dehazing, training is on SOTS (72,135 images)

Table 1: Comparison in the All-in-One restoration setting. Results are reported as PSNR/SSIM. Results are reported as PSNR/SSIM. Within each block (single-scale vs. multi-scale), the best and second-best are **boldfaced** and underlined, and gray shading indicates white-box models. Overall, WBPF yields a marked improvement over WBP, while WBPF† achieves performance comparable to PromptIR and surpasses it on several tasks.

Method	Dehazing SOTS	Deraining Rain100L	Denoising (BSD68) σ=15	σ=25	σ=50	Avg.
Single-scale methods						
BRDNet	23.23/0.895	27.42/0.895	32.26/0.898	29.76/0.836	26.34/**0.836**	27.80/0.843
FDGAN	24.71/0.924	29.89/0.933	30.25/0.910	28.81/0.868	26.43/0.776	28.02/0.883
AirNet	27.94/0.962	34.90/0.967	**33.92/0.933**	**31.26/0.888**	**28.00**/0.797	**31.20/0.910**
WBT	27.40/0.958	32.13/0.940	33.17/0.923	30.68/0.875	27.41/0.770	30.16/0.893
WBPT	**29.31/0.972**	**35.93/0.971**	33.66/0.929	31.01/0.881	27.72/0.781	31.02/0.907
Multi-scale methods						
LPNet	20.84/0.828	24.88/0.784	26.47/0.778	24.77/0.748	21.26/0.552	23.64/0.738
MPRNet	25.28/0.954	33.57/0.954	33.54/0.927	30.89/0.880	27.56/0.779	30.17/0.899
DL	26.92/0.391	32.62/0.931	33.05/0.914	30.41/0.861	26.90/0.740	29.98/ 0.875
PromptIR	**30.58/0.974**	36.37/0.972	**33.98**/0.933	**31.31**/0.888	28.06/0.799	**32.06**/0.913
WBPT†	29.94/0.970	**37.08/0.974**	33.86/**0.934**	31.28/**0.890**	**28.08/0.801**	32.05/**0.914**

Table 2: Deraining results on Rain100L in the single-task setting. Within each scale group (Single or Multi), best results are **boldfaced** and second-best are underlined; Gray indicates white-box models.

Scale	*Single-scale methods*				*Multi-scale methods*				
Method	SIRR	AirNet	WBT	WBPT	MSPFN	LPNet	Restormer	PromptIR	WBPT†
PSNR	32.37	34.90	36.77	**38.70**	33.50	33.61	36.74	37.04	**38.54**
SSIM	0.926	0.977	0.977	**0.983**	0.948	0.958	0.978	0.979	**0.984**

and evaluation on SOTS (500 images) (Li et al., 2018). In the All-in-One setting, these datasets are combined to train a unified model, following the protocol of (Potlapalli et al., 2023).

Model and Training. WBPT is trained end-to-end using a 10-iteration white-box framework, where each iteration integrates a learnable gradient update with a Transformer-based prompt branch. Prompts are injected specifically at the sixth Transformer block in each iteration. Training is performed with the standard Adam optimizer (β_1=0.9, β_2=0.999) at a fixed learning rate of 1×10^{-4} for 120 epochs in total. We use random 128×128 crops with rotations and flips, optimizing with L2 (MSE) loss. The best checkpoint is selected based on validation performance.

3.1 Multiple Degradation All-in-One Results

We compare our white-box models with general-purpose restoration approaches and specialized All-in-One methods (Table 1). Averaged across tasks, WBPT raises the mean PSNR from 30.16 to 31.02 dB, narrowing the gap to PromptIR while preserving transparency and controllability. Moreover, our multi-scale white-box variant, WBPT†, performs on par with PromptIR and even surpasses it on certain tasks. On Rain100L for deraining, WBPT† outperforms PromptIR; visual comparisons in Fig. 2 show that, relative to WBT, WBPT more effectively removes rain streaks of diverse orientations, yielding cleaner rain-free results. On SOTS for dehazing, although WBPT† does not surpass PromptIR, it confirms the benefits of the multi-scale design; examples in Fig. 3 indicate clearer haze removal and more faithful scene restoration. On BSD68 for denoising, WBPT† attains PSNR comparable to PromptIR while overall delivering higher SSIM.

3.2 Single Degradation One-by-One Results

We evaluate PromptIR under the single-task setting, where a separate model is trained for each restoration task. This setting is intended to empirically verify that content-adaptive prompting via the prompt block is also effective for single-task networks. Table 2 reports the deraining results on standard datasets: our single-scale white-box WBPT consistently achieves the best performance, surpassing PromptIR and the multi-scale white-box variant WBPT†; relative to WBT (without prompts), WBPT delivers a 1.93 dB gain in PSNR. For dehazing and denoising, although WBPT† does not surpass PromptIR, it achieves comparable performance; see Tables 3 and 4.

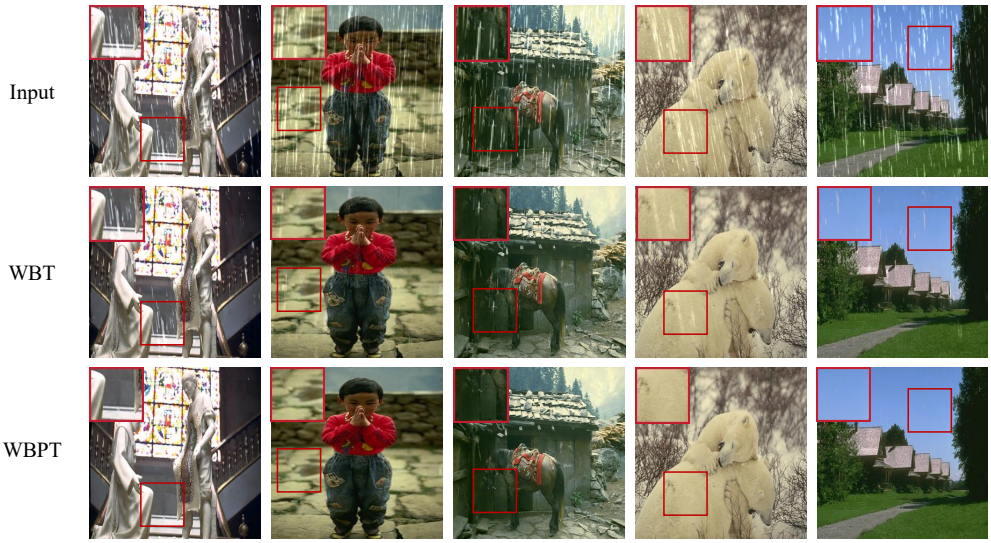

Figure 2: Deraining results under the All-in-One setting. Compared with WBT, our WBPT removes numerous residual rain streaks that WBT fails to eliminate, yielding cleaner backgrounds and sharper details (see red zoom-in boxes).

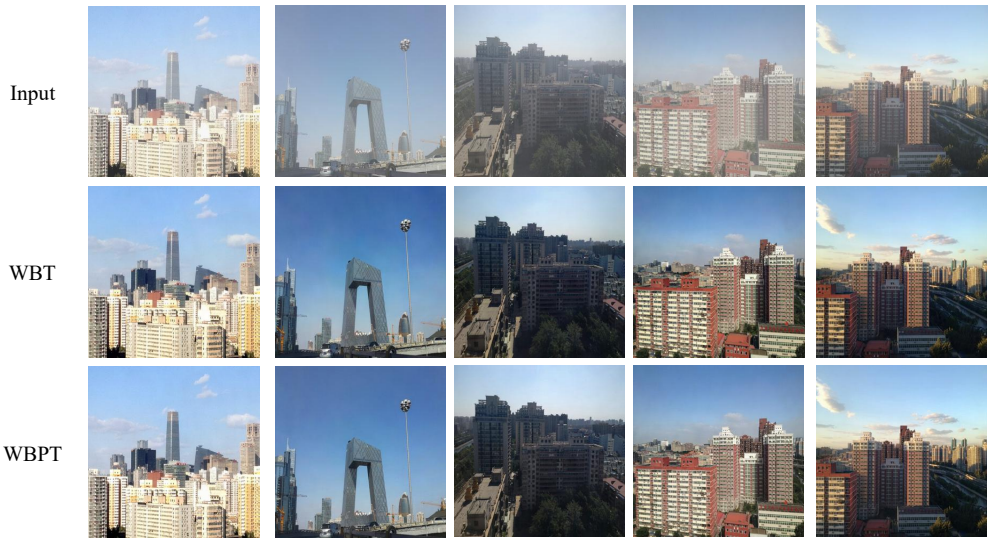

Figure 3: Dehazing results for all-in-one methods.Compared with WBT, our WBPT recovers clearer sky regions and sharper building edges, suppresses veiling glare and color cast, and yields more natural contrast and details across diverse urban scenes.

3.3 Attention Visualization

To examine differences in model focus, we visualize the last-layer multi-head attention of the white-box reconstruction model (WBPT) and the black-box method PromptIR. For each model, we extract the last-layer attention tensor $A \in \mathbb{R}^{H \times N \times N}$, average across heads, and further aggregate along the query dimension to obtain a single-channel response for each window. The window-wise responses are then reassembled into a full-image heatmap via reverse window stitching. To ensure a fair comparison, both models use identical inputs and visualization settings. As shown in Fig. 5, WBPT exhibits strong responses on object boundaries and structural regions, indicating a preference for image geometry and semantic content rather than directly following degradation textures; notably, this boundary-centric attention aligns with our STV objective (Sec. 2.1). In contrast, PromptIR's responses align with rain streaks and are more tightly coupled to the degradation pattern, suggesting a greater reliance on degradation-pattern detection.

Figure 4: Denoising results for all-in-one methods.

Table 3: Dehazing results on SOTS dataset in the single-task setting. Within each block (single-scale vs. multi-scale), the best and second-best are **boldfaced** and underlined, and gray shading indicates white-box models. PSNR/SSIM reported; higher is better.

Scale	*Single-scale methods*					*Multi-scale methods*			
Method	AODNet	FDGAN	AirNet	WBT	WBPT	EPDN	Restormer	PromptIR	WBPT†
PSNR	20.29	23.15	23.18	**28.72**	28.33	22.57	30.87	**31.31**	30.47
SSIM	0.877	0.921	0.900	0.961	**0.967**	0.863	0.969	**0.973**	0.972

3.4 t-SNE Analysis of Intermediate Representations

To analyze the intermediate representations of an all-in-one model across different degradations, we tap the *input* to the final convolution layer of the Transformer backbone during the forward pass. For each image, we apply global average pooling over the spatial dimensions to obtain a channel-wise embedding vector. We collect embeddings from three standard test sets—BSD68 denoising ($\sigma = 25$), Rain100L deraining, and SOTS-Outdoor dehazing—and project them to 2D using t-SNE.

Figure 6 compares the black-box PromptIR with our white-box WBPT under identical preprocessing and t-SNE settings: PromptIR (left) yields highly entangled embeddings with substantial cross-task overlap, whereas WBPT (right) forms well-separated clusters for the Noisy, Hazy, and Rainy samples, exhibiting tighter intra-cluster compactness and clearer inter-cluster margins. These results indicate that WBPT learns more discriminative, task-aware representations in the all-in-one setting.

3.5 Stability under prompt-parameter perturbations

To verify the stability of WBPT under prompt-parameter perturbations, we conduct a perturbation-sensitivity study in a controlled experimental setting and compare it with PromptIR. The test datasets are BSD68 (denoising), Rain100L (deraining), and SOTS-Outdoor (dehazing). We inject additive Gaussian noise *only* into the prompt parameters ($\sigma \in [0.001, 0.1]$), while keeping all other settings (e.g., the prompt insertion layer) identical to the previous configuration. The evaluation metric is the *average performance drop percentage* (lower indicates higher stability), averaged over multiple perturbation severities and the three datasets. The corresponding averages are summarized in Table 5; a representative qualitative comparison at $\sigma = 0.1$ is shown in Fig. 7.

From the visual results, PromptIR consistently exhibits *systematic contrast and color shifts* after perturbing the prompt, suggesting an undesirable coupling between the prompt representation and global imaging attributes. Under prompt-only perturbations, such global tone/contrast changes are not what degradation awareness is expected to primarily induce. In contrast, WBPT with $\sigma = 0.1$ still removes rain effectively while maintaining remarkably stable contrast and colors, indicating stronger robustness and better degradation–prompt decoupling.

Table 4: Denoising comparisons in the single-task setting on BSD68 and Urban100. Results are reported as PSNR/SSIM. Within each block (single-scale vs. multi-scale), the best and second-best are **boldfaced** and underlined, respectively. gray shading indicates white-box models. At the challenging noise level of $\sigma = 50$ on Urban100, our WBPT achieves a 0.39 dB improvement over WBT. Meanwhile, WBPT† attains performance comparable to PromptIR.

Method	BSD68			Urban100		
	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
Single-scale methods						
CBM3D	33.50/0.922	30.69/0.868	27.36/0.763	33.93/0.941	31.36/0.909	27.93/0.840
DnCNN	33.89/0.930	31.23/0.883	27.92/0.789	32.98/0.931	30.81/0.902	27.59/0.833
IRCNN	33.87/0.929	31.18/0.882	27.88/0.790	27.59/0.833	31.20/0.909	27.70/0.840
FFDNet	33.87/0.929	31.21/0.882	27.96/0.789	33.83/0.942	31.40/0.912	28.05/0.848
BRDNet	34.10/0.929	31.43/0.885	28.16/0.794	**34.42**/0.946	31.99/0.919	28.56/0.858
AirNet	**34.14/0.936**	**31.48/0.893**	**28.23/0.806**	34.40/0.949	**32.10**/0.924	**28.88/0.871**
WBT	33.59/0.930	30.92/0.882	27.85/0.793	33.43/0.956	30.42/0.924	27.18/0.858
WBPT	34.02/0.935	31.35/0.891	28.03/0.797	34.15/**0.963**	31.38/**0.937**	27.57/0.870
Multi-scale methods						
Restormer	34.29/0.937	31.64/0.895	28.41/0.810	34.67/**0.969**	32.41/0.927	29.31/0.878
PromptIR	**34.34**/0.938	**31.71/0.897**	**28.49/0.813**	**34.77**/0.952	**32.49/0.929**	**29.39/0.881**
WBPT†	34.31/**0.938**	31.60/0.895	28.36/0.811	34.76/0.952	32.27/0.927	29.08/0.877

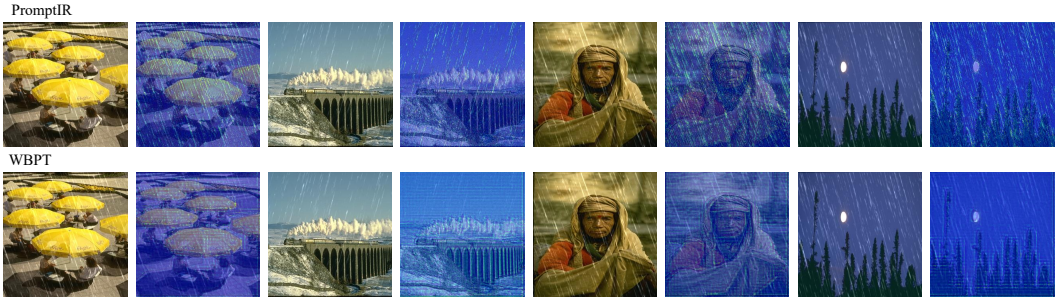

Figure 5: Attention-map visualizations for WBPT and PromptIR. Odd-numbered columns show input images; even-numbered columns show the corresponding attention maps. Top row: PromptIR; bottom row: WBPT. Attention heads and queries from the final layer are aggregated, and full-image heatmaps are reconstructed via reverse window stitching. WBPT focuses on object boundaries and main structures, whereas PromptIR emphasizes rain streaks.

Table 5: Evaluation of stability under prompt-parameter perturbations, reported as relative drops in PSNR and SSIM (lower is better). Gaussian noise with $\sigma \in [0.001, 0.1]$ is injected exclusively into the prompt parameters. Results are averaged over multiple severity levels on BSD68, Rain100L, and SOTS-Outdoor. WBPT exhibits smaller drops than PromptIR, indicating greater stability.

Model	Denoising		Deraining		Dehazing		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
PromptIR	-10.2%	-13.0%	-13.8%	-12.4%	-13.0%	-12.2%	-12.3%	-12.5%
WBPT	-3.05%	-0.35%	-1.02%	-0.64%	-2.18%	-0.05%	-2.08%	-0.31%

3.6 Guidance Modality Validation: Soft Prompt vs Real Guidance

We compare a learnable *soft prompt* with a proxy of real guidance (*hard*: image gradients $\rightarrow$ edge map plus Gaussian noise, $\sigma \in \{0.01, 0.02\}$). To control compute and isolate the modality effect, the *backbone is frozen* and only the prompt and fusion parameters are finetuned. For each test image, we report the paired difference $\Delta = \text{metric}_{\text{soft}} - \text{metric}_{\text{hard}}$; our goal is to assess the *relative* gap between soft and hard rather than absolute gains. Equivalence margins are pre-registered as ± 0.02 dB (PSNR) and ± 0.002 (SSIM), within which soft and hard are deemed practically equivalent.

Under the finetune-only setting (backbone frozen), soft and hard guidance behave nearly identically across denoising, deraining, and dehazing in our controlled evaluations. The paired differences

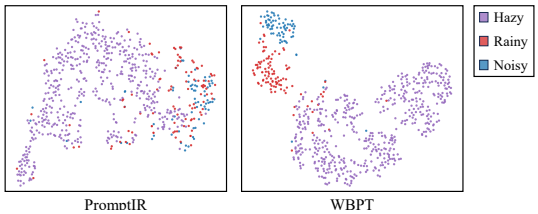

Figure 6: t-SNE visualization of degradation embeddings from PromptIR and WBPT. Colors denote degradation types. With identical inputs and t-SNE settings, WBPT yields clearly separable clusters by degradation type, whereas PromptIR is more entangled.

Figure 7: Qualitative comparison under Gaussian perturbation of prompt parameters ($\sigma = 0.1$). Odd-numbered columns show input images; even-numbered columns show restored results. Top row: PromptIR; bottom row: WBPT. WBPT preserves deraining quality, contrast, and colors, whereas PromptIR exhibits noticeable shifts, indicating lower robustness.

Δ = soft − hard are consistently tiny and remain within the pre-registered equivalence margins (± 0.02,dB PSNR / ± 0.002 SSIM) for both $\sigma = 0.01$ and $\sigma = 0.02$. While dehazing yields lower absolute scores—reflecting its higher difficulty—the relative gap between soft and hard stays stable, which is precisely the comparison this experiment aims to isolate.

Table 6: Soft vs. hard guidance under the finetune-only setting (backbone frozen). Metrics are PSNR and SSIM in separate columns; parentheses denote the change relative to the baseline in the same column. Gray denote the paired difference Δ = soft − hard; values within the pre-registered equivalence margins (±0.10 dB PSNR, ±0.002 SSIM) indicate practical equivalence. Results are reported for $\sigma \in \{0.01, 0.02\}$. While absolute scores for dehazing are lower due to its higher difficulty, the soft–hard gap remains small and stable across σ.

		Denoise		Derain		Dehaze	
σ	Task	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
0.01	WBT	33.59	0.930	36.77	0.977	28.72	0.961
	WBT+soft	33.61 (+0.02)	0.929 (−0.1%)	36.90 (+0.13)	0.978 (+0.1%)	27.83 (−0.89)	0.956 (−0.5%)
	WBT+hard	33.63 (+0.04)	0.929 (−0.1%)	36.89 (+0.12)	0.978 (+0.1%)	27.87 (−0.85)	0.958 (−0.3%)
	Δ (soft−hard)	−0.02	0.000	+0.01	0.000	−0.04	−0.002
0.02	WBT	33.59	0.930	36.77	0.977	28.72	0.961
	WBT+soft	33.62 (+0.03)	0.930 (+0.0%)	36.89 (+0.12)	0.978 (+0.1%)	27.83 (−0.89)	0.956 (−0.5%)
	WBT+hard	33.60 (+0.01)	0.929 (−0.1%)	36.89 (+0.12)	0.978 (+0.1%)	27.86 (−0.86)	0.958 (−0.3%)
	Δ (soft−hard)	+0.02	0.000	+0.00	0.000	−0.03	−0.002

4 Conclusion

In this work, we revisited prompt-based Transformers from a variational perspective and established a principled connection between prompts and attention. By casting guided image restoration as a *STV* problem, we derived a white-box attention mechanism that offers an interpretable foundation for prompt–attention coupling. Building on this formulation, we unrolled the gradient flow into WBPT, a cascaded architecture that integrates variational principles with modern prompt learning. Extensive experiments across diverse restoration tasks demonstrate that WBPT delivers competitive performance while maintaining transparent and robust prompt–attention dynamics. These findings point to a new direction for designing interpretable and controllable prompt-based models, with implications extending beyond image restoration to broader areas of vision and multimodal learning.

REFERENCES

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018.

Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. Modl: model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2019.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2010.

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

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12299–12310, 2021.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pp. 17–33. Springer, 2022.

Yunjin Chen, Wei Yu, and Thomas Pock. On learning optimized reaction diffusion processes for effective image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5261–5269, 2015.

Hyungjin Chung, Byeonghu Sim, Minyong Ryu, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Matthias J Ehrhardt and Marta M Betcke. Multicontrast MRI reconstruction with structure-guided total variation. *SIAM Journal on Imaging Sciences*, 9(3):1084–1106, 2016a.

Matthias J Ehrhardt and Marta M Betcke. Multicontrast MRI reconstruction with structure-guided total variation. *SIAM Journal on Imaging Sciences*, 9(3):1084–1106, 2016b.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Chun-Mei Feng, Huazhu Fu, Tianfei Zhou, Yong Xu, Ling Shao, and David Zhang. Deep multimodal aggregation network for MR image reconstruction with auxiliary modality. *arXiv Preprint arXiv:2110.08080*, 2021.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 469–477. Springer, 2016.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2012.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv Preprint arXiv:1902.10186*, 2019.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Ulugbek S Kamilov, Charles A Bouman, Gregery T Buzzard, and Brendt Wohlberg. Plug-and-play methods for integrating physical and learned models in computational imaging: theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1):85–97, 2023.

Yonatan Kawar, Michael Elad, Tomer Michaeli, and Stefano Ermon. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12809–12819, 2025.

Stamatios Lefkimmiatis, Anastasios Roussos, Petros Maragos, and Michael Unser. Structure tensor total variation. *SIAM Journal on Imaging Sciences*, 8(2):1090–1122, 2015.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2018.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17452–17462, 2022.

Xiang Lisa Li and Percy Liang. Prefix-tuning: optimizing continuous prompts for generation. *arXiv Preprint arXiv:2101.00190*, 2021.

Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *European Conference on Computer Vision*, pp. 154–169. Springer, 2016.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1833–1844, 2021.

Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 2007.

Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: new challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016.

David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, volume 2, pp. 416–423. IEEE, 2001.

Junying Meng, Faqiang Wang, and Jun Liu. Learnable nonlocal self-similarity of deep features for image denoising. *SIAM Journal on Imaging Sciences*, 17(1):441–475, 2024.

Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: prompting for all-in-one image restoration. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71275–71293, 2023.

Klaas P. Pruessmann, Markus Weiger, Michael B. Scheidegger, and Peter Boesiger. Sense: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 1999.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Proceedings of the ACM SIGGRAPH*, 2022.

Jo Schlemper, Jose Caballero, Joseph V. Hajnal, Anthony Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, 2018.

Jian Sun, Huibin Li, Zongben Xu, et al. Deep ADMM-net for compressive sensing MRI. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Yuchuan Tian, Jianhong Han, Hanting Chen, Yuanyuan Xi, Ning Ding, Jie Hu, Chao Xu, and Yunhe Wang. Instruct-IPT: all-in-one image processing transformer via weight modulation. *arXiv Preprint arXiv:2407.00676*, 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.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model-based reconstruction. In *Proceedings of the IEEE Global Conference on Signal and Information Processing*, pp. 945–948. IEEE, 2013.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: a general U-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17683–17693, 2022.

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5791–5800, 2020.

Haidi Yang, Chuang Zhang, Chun-Mei Feng, Jianfu Zhang, and Huazhu Fu. Multi-modal guidance-based deep unfolding network for MRI reconstruction. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9422–9457, 2023.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Wei Zhou, Xiaoyu Wang, Qi Zhang, Xiang Li, and Ke Chen. Multi-level modality fusion network for multi-contrast MRI reconstruction. *Computerized Medical Imaging and Graphics*, 2024.

APPENDIX

A Related Work **13**

B Ablation Experiment **14**

B.1 SPSA Component Ablations 14

B.2 Position of prompt blocks. 14

B.3 Complementarity between Prompt and Data-Consistency Modules. 14

C Variational Derivation of the Energy Functional **15**

C.1 Variation of the First Term 15

C.2 Variation of the Second Term 16

C.3 Combining Both Terms 17

C.4 Variational Derivative Approximation 17

D Limitation **17**

E Qualitative results **18**

F Algorithm **18**

G Transformer and Prompt Blocks in WBPT **19**

G.1 Prompt Block in WBPT Framework 20

G.2 Transformer Block in WBPT Framework 21

H Reproducibility Statement **22**

I The Use of Large Language Models (LLMs) **22**

A RELATED WORK

Transformer-based image restoration. Transformer architectures have advanced image restoration by modeling long-range dependencies, with strong single-task systems such as SwinIR, Restormer, and successors (Chen et al., 2021; Liang et al., 2021; Zamir et al., 2022; Chen et al., 2022; Wang et al., 2022). To curb task specialization, all-in-one models handle multiple degradations with a unified backbone (e.g., AirNet (Li et al., 2022)). We likewise pursue unified restoration but differ in how conditioning is defined and used within the network. Parallel to Transformer priors, diffusion-based restoration has recently shown competitive performance and broad applicability(Kawar et al., 2022; Chung et al., 2023; Saharia et al., 2022), and our formulation is complementary: it can inject structure-aware conditioning regardless of the underlying prior family.

Prompt-based conditioning for restoration. Prompt-based Transformers inject task or condition cues (e.g., degradation type, maps, or control signals) into a shared backbone to enable multi-task adaptation (Potlapalli et al., 2023; Li et al., 2022; Tian et al., 2024; Kong et al., 2025). Prevailing designs treat prompts as black-box tokens or channel-wise modulations (e.g., visual prompt tuning/adapters) that are concatenated to features and learned end-to-end, which limits interpretability and spatial control(Jia et al., 2022). In contrast, we treat prompts as **structural priors** and derive structure-aware prompt attention from a variational formulation via **learnable regularization gradients (LRG)**, providing mechanistic interpretability and spatial controllability (Yu et al., 2023).

Model-based deep learning and unrolled networks. Model-based approaches explicitly couple data fidelity with learned priors through algorithm unrolling (e.g., ADMM-Net, MoDL, VarNet) (Sun et al., 2016; Aggarwal et al., 2019; Hammernik et al., 2018; Adler & Öktem, 2018). Plug-and-play and RED families further connect optimization with learned denoisers (Venkatakrishnan et al., 2013; Yang et al., 2022; Ulyanov et al., 2018; Kamilov et al., 2023). Our design follows this lineage: we unfold a variational objective, keep a physics-consistency step, and introduce a **learnable data-consistency (LDC)** module that complements the physics-consistency step to suppress residual artifacts. A key difference is that our conditioning instead arises from a regularizer-driven decomposition and directly parameterizes attention via LRG.

Structure-guided and cross-modal reconstruction (MRI). Guided reconstruction leverages side information to align the target with structures visible in a guide modality; in MRI, T1 can guide T2 reconstruction under multi-contrast or cross-modal priors (Ehrhardt & Betcke, 2016b; Feng et al., 2021; Yang et al., 2022; Zhou et al., 2024). This line builds upon classical physics-consistent MRI, including parallel imaging and compressed sensing(Pruessmann et al., 1999; Lustig et al., 2007), and deep cascades that interleave learned priors with data consistency(Schlemper et al., 2018). We reinterpret guidance as prompting within a white-box formulation: the prompt enters the variational gradient as a structure-aligned term that controls attention, linking classical guided reconstruction with modern prompt-based Transformers.

Summary. Compared to black-box prompt injection, our **White-Box Prompt Transformer (WBPT)** provides a variationally grounded route to structure-aware attention (via LRG) while maintaining faithful reconstruction through an LDC-augmented fidelity step, unifying multi-task restoration and guided MRI within the same unfolded architecture.

B Ablation Experiment

B.1 SPSA Component Ablations

We ablate the **SPSA** operator defined in Eq. 4 by removing its last two terms: A $(-\gamma\mu\,\boldsymbol{\xi}_i \cdot \mathrm{softmax}((\mathbf{W}_i\mathbf{x})^\top\boldsymbol{\xi}_i))$ and B $(+\mu\,\boldsymbol{\xi}_i)$. All other implementation details, training protocol, and hyperparameters (including γ, μ) follow the main setup. On BSD68 at $\sigma = 50$, removing either component degrades performance, while the full model (Eq. 4) attains the best average results, indicating that the two terms play complementary roles.

B.2 Position of prompt blocks.

In the hierarchical decoder, we ablate where to inject the prompt blocks. Table 8 compares placing prompts at blocks 1&2, at block 6, and at all decoder blocks on the denoising task with $\sigma = 15$. While placing prompts at all blocks yields only marginal gains over block 6 (up to $0.04\,\mathrm{dB}$ in PSNR and 0.001 in SSIM), it introduces nontrivial computational and latency overhead. We therefore adopt the single-block design at block 6 as the default, which closely matches the all-block variant while reducing wall-clock time and memory footprint.

B.3 Complementarity between Prompt and Data-Consistency Modules.

To avoid attributing the overall improvement to a single component, we conduct a systematic ablation on the deraining task, comparing four configurations: removing the Prompt (w/o Prompt), removing the data-consistency module (w/o DC), removing both components as the cascaded baseline (w/o Prompt & DC), and enabling both components (Prompt+DC). To ensure fairness, all other settings—the number of unrolled steps K, step size η, training schedule, and parameter budget—are kept identical across variants. As summarized in Table 9, relative to the baseline without both modules (w/o Prompt DC), introducing either module alone yields consistent gains in PSNR/SSIM; enabling both simultaneously (Prompt+DC) produces the largest improvement, confirming the strong complementarity between the structural prior provided by the Prompt and the observation-consistency constraint enforced by the DC module.

Table 7: Ablation study of **SPSA** components in Eq. 4 on BSD68 with $\sigma = 50$. The complete formulation (Eq. 4) achieves the best overall performance. Here, A corresponds to $-\gamma\mu\,\boldsymbol{\xi}_i \cdot \mathrm{softmax}((\mathbf{W}_i\mathbf{x})^\top \boldsymbol{\xi}_i)$, and B corresponds to $+\mu\,\boldsymbol{\xi}_i$.

Variant	PSNR	SSIM
w/o A	27.57	0.753
w/o B	27.97	0.798
Full	28.33	0.967

Table 8: Ablation of prompt-injection positions on BSD68 at $\sigma = 15$. Injecting at block 6 matches all-block injection while substantially reducing computation. Results are shown for blocks 1&2, block 6, and all blocks.

Placement	PSNR	SSIM
block 6	34.02	0.963
blocks 1&2	33.97	0.962
all blocks	34.06	0.964

Table 9: Ablation study of **Prompt** and **Data-Consistency (DC)** components within the cascaded Transformer unrolled from Eq. 8, evaluated on the Rain100L dataset.

Variant	PSNR	SSIM
w/o DC&prompt	36.77	0.977
w/o DC	37.07	0.978
w/o prompt	37.46	0.979
Prompt&DC	38.70	0.983

C Variational Derivation of the Energy Functional

For notational simplicity, the subscript i is omitted throughout this section. To develop the optimization algorithm for the proposed model, we consider the variational derivative of the energy functional $\mathcal{R}(\mathbf{x})$, defined as:

$$\begin{aligned}\mathcal{R}(\mathbf{x}) = \frac{1}{2}\int_\Omega \mathrm{Tr}\left(\ln\left(\mathbf{I} + (\mathbf{W}\mathbf{x}(s))(\mathbf{W}\mathbf{x}(s))^\top\right)\right) ds \\ - \mu \int_\Omega \mathrm{Tr}\left(\ln\left(\mathbf{I} + (\mathbf{W}\mathbf{x}(s))(\boldsymbol{\xi}(s))^\top\right)\right) ds\end{aligned} \tag{9}$$

where $\mathbf{x}(s)$ is the optimization variable, $\boldsymbol{\xi}(s)$ denotes the structural prompt, $\mathbf{W}$ is a linear transformation operator, and μ is a regularization weight.

Let us define $\mathbf{y}(s) = \mathbf{W}\mathbf{x}(s)$ and $\tilde{\mathbf{y}}(s) = \boldsymbol{\xi}(s)$.

C.1 Variation of the First Term

Consider the first term:

$$\mathcal{R}_1(\mathbf{x}) = \frac{1}{2}\int_\Omega \mathrm{Tr}\left(\ln\left(\mathbf{I} + \mathbf{y}(s)\mathbf{y}(s)^\top\right)\right) ds \tag{10}$$

Using the matrix differential identity:

$$d\,\mathrm{Tr}\left(\ln\left(\mathbf{I} + \mathbf{y}\mathbf{y}^\top\right)\right) = \mathrm{Tr}\left(\left(\mathbf{I} + \mathbf{y}\mathbf{y}^\top\right)^{-1} d(\mathbf{y}\mathbf{y}^\top)\right) \tag{11}$$

Since $d(\mathbf{y}\mathbf{y}^\top) = d\mathbf{y}\cdot\mathbf{y}^\top + \mathbf{y}\cdot d\mathbf{y}^\top$, we have:

$$\begin{aligned} d\,\mathrm{Tr}\left(\ln\left(\mathbf{I} + \mathbf{y}\mathbf{y}^\top\right)\right) &= \mathrm{Tr}\left(\left(\mathbf{I} + \mathbf{y}\mathbf{y}^\top\right)^{-1}\left(d\mathbf{y}\cdot\mathbf{y}^\top + \mathbf{y}\cdot d\mathbf{y}^\top\right)\right) \\ &= 2\mathrm{Tr}\left(\left(\mathbf{I} + \mathbf{y}\mathbf{y}^\top\right)^{-1}\mathbf{y}\,d\mathbf{y}^\top\right) \\ &= 2\left(\left(\mathbf{I} + \mathbf{y}\mathbf{y}^\top\right)^{-1}\mathbf{y}\right)^\top d\mathbf{y}\end{aligned} \tag{12}$$

where the last equality follows from the identity $\mathrm{Tr}(A^\top) = \mathrm{Tr}(A)$ and $\mathrm{Tr}(A^\top B) = \langle A, B\rangle_F$.

Thus, the variation is:

$$d\,\mathrm{Tr}\left(\ln\left(\mathbf{I}+\mathbf{y}\mathbf{y}^\top\right)\right)=\left\langle 2\left(\mathbf{I}+\mathbf{y}\mathbf{y}^\top\right)^{-1}\mathbf{y},d\mathbf{y}\right\rangle_F \tag{13}$$

Substituting $\mathbf{y}(s)=\mathbf{W}\mathbf{x}(s)$ and $d\mathbf{y}(s)=\mathbf{W}d\mathbf{x}(s)$, we obtain:

$$\begin{aligned} d\mathcal{R}_1 &= \frac{1}{2}\int_\Omega \left\langle 2\left(\mathbf{I}+\mathbf{y}(s)\mathbf{y}(s)^\top\right)^{-1}\mathbf{y}(s),\mathbf{W}d\mathbf{x}(s)\right\rangle_F ds \\ &= \int_\Omega \left\langle \left(\mathbf{I}+\mathbf{y}(s)\mathbf{y}(s)^\top\right)^{-1}\mathbf{y}(s),\mathbf{W}d\mathbf{x}(s)\right\rangle_F ds \end{aligned} \tag{14}$$

Using the definition of the adjoint operator $\mathbf{W}^*$:

$$d\mathcal{R}_1=\int_\Omega \left\langle \mathbf{W}^*\left(\left(\mathbf{I}+\mathbf{y}(s)\mathbf{y}(s)^\top\right)^{-1}\mathbf{y}(s)\right),d\mathbf{x}(s)\right\rangle_F ds \tag{15}$$

Therefore, the variational derivative is:

$$\frac{\delta\mathcal{R}_1}{\delta\mathbf{x}}(s)=\mathbf{W}^*\left(\left(\mathbf{I}+\mathbf{y}(s)\mathbf{y}(s)^\top\right)^{-1}\mathbf{y}(s)\right) \tag{16}$$

C.2 Variation of the Second Term

Now consider the second term:

$$\mathcal{R}_2(\mathbf{x})=\int_\Omega \mathrm{Tr}\left(\ln\left(\mathbf{I}+\mathbf{y}(s)\tilde{\mathbf{y}}(s)^\top\right)\right)ds \tag{17}$$

Using the matrix differential identity:

$$d\,\mathrm{Tr}\left(\ln\left(\mathbf{I}+\mathbf{y}\tilde{\mathbf{y}}^\top\right)\right)=\mathrm{Tr}\left(\left(\mathbf{I}+\mathbf{y}\tilde{\mathbf{y}}^\top\right)^{-1}d(\mathbf{y}\tilde{\mathbf{y}}^\top)\right) \tag{18}$$

$$=\mathrm{Tr}\left(\left(\mathbf{I}+\mathbf{y}\tilde{\mathbf{y}}^\top\right)^{-1}d\mathbf{y}\tilde{\mathbf{y}}^\top\right) \tag{19}$$

where the second equality holds because $\tilde{\mathbf{y}}$ is independent of $\mathbf{x}$.

Using the cyclic property of the trace:

$$\mathrm{Tr}\left(\left(\mathbf{I}+\mathbf{y}\tilde{\mathbf{y}}^\top\right)^{-1}d\mathbf{y}\tilde{\mathbf{y}}^\top\right)=\mathrm{Tr}\left(\tilde{\mathbf{y}}^\top\left(\mathbf{I}+\mathbf{y}\tilde{\mathbf{y}}^\top\right)^{-1}d\mathbf{y}\right) \tag{20}$$

This can be written as an inner product:

$$d\,\mathrm{Tr}\left(\left(\mathbf{I}+\mathbf{y}\tilde{\mathbf{y}}^\top\right)^{-1}d\mathbf{y}\tilde{\mathbf{y}}^\top\right)=\left\langle\left(\mathbf{I}+\mathbf{y}\tilde{\mathbf{y}}^\top\right)^{-\top}\tilde{\mathbf{y}},d\mathbf{y}\right\rangle_F \tag{21}$$

Substituting $\mathbf{y}(s)=\mathbf{W}\mathbf{x}(s)$ and $d\mathbf{y}(s)=\mathbf{W}d\mathbf{x}(s)$, we obtain:

$$d\mathcal{R}_2=\int_\Omega\left\langle\left(\mathbf{I}+\mathbf{y}(s)\tilde{\mathbf{y}}(s)^\top\right)^{-\top}\tilde{\mathbf{y}}(s),\mathbf{W}d\mathbf{x}(s)\right\rangle_F ds \tag{22}$$

Using the adjoint operator $\mathbf{W}^*$:

$$d\mathcal{R}_2=\int_\Omega\left\langle\mathbf{W}^*\left(\left(\mathbf{I}+\mathbf{y}(s)\tilde{\mathbf{y}}(s)^\top\right)^{-\top}\tilde{\mathbf{y}}(s)\right),d\mathbf{x}(s)\right\rangle_F ds \tag{23}$$

Therefore, the variational derivative is:

$$\frac{\delta\mathcal{R}_2}{\delta\mathbf{x}}(s)=\mathbf{W}^*\left(\left(\mathbf{I}+\mathbf{y}(s)\tilde{\mathbf{y}}(s)^\top\right)^{-\top}\tilde{\mathbf{y}}(s)\right) \tag{24}$$

C.3 Combining Both Terms

Combining both components with their respective coefficients and regularization weight μ, we derive the complete variational derivative:

$$\begin{aligned}\frac{\delta \mathcal{R}}{\delta \mathbf{x}}(s) &= \frac{\delta \mathcal{R}_1}{\delta \mathbf{x}}(s) - \mu \frac{\delta \mathcal{R}_2}{\delta \mathbf{x}}(s) \\ &= \mathbf{W}^* \left(\left(\mathbf{I} + \mathbf{y}(s)\mathbf{y}(s)^\top \right)^{-1} \mathbf{y}(s) \right) - \mu \mathbf{W}^* \left(\left(\mathbf{I} + \mathbf{y}(s)\tilde{\mathbf{y}}(s)^\top \right)^{-\top} \tilde{\mathbf{y}}(s) \right)\end{aligned} \tag{25}$$

This gradient form supports the structure-aware optimization in the main algorithm and highlights the explicit interaction between the target variable $\mathbf{x}$ and the structural prompt $\boldsymbol{\xi}$ within the proposed framework.

C.4 Variational Derivative Approximation

Consider the variational derivative of the energy functional $\mathcal{R}[\mathbf{x}]$ with respect to $\mathbf{x}(s)$:

$$\frac{\delta \mathcal{R}}{\delta \mathbf{x}}(s) = \mathbf{W}^* \Big((\mathbf{I} + \mathbf{y}(s)\mathbf{y}(s)^\top)^{-1} \mathbf{y}(s) \Big) - \mu \mathbf{W}^* \Big((\mathbf{I} + \mathbf{y}(s)\tilde{\mathbf{y}}(s)^\top)^{-\top} \tilde{\mathbf{y}}(s) \Big), \tag{26}$$

where

$$\mathbf{y}(s) = \mathbf{W}\mathbf{x}(s), \quad \tilde{\mathbf{y}}(s) = \boldsymbol{\xi}(s). \tag{27}$$

Using the Neumann series expansion,

$$(\mathbf{I} + A)^{-1} = \mathbf{I} - A + A^2 - \cdots \approx \mathbf{I} - A, \tag{28}$$

which holds for matrices with small spectral norm. Therefore,

$$(\mathbf{I} + \mathbf{y}\mathbf{y}^\top)^{-1} \approx \mathbf{I} - \mathbf{y}\mathbf{y}^\top, \quad (\mathbf{I} + \mathbf{y}\tilde{\mathbf{y}}^\top)^{-T} \approx \mathbf{I} - \tilde{\mathbf{y}}\mathbf{y}^\top. \tag{29}$$

Substituting the above approximation into the variational derivative:

$$\frac{\delta \mathcal{R}}{\delta \mathbf{x}} \approx \mathbf{W}^* \big((\mathbf{I} - \mathbf{y}\mathbf{y}^\top)\mathbf{y} \big) - \mu \mathbf{W}^* \big((\mathbf{I} - \tilde{\mathbf{y}}\mathbf{y}^\top)\tilde{\mathbf{y}} \big). \tag{30}$$

Expanding the matrix products gives

$$\frac{\delta \mathcal{R}}{\delta \mathbf{x}} \approx \mathbf{W}^* (\mathbf{y} - \mathbf{y}(\mathbf{y}^\top \mathbf{y})) - \mu \mathbf{W}^* (\tilde{\mathbf{y}} - \tilde{\mathbf{y}}(\mathbf{y}^\top \tilde{\mathbf{y}})). \tag{31}$$

This derivation is rigorous and relies solely on the first-order Neumann expansion and standard matrix algebra.

To obtain a more intuitive subspace membership interpretation, one can heuristically replace the inner product terms $\mathbf{y}^\top \mathbf{y}$ and $\mathbf{y}^\top \tilde{\mathbf{y}}$ with a softmax-normalized form:

$$\frac{\delta \mathcal{R}}{\delta \mathbf{x}} \approx \mathbf{W}^* \Big(\mathbf{y} - \gamma \mathbf{y} \cdot \mathrm{softmax}(\mathbf{y}^\top \mathbf{y}) \Big) - \mu \mathbf{W}^* \Big(\tilde{\mathbf{y}} - \gamma \tilde{\mathbf{y}} \cdot \mathrm{softmax}(\mathbf{y}^\top \tilde{\mathbf{y}}) \Big), \tag{32}$$

where γ is a scaling factor. This softmax replacement is heuristic and provides an interpretation of the inner product terms as subspace membership weights, rather than being a mathematically rigorous derivation.

D Limitation

Despite achieving competitive results across restoration tasks and affording transparent prompt–attention dynamics, one aspect merits further investigation: *computational efficiency*. Specifically, unrolling the gradient flow and stacking Transformer blocks increase the backpropagation memory footprint and wall-clock training time; at inference, the near-quadratic complexity of attention with respect to sequence length (or spatial resolution) can exacerbate latency. We view this as an engineering trade-off rather than a fundamental limitation, and it does not compromise our core conclusion of a variationally anchored, interpretable prompt–attention coupling; nevertheless, further optimization is warranted for resource- and latency-constrained deployments.

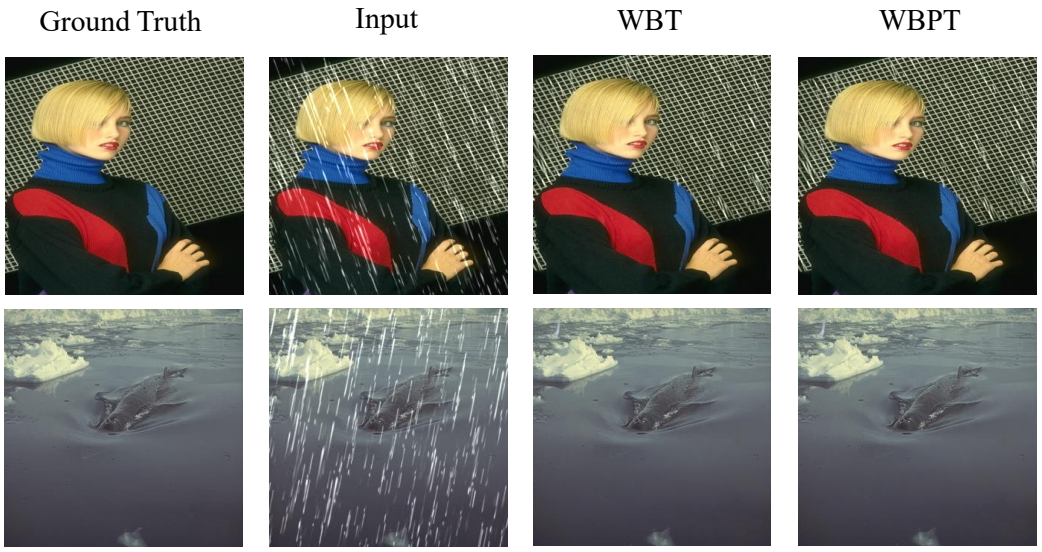

Figure 8: Deraining results for all-in-one methods.

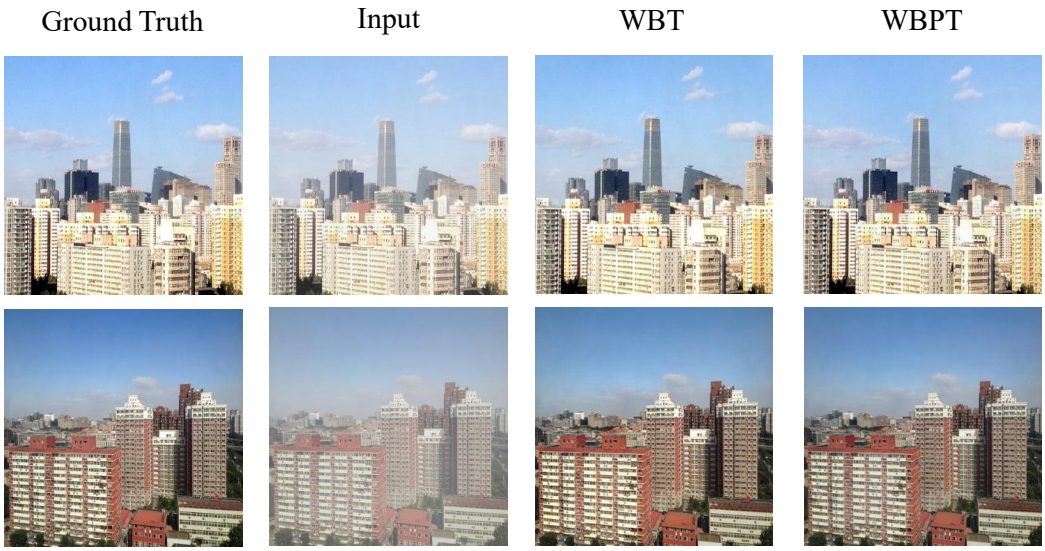

Figure 9: Dehazing results for all-in-one methods.

E QUALITATIVE RESULTS

We present additional qualitative results under the single-task setting to further demonstrate the effectiveness of *prompt-block*. The presented examples correspond one-to-one with the three quantitative single-task tables in the main text (Tables 2, 3, 4), serving to complement and visually illustrate the trends reported in the main paper.

F ALGORITHM

This part provides PyTorch-style pseudocode for the WBPT. Alg. 1 outlines the overall training loop with T-stage learnable gradient updates; Alg. 2 details one iteration stage with the SwinIR backbone and prompt injection; Alg. 3 specifies the prompted window attention and feedforward modules. Unless otherwise noted, we use the following defaults in the pseudocode: *epochs* $=$ 120, *stages*

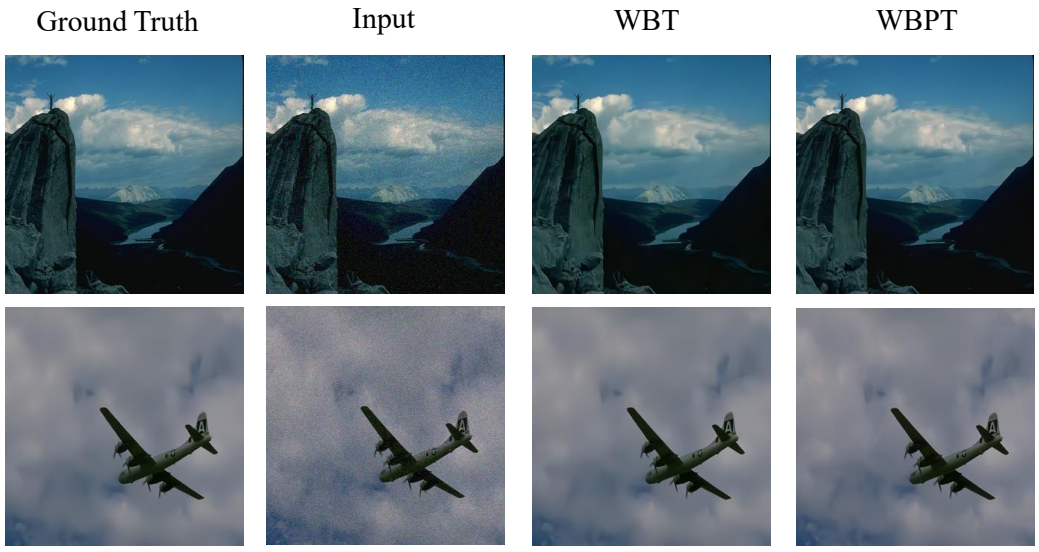

Figure 10: Denoising results for all-in-one methods.

Algorithm 1: PyTorch-style pseudocode for WBPT training (overall)

Input: Dataset $\mathcal{D}$; epochs E (default 120); stages K (default 10); hyperparameters and prompt config P

Output: Trained parameters $\theta^\star$; best validation metrics (PSNR/SSIM)

Optimizer & Loss: Adam(β_1=0.9, β_2=0.999), lr 1×10^{-4}; reconstruction loss $\mathcal{L} = \mathrm{MSE}$

for epoch $\leftarrow 1$ **to** E **do**

  foreach $(\mathbf{x}_{\text{noisy}}, \mathbf{x}_{\text{clean}}) \in \mathrm{DataLoader}(\mathcal{D})$ **do**

    $\mathbf{x}_0 \leftarrow \mathbf{x}_{\text{noisy}}$; $\mathbf{W}_{\text{hist}} \leftarrow [\,]$; $\boldsymbol{\xi}_{\text{hist}} \leftarrow [\,]$

    for $k \leftarrow 1$ **to** K **do**

      $\{\mathbf{W}_i\}_{i=1}^{N}, \{\boldsymbol{\xi}_i\}_{i=1}^{N} \leftarrow \mathrm{SwinIR}(\mathbf{x}_{k-1};\, P,\, k)$ `# with optional prompt injection`

      Append $\{\mathbf{W}_i\}_{i=1}^{N}$ to $\mathbf{W}_{\text{hist}}$ and $\{\boldsymbol{\xi}_i\}_{i=1}^{N}$ to $\boldsymbol{\xi}_{\text{hist}}$; drop oldest if length > 5

      $\mathbf{x}_k \leftarrow \mathrm{MPSA}\big(\mathbf{x}_{k-1},\, \mathbf{y}{=}\mathbf{x}_{\text{clean}},\, \{\mathbf{W}_i\},\, \{\boldsymbol{\xi}_i\},\, k\big)$

    $\mathcal{L} \leftarrow \mathrm{MSE}(\mathbf{x}_K, \mathbf{x}_{\text{clean}})$

    optimizer.zero_grad(); $\mathcal{L}$.backward(); optimizer.step()

  ValidateAndSaveIfBest(θ) `# compute PSNR/SSIM on val, save best`

Complexity: $\mathcal{O}\big(E \cdot N \cdot K \cdot (F_{\text{swin}} + F_{\text{mpsa}})\big)$ `# Here` N `denotes the number of batches per epoch.`

$T = 10$, `embed_dim` $= 96$, `num_heads` $= [6, 6, 6]$, `window` $= 8$, `prompt_len` $= 5$, Adam with (β_1=0.9, β_2=0.999), learning rate 1×10^{-4}, MSE reconstruction loss, and 128×128 random crops with rotation/flip augmentations. The pseudocode focuses on core computations; engineering aspects (I/O, multi-GPU, logging, checkpointing) are omitted for brevity. Complexity expressions report the dominant terms (attention and MLP). Notation: $x_{\text{noisy}}/x_{\text{clean}}$ denote inputs/targets, z the current state, U_k the backbone output/prompt at stage k, and t the iteration index. Prompt injection is fixed at block indices $\{2, 4, 6\}$ within the SwinIR backbone.

G Transformer and Prompt Blocks in WBPT

As stated in Section 2.2 of the main manuscript, we present in Fig. 11 the block diagram of the Prompt block corresponding to ξ_i, and further elaborate on the implementation details of the Transformer block used within this Prompt block in Fig. 12. The Prompt block and the Transformer block

Algorithm 2: One WBPT iteration stage (SwinIR with prompt injection)

Input: Current state $\mathbf{x}_{k-1}$; prompt config P; SwinIR depths $[2, 2, 2]$, channels $C{=}96$, window $M{=}8$
Output: $\{\mathbf{W}_i\}_{i=1}^N$, $\{\boldsymbol{\xi}_i\}_{i=1}^N$ (transformations and prompts for MPSA)

$\mathbf{f} \leftarrow \mathrm{Conv3x3}(\mathbf{x}_{k-1})$ `# shallow feature`
for $\ell \leftarrow 1$ **to** 3 **do**
  $(\mathbf{f}, \mathrm{size}) \leftarrow \mathrm{PatchEmbed}(\mathbf{f})$
  for $b \leftarrow 1$ **to** 2 **do**
    if $\mathrm{ShouldUsePrompt}(\ell, b, P)$ **then**
      $\{\boldsymbol{\xi}_i\}_{i=1}^N \leftarrow \mathrm{PromptGenBlock}(\mathbf{f})$; $\mathbf{f} \leftarrow \mathbf{f} + \mathrm{Inject}(\{\boldsymbol{\xi}_i\})$ `# e.g., at fixed block indices {2,4,6}`
    $\mathbf{f} \leftarrow \mathrm{SwinTransformerBlock}(\mathbf{f}, \mathrm{size})$
  if $\ell < 3$ **then**
    $\mathbf{f} \leftarrow \mathrm{PatchMerging}(\mathbf{f}, \mathrm{size})$
$\{\mathbf{W}_i\}_{i=1}^N$, $\{\boldsymbol{\xi}_i\}_{i=1}^N \leftarrow \mathrm{ExtractTransformationsAndPrompts}(\mathbf{f})$
return $\{\mathbf{W}_i\}_{i=1}^N$, $\{\boldsymbol{\xi}_i\}_{i=1}^N$
Complexity: $\sum_{\ell=1}^{3}\sum_{b=1}^{2}\big(F_{\text{W-MSA}} + F_{\text{MLP}}\big) \approx \mathcal{O}(n_W \cdot B \cdot M^2 C + LC^2)$

Algorithm 3: Multi-Prompted Structure Attention (MPSA) with Learnable Data Consistency

Input: $\mathbf{x}_k \in \mathbb{R}^{B\times H\times W\times C}$; transforms $\{\mathbf{W}_i\}_{i=1}^N$; prompts $\{\boldsymbol{\xi}_i\}_{i=1}^N$; measurements/targets $\mathbf{y}$
Output: $\mathbf{x}_{k+1} \in \mathbb{R}^{B\times H\times W\times C}$

Multi-Prompted Structure Attention:
  for $i \leftarrow 1$ **to** N **do**
    $\mathrm{spsa}_i \leftarrow \mathrm{SPSA}(\mathbf{x}_k \mid \mathbf{W}_i, \boldsymbol{\xi}_i, \gamma, \mu)$ `# Eq. 4`
  $\mathrm{mpsa\_out} \leftarrow [\mathbf{W}_1^* \ \cdots \ \mathbf{W}_N^*] \begin{bmatrix} \mathrm{spsa}_1 \\ \vdots \\ \mathrm{spsa}_N \end{bmatrix}$ `# Eq. 5`
Learnable Data Consistency:
  $\Delta_\phi(\mathbf{x}_k, \mathbf{y}) \leftarrow \mathrm{LearntGradient}(\mathbf{x}_k, \mathbf{y})$
Gradient Flow Update:
  $\mathbf{x}_{k+1} \leftarrow \big(\boldsymbol{I} - \eta \sum_{i=1}^N \mathbf{W}_i^*\mathbf{W}_i\big)\,\mathbf{x}_k \ - \ \eta\,\mathrm{mpsa\_out} \ - \ \eta\,\Delta_\phi(\mathbf{x}_k, \mathbf{y})$ `# Eq. 8`
return $\mathbf{x}_{k+1}$
Complexity: MPSA $\mathcal{O}(N \cdot HWC^2)$; data consistency $\mathcal{O}(HWC^2)$

follow the design and hyper-parameter settings outlined in Potlapalli et al. (2023) and Zamir et al. (2022), respectively.

G.1 Prompt Block in WBPT Framework

G.1.1 Prompt Block Overview

Given prompt components $\mathbf{P_c} \in \mathbb{R}^{N\times\hat{H}\times\hat{W}\times\hat{C}}$ and input features $\mathbf{F_1} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$, the prompt block refines the input via:

$$\hat{\mathbf{F}}_\mathbf{1} = \mathtt{PIM}(\mathtt{PGM}(\mathbf{P_c}, \mathbf{F_1}), \mathbf{F_1}) \tag{33}$$

The block contains two modules: the Prompt Generation Module (PGM) and the Prompt Interaction Module (PIM), detailed below.

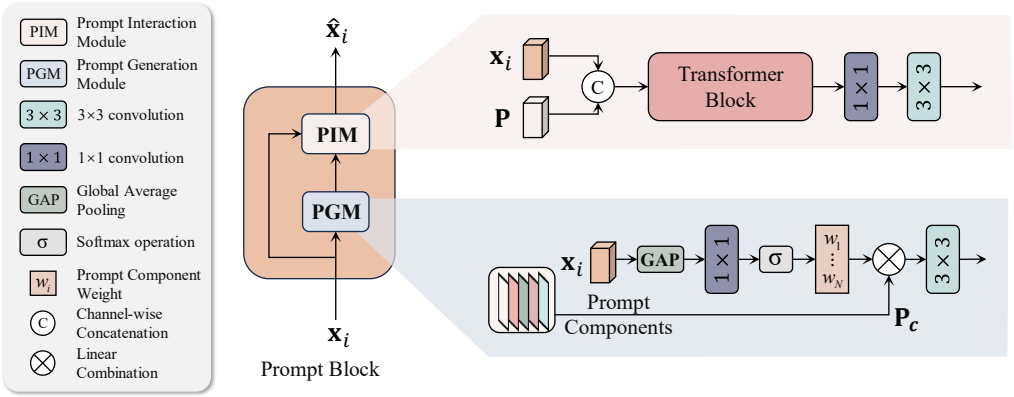

Figure 11: Overview of the Prompt block used in the WBPT framework. The Prompt block is composed of two sub modules,the Prompt Generation Module (PGM) and the Prompt Interaction Module (PIM).

G.1.2 Prompt Generation Module (PGM)

PGM dynamically generates input-conditioned prompts. First, global average pooling (GAP) is applied on $\mathbf{F_1}$ to produce a channel-wise descriptor $\mathbf{v} \in \mathbb{R}^{\hat{C}}$. A 1×1 convolution and softmax yield prompt weights $w \in \mathbb{R}^N$:

$$w_i = \texttt{Softmax}(\texttt{Conv}_{1\times1}(\texttt{GAP}(\mathbf{F_1}))) \tag{34}$$

These weights modulate the learned prompt components to form the final prompt $\mathbf{P}$:

$$\mathbf{P} = \texttt{Conv}_{3\times3}\left(\sum_{c=1}^{N} w_i \mathbf{P}_c\right) \tag{35}$$

To support variable-resolution inputs, prompt components are upsampled to match the spatial size of $\mathbf{F_1}$ via bilinear interpolation.

G.1.3 Prompt Interaction Module (PIM)

PIM fuses the generated prompt $\mathbf{P}$ with features $\mathbf{F_1}$ via channel-wise concatenation, followed by a Transformer block:

$$\hat{\mathbf{F}}_\mathbf{1} = \texttt{Conv}_{3\times3}(\texttt{GDFN}(\texttt{MDTA}([\mathbf{F_1}; \mathbf{P}]))) \tag{36}$$

G.2 Transformer Block in WBPT Framework

MDTA Module. Let the input feature map be $\mathbf{X} \in \mathbb{R}^{H_l \times W_l \times C_l}$. MDTA first applies Layer Normalization, then projects the input to query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) tensors using a sequence of 1×1 pointwise convolution followed by 3×3 depth-wise convolution, all bias-free. To enable channel-wise attention, $\mathbf{Q}$ and $\mathbf{K}$ are reshaped to $\mathbb{R}^{H_l W_l \times C_l}$ and $\mathbb{R}^{C_l \times H_l W_l}$ respectively, resulting in a transposed attention map of shape $C_l \times C_l$ via dot-product interaction. Multi-head computation is performed in parallel.

GDFN Module. The GDFN submodule begins with a channel expansion using a 1×1 convolution by a factor γ. The expanded features are split into two parallel branches, each followed by a 3×3 depth-wise convolution. One branch passes through a GeLU activation while the other remains linear. The outputs are combined via element-wise multiplication, and finally projected back to the original channel dimension through a 1×1 convolution. Residual connections are maintained throughout the block.

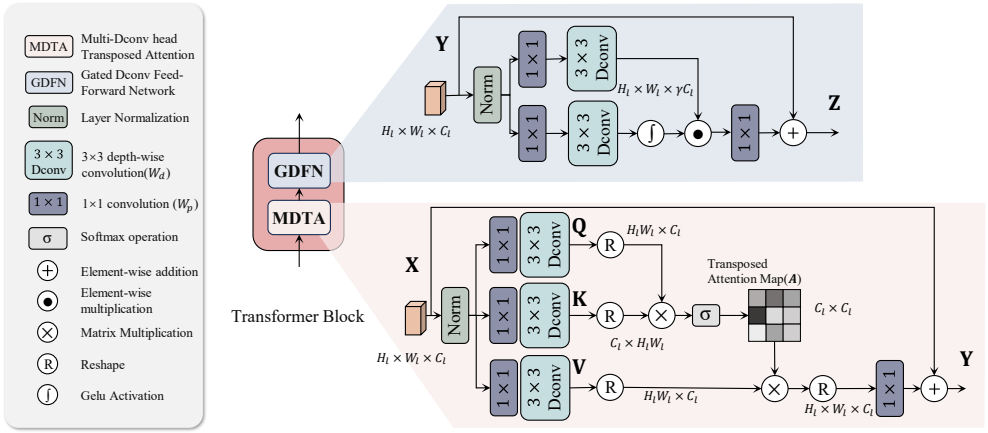

Figure 12: Overview of the Transformer block used in the Prompt block. The Transformer block is composed of two sub modules,the Multi Dconv head transposed attention module(MDTA) and the Gated Dconv feed-forward network(GDFN).

MDTA performs self-attention along channels:

$$\mathbf{Y} = W_p\mathbf{V} \cdot \texttt{Softmax}(\mathbf{K} \cdot \mathbf{Q}/\alpha) + \mathbf{X}$$

GDFN transforms the result as:

$$\mathbf{Z} = W_p^0 \left(\phi(W_d^1 W_p^1(\texttt{LN}(\mathbf{Y}))) \odot W_d^2 W_p^2(\texttt{LN}(\mathbf{Y}))\right) + \mathbf{Y}$$

Here, LN is layer normalization, ϕ denotes GELU activation, and $\odot$ is element-wise multiplication.

H Reproducibility Statement

We provide all necessary details to support reproducibility. All experiments are conducted on publicly available datasets, and the model architectures, hyperparameters, training protocols, and evaluation metrics are specified in the paper. We will release our codebase, training scripts, and pretrained checkpoints on GitHub upon acceptance.

I The Use of Large Language Models (LLMs)

We used large language models only for light editorial assistance during manuscript preparation (grammar and wording refinement, minor style/formatting suggestions). No LLMs were used for research ideation, dataset curation, modeling, experiment design, analysis, or drafting substantive sections.