# HYBRID GENERALIZATION-AWARE PARAMETER SELECTION TEST-TIME ADAPTATION FOR ALL-IN-ONE IMAGE RESTORATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

All-in-one image restoration (AiOIR) methods achieve impressive performance across diverse degradations by learning a unified restoration model. However, they remain vulnerable to test-time distribution shifts. Existing test-time adaptation (TTA) methods stabilize online parameter updates via Fisher-based selective adaptation, yet this selection criterion is inherently one-sided: it only identifies which parameters to preserve from updates, but fails to distinguish which of the remaining parameters merit adaptation to the target degradation. To address this limitation, we propose ***H**ybrid **G**eneralization-aware **P**arameter **S**election* (HGPS), a budget-neutral, per-tensor selection principle. Specially designed to complement Fisher stability, HGPS employs cross-view gradient agreement as a proxy for target-relevant update utility. Within each tensor's original update budget, HGPS hard-protects high-Fisher elements, assigns part of the update slots to parameters with consistent cross-view gradients, and fills the rest with low-Fisher elements. Furthermore, the mask is computed once before adaptation and can be plugged into Fisher-based TTA pipelines without altering their optimization or restoration schedule. Extensive experiments across multiple restoration tasks and distribution shifts demonstrate that HGPS consistently improves state-of-the-art restoration TTA, surpassing the strongest existing TTA method by 2.69 dB PSNR on Kodak24.

## 1 INTRODUCTION

All-in-one image restoration (AiOIR) seeks a single model that can handle multiple degradations such as rain, haze, noise, blur, and low light. Prompt-based and unified restoration networks have made this setting increasingly practical, but their behavior can deteriorate when test degradations depart from the training distribution. Test-time adaptation (TTA) offers a source-free mechanism for responding to such shifts by updating a pretrained model on unlabeled test data. Yet restoration TTA is unusually fragile: unlike classification, it optimizes a dense regression model whose online updates can introduce artifacts or erase pretrained restoration knowledge.

To mitigate this fragility, state-of-the-art restoration TTA couples degradation-consistent pseudo-supervision with Fisher-based selective adaptation: high-importance parameters are protected and restored after each optimizer step, as instantiated by the TIPS module in DCTTA. Such protection stabilizes adaptation—removing it severely degrades restoration quality—yet the selection criterion itself has received little scrutiny.

As illustrated in Figure 1(a), Fisher importance answers only one side of the selection problem: it tells us which parameters should not be changed. A low-Fisher element is comparatively safe to update, but safety is not utility—such an element may have near-zero influence on target-domain adaptation, or its update direction may be inconsistent across different valid views of the same degraded input. Consequently, the update budget of Fisher-only selection is partially spent on elements whose updates bring little generalization value. This distinction echoes evidence from the broader TTA literature: EATA constrains important weights with Fisher information (Niu et al., 2022), SAR rejects samples that induce unstable gradients (Niu et al., 2023), and GraTa explicitly exploits gradient alignment to improve adaptation directions (Chen et al., 2025). Together, these observations call



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.

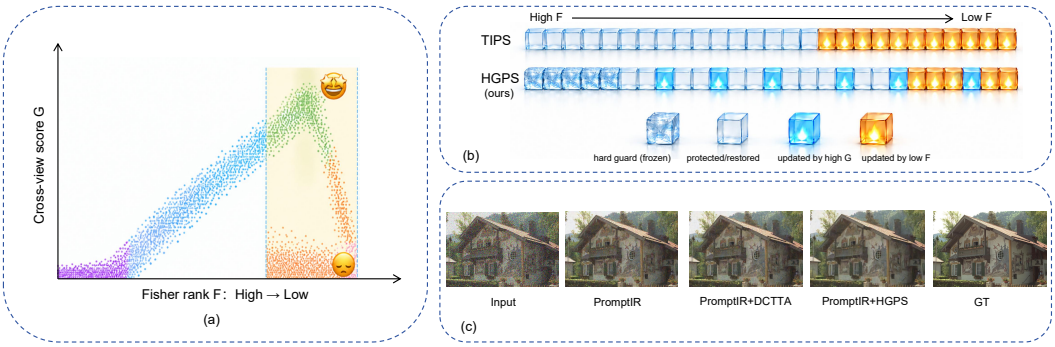


Figure 1: Overview of the proposed hybrid generalization-aware parameter selection (HGPS). **(a)** Elements within a parameter tensor, ranked by their Fisher score $F$ (High $\rightarrow$ Low) and colored by the cross-view update-utility score $G$. The low-$F$ region filled by Fisher-only selection is heterogeneous: it mixes elements with consistent and non-vanishing cross-view gradients (green), whose updates generalize to the target degradation, with elements showing inconsistent or near-zero gradients (orange), whose updates contribute little. **(b)** Per-tensor update-slot assignment of TIPS and HGPS under an identical budget $K_t$ (Eq. (2)). TIPS updates exactly the $K_t$ lowest-$F$ elements and restores all the rest. HGPS freezes the $H_t$ highest-$F$ elements as a hard guard, then assigns $\mathrm{round}(\alpha K_t)$ update slots to the non-guarded elements with the largest $G$, and fills the remaining slots with the lowest-$F$ elements; the per-tensor budget, ranking granularity, and post-step restoration are unchanged. **(c)** Visual comparison on a degraded test image: updating parameters selected by cross-view agreement (PromptIR + HGPS) recovers details closer to the GT than updating the lowest-Fisher elements alone (PromptIR + DCTTA).

for a complementary measure of update utility that operates *within*, rather than instead of, a strong anti-forgetting boundary.

Motivated by this, we propose HGPS, a per-tensor, budget-preserving hybrid selection scheme, as shown in Figure 1(b). For each tensor $t$, HGPS retains the host scheme's original update budget $K_t = n_t - \lfloor \rho n_t \rfloor$, where $\rho$ is the Fisher protection ratio, and changes only which elements within the tensor are updated. HGPS combines two element-wise scores: Fisher importance $F$, which determines what to protect, and a new cross-view score $G$, which measures whether two groups of geometric views induce aligned and non-vanishing student gradients under EMA-teacher pseudo-supervision. A Fisher hard guard first excludes the most critical elements from the $G$ branch; the $K_t$ update slots are then split between high-$G$ and low-$F$ elements via a single ratio $\alpha$, the only new hyperparameter HGPS introduces. As shown in Figure 1(c), this generalization-aware re-selection recovers details closer to the ground truth than updating the lowest-Fisher elements alone, while leaving the host pipeline's optimization and restoration schedule untouched.

Our main contributions are summarized as follows:

- We identify a one-sidedness in Fisher-based parameter selection for restoration TTA: importance scores determine which parameters to protect, but not which of the remaining ones are worth updating. We show that parameter stability and target-domain update utility are complementary rather than interchangeable, motivating a fixed-budget hybrid criterion instead of a simple change of the update ratio.
- An element-wise cross-view gradient-agreement score is introduced, relying solely on unlabeled target images and EMA-teacher pseudo-supervision. It rewards update directions that are consistent across view groups while suppressing near-vanishing gradients.
- We design a per-tensor allocation rule that combines a Fisher hard guard with a split of update slots between high-agreement and low-Fisher elements. The rule is budget-neutral—it preserves the per-tensor update count and the post-step restoration semantics of the host pipeline—thereby isolating the effect of member selection from capacity reallocation and optimizer dynamics.

# 2 Related Work

## 2.1 Test-Time Adaptation and Stability

Fully test-time adaptation (TTA) updates a source-trained model using only unlabeled target data. Tent (Wang et al., 2021) minimizes prediction entropy on test batches while updating only normalization parameters. Later work shows that unconstrained online updates accumulate errors and forget source knowledge, and stabilizes adaptation at different granularities: EATA (Niu et al., 2022) restricts updates to reliable samples and regularizes Fisher-important weights; CoTTA (Wang et al., 2022) combines augmentation-averaged teacher predictions with stochastic weight restoration; SAR (Niu et al., 2023) removes samples with large, unstable gradients and takes sharpness-aware steps. Practical TTA thus requires both an adaptation signal and an explicit stabilization mechanism.

Agreement across valid views of the same input is another useful test-time signal: MEMO (Zhang et al., 2022) minimizes the marginal entropy over augmented views of a single test sample, and GraTa (Chen et al., 2025) aligns a pseudo gradient with an auxiliary gradient in medical image segmentation, scaling the learning rate by their cosine similarity. HGPS differs from these uses in both role and granularity: it modifies neither the online descent direction nor the learning rate, but computes a parameter-element score *before* adaptation to decide which elements may retain optimizer updates—gradient agreement as a selection score, not as an online loss.

## 2.2 TTA for All-in-One Image Restoration

All-in-one image restoration (AiOIR) seeks a single shared-parameter network that handles diverse degradations. Prompt-based and unified architectures (e.g., PromptIR (Potlapalli et al., 2023) and its successors) achieve strong in-distribution performance but degrade noticeably under unseen or out-of-distribution corruptions, motivating TTA for restoration models.

This setting remains largely unexplored. TAO (Gou et al., 2024) adapts a pretrained, degradation-agnostic diffusion model to unseen degradations with a learnable adapter, but relies on diffusion-based generation at inference rather than adapting an AiOIR network; diffusion-based TTA has also been explored for video adverse-weather removal (Yang et al., 2024). DCTTA (Tang et al., 2026) is, to our knowledge, the only dedicated source-free TTA framework for AiOIR: it builds degradation-consistent pseudo-supervision via test-time redegradation, and its TIPS module estimates a diagonal Fisher importance from an augmentation-consistency objective and restores high-importance elements to their pretrained values after each optimizer step. This protection is load-bearing: removing TIPS drops PSNR/SSIM on PromptIR/Rain100H from 20.21/0.663 to 14.24/0.487—*below* the unadapted model itself (15.64/0.498) (Tang et al., 2026)—indicating that in restoration TTA, *which* parameters may change matters as much as the adaptation signal itself.

HGPS targets exactly this selection step with deliberately minimal scope: it leaves the host pipeline's degradation modeling, adaptation losses, EMA teacher update, optimizer step, and post-step restoration intact, and changes only the membership of the per-tensor update set. The criterion itself is host-agnostic—we instantiate it on DCTTA because it is the current state of the art, but it applies to any Fisher-based selective adaptation scheme. This controlled design isolates the effect of the selection criterion and underpins the causal interpretation of our experiments.

# 3 Method

We propose *hybrid generalization-aware parameter selection* (HGPS), a budget-neutral, per-tensor selection principle for Fisher-based test-time adaptation (TTA). HGPS is host-agnostic: it computes an element-wise update-utility score once, before adaptation, and re-selects only the membership of each tensor's update set, without altering the host pipeline's adaptation objective, update budget, optimizer, or restoration schedule. Figure 2 illustrates how HGPS plugs into the host degradation-guided restoration pipeline. We first formalize the host objective and the Fisher-based selective adaptation scheme that HGPS plugs into (Section 3.1), then define the cross-view update-utility score (Section 3.2), and finally present the hybrid selection rule (Section 3.3).

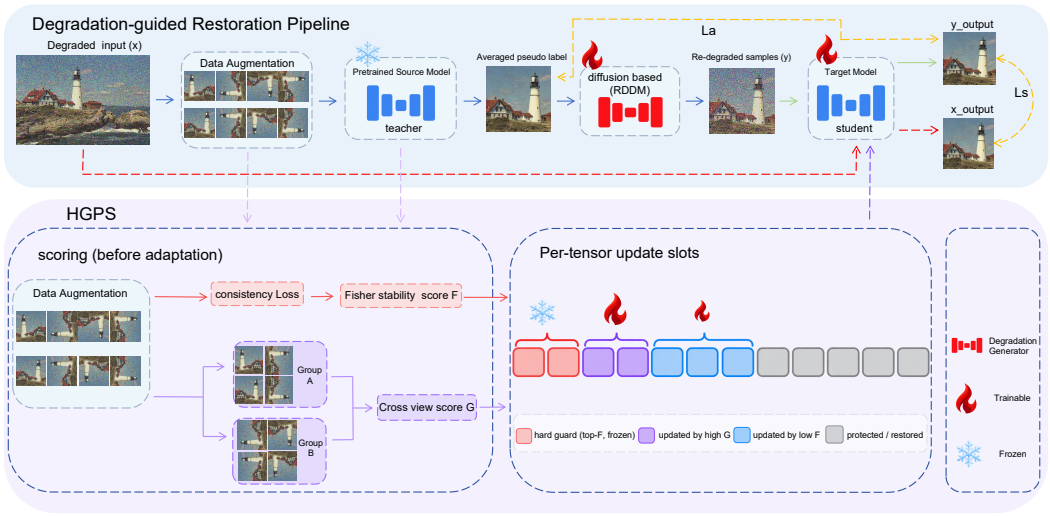


Figure 2: Pipeline of HGPS on top of the host degradation-guided restoration framework (DCTTA). **Adaptation:** an EMA teacher restores the dihedral views of the degraded input $x$, which are averaged in canonical coordinates into a pseudo-label with a per-pixel confidence map (Eq. (A.1)); a diffusion-based re-degradation generator produces a re-degraded sample sharing the target degradation, and the student is optimized with the degradation-consistency objective $\mathcal{L}_{\text{adapt}}$ (Eq. (A.2)). **Selection:** once before adaptation, HGPS accumulates the Fisher score $F$ (Eq. (1)) and the cross-view score $G$ from two complementary view groups (Eqs. (4)–(6)), and then fills the $K_t$ update slots of every tensor with a Fisher hard guard, high-$G$ elements, and low-$F$ elements (Section 3.3); all remaining positions are restored to their pretrained values after every optimizer step (Eq. (3)).

### 3.1 Preliminaries: Degradation-Consistent Adaptation with a Fisher Mask

**Problem setup.** Let $f_{\boldsymbol{\theta}^0}$ be an all-in-one restoration model pretrained on source data, which is adapted at test time on an unlabeled target stream $\{x_s\}$ of degraded images; ground-truth images are never accessed. Following the host pipeline, adaptation is restricted to a candidate set of parameter tensors (prompt, encoder, and decoder modules); we denote by $\theta_{t,i}$ the $i$-th element of the $t$-th candidate tensor, which has $n_t$ elements in total.

**Host adaptation objective.** We instantiate the host pipeline with DCTTA (Tang et al., 2026), whose per-iteration objective couples an EMA teacher $f_{\bar{\boldsymbol{\theta}}}$ with test-time re-degradation. Given the dihedral view group $\mathcal{T} = \{\tau_m\}_{m=1}^{8}$, the teacher's restorations of all views are mapped back to canonical coordinates and, over $R$ independent rounds, averaged into a pseudo-label $\bar{y}(x)$ with a per-pixel confidence map $c(x)$; a lightweight diffusion-based re-degradation generator $D_\psi$ is then updated on the fly with the pair $(\bar{y}(x), x)$ and produces a re-degraded sample $\tilde{x}$ that shares the target degradation of $x$. The student is adapted on $\tilde{x}$ with a degradation-consistency objective $\mathcal{L}_{\text{adapt}}$, which enforces consistency with the student's own restoration of $x$ and distills the teacher pseudo-label with weight $\lambda_{\text{tea}} = 5$; after each optimizer step, the teacher is updated by exponential moving average. All components of the host objective are kept strictly unchanged—the full equations are deferred to Appendix A (Eqs. (A.1)–(A.3))—and HGPS only redefines how the restore mask $M_t$ in Eq. (3) below is constructed.

**Selective adaptation with a Fisher mask.** Fisher-based selective adaptation, instantiated by the TIPS module of DCTTA (Tang et al., 2026), first estimates an importance score for every trainable parameter element $\theta_{t,i}$ over a scoring set of $S$ target samples. The scoring objective is an augmentation-consistency loss that involves neither the teacher nor the re-degradation generator: the student restores all views of $\mathcal{T}$, the results are mapped back to canonical coordinates and averaged into a self-ensemble pseudo-label $\bar{y}_{\text{aug}}(x;\boldsymbol{\theta}) = \frac{1}{|\mathcal{T}|}\sum_m \tau_m^{-1}(f_{\boldsymbol{\theta}}(\tau_m(x)))$, and the differentiable anchor view is supervised with a Charbonnier loss $\ell_{\text{ch}}(a,b) = \sqrt{(a-b)^2 + \epsilon_{\text{ch}}}$ ($\epsilon_{\text{ch}} = 10^{-3}$, averaged

over pixels),

$$\mathcal{L}_{\text{aug}}(x;\boldsymbol{\theta}) \;=\; \ell_{\text{ch}}\big(f_{\boldsymbol{\theta}}(x),\, \text{sg}[\bar{y}_{\text{aug}}(x;\boldsymbol{\theta})]\big), \qquad F_{t,i} \;=\; \sum_{s=1}^{S}\Big(\frac{\partial \mathcal{L}_{\text{aug}}(x_s;\boldsymbol{\theta})}{\partial \theta_{t,i}}\Big)^2. \tag{1}$$

Selection is performed *independently inside each tensor*: for tensor $t$ with $n_t$ elements, the $\lfloor \rho n_t \rfloor$ highest-Fisher elements ($\rho = 0.6$ by default) are protected, and only the remaining

$$K_t \;=\; n_t - \lfloor \rho n_t \rfloor \tag{2}$$

lowest-Fisher elements may retain their updates. Every optimizer step is then followed by a *post-step restoration*: with a binary mask $M_t \in \{0,1\}^{n_t}$ ($1 =$protected),

$$\theta_{t,i} \;\longleftarrow\; M_{t,i}\,\theta^0_{t,i} \;+\; \big(1 - M_{t,i}\big)\,\theta_{t,i}, \tag{3}$$

i.e., protected positions are reset to the pretrained weights $\boldsymbol{\theta}^0$, so adaptation takes place on a per-tensor subspace of size $K_t$.

**Limitation.** The Fisher criterion answers *which parameters to protect*, but fills the $K_t$ update slots only implicitly, by ascending Fisher score. This is not equivalent to selecting *valuable* updates: an element attains a near-zero Fisher score whenever its gradient magnitude is tiny, so part of the budget is spent on updates that contribute almost nothing; and Fisher magnitude says nothing about whether an update direction generalizes beyond the particular views and samples that produced it. HGPS therefore keeps Eqs. (2)–(3) exactly as they are, and changes only the membership of the update set.

### 3.2 Cross-View Update-Utility Score

We seek a complementary score identifying elements whose updates (i) point in a direction *consistent* across independent evidence and (ii) have *non-negligible magnitude*, using only unlabeled target images. The host pipeline already provides the two ingredients we need: a set of geometric views and an EMA teacher. We partition the eight views of $\mathcal{T}$ into two fixed complementary groups of four, $\mathcal{A} = \{0,3,4,7\}$ and $\mathcal{B} = \{1,2,5,6\}$ (indices into the op list $[\varnothing, v, h, t, vh, vt, ht, vht]$). For each group $V \in \{\mathcal{A},\mathcal{B}\}$, the EMA teacher produces a pseudo-label $\hat{y}^V_s = \frac{1}{|V|}\sum_{m\in V} \tau_m^{-1}(f_{\bar{\boldsymbol{\theta}}}(\tau_m(x_s)))$ in canonical coordinates, and the student is supervised by the group consistency loss, built from the same Charbonnier loss $\ell_{\text{ch}}$ as Eq. (1),

$$\mathcal{L}^V_s(\boldsymbol{\theta}) \;=\; \frac{1}{|V|}\sum_{m\in V} \ell_{\text{ch}}\big(\tau_m^{-1}\big(f_{\boldsymbol{\theta}}(\tau_m(x_s))\big),\, \hat{y}^V_s\big), \qquad g^V_s \;=\; \frac{\partial \mathcal{L}^V_s}{\partial \boldsymbol{\theta}}. \tag{4}$$

The two groups thus provide, for every element $\theta_{t,i}$, two independent gradient estimates $g^{\mathcal{A}}_{t,i,s}$ and $g^{\mathcal{B}}_{t,i,s}$ of how it should change to satisfy view consistency. We accumulate the cross-product and per-group squared norms over the same $S$ scoring samples used for $F$ (no ground-truth image is ever read),

$$S^{\mathcal{AB}}_{t,i} = \sum_s g^{\mathcal{A}}_{t,i,s}\, g^{\mathcal{B}}_{t,i,s}, \qquad S^{\mathcal{AA}}_{t,i} = \sum_s \big(g^{\mathcal{A}}_{t,i,s}\big)^2, \qquad S^{\mathcal{BB}}_{t,i} = \sum_s \big(g^{\mathcal{B}}_{t,i,s}\big)^2, \tag{5}$$

and define the cross-view update-utility score as

$$C_{t,i} \;=\; \frac{S^{\mathcal{AB}}_{t,i}}{\sqrt{S^{\mathcal{AA}}_{t,i}\, S^{\mathcal{BB}}_{t,i} + \epsilon}}, \qquad G_{t,i} \;=\; \max\big(C_{t,i},\, 0\big)\cdot\sqrt{\tfrac{1}{2}\big(S^{\mathcal{AA}}_{t,i} + S^{\mathcal{BB}}_{t,i}\big)}, \tag{6}$$

with $\epsilon = 10^{-12}$ for numerical stability. $C_{t,i}$ is a per-element, cosine-like agreement between the update directions suggested by the two view groups: it is positive only when both groups consistently push the element in the same direction, and the clipping $\max(C_{t,i}, 0)$ discards elements on which the groups disagree. The square-root term re-weights the agreement by the typical gradient magnitude, so that elements with consistent-but-negligible gradients do not rank highly. Since $G_{t,i}$ is only ever compared *within* tensor $t$, no cross-layer scale calibration is needed. We stress that $G$ is a *selection score*, not an online loss: it is computed once before adaptation and never enters the optimization objective.

### 3.3 HYBRID GENERALIZATION-AWARE PARAMETER SELECTION

**Per-tensor selection rule.** Given $F$ (Eq. (1)) and $G$ (Eq. (6)), HGPS fills the $K_t$ update slots of every tensor $t$ independently, in three steps. First, the

$$H_t \;=\; \max\big(n_t - 2K_t,\; 0\big) \tag{7}$$

highest-Fisher elements form a hard guard set $\mathcal{H}_t$ ($H_t \approx 0.2\, n_t$ under the default $\rho = 0.6$), preventing the highest-Fisher elements from being selected by the utility branch. Second, among the non-guarded elements, the $K_{G,t} = \mathrm{round}(\alpha K_t)$ elements with the largest $G_{t,i}$ form the update subset $\mathcal{U}_{G,t}$, where $\alpha \in [0, 1]$ ($\alpha = 0.5$ by default) is the only new hyperparameter. Third, the remaining $K_{F,t} = K_t - K_{G,t}$ slots are filled, from the candidates left, by the elements with the *smallest* $F_{t,i}$, forming $\mathcal{U}_{F,t}$ as a stability-driven fallback. The update set $\mathcal{U}_t = \mathcal{U}_{G,t} \cup \mathcal{U}_{F,t}$ has exactly $K_t$ elements by construction, and the restore mask in Eq. (3) is $M_{t,i} = \mathbb{1}[\theta_{t,i} \notin \mathcal{U}_t]$. No ranking, quota transfer, or comparison ever crosses tensor boundaries: the per-tensor budget, ranking granularity, and restoration semantics are identical to the host, so any empirical gain can be attributed to the membership criterion alone. At $\alpha = 0$, $\mathcal{U}_t$ reduces to the $K_t$ lowest-Fisher elements, i.e., exactly the host's Fisher-only selection, which we use as an implementation-level consistency check; at $\alpha = 1$ (denoted HGPS-G) all slots are filled by $G$ but the hard guard is retained, which is distinct from a guard-free Gen-only ablation that selects the top-$K_t$ elements by $G$ alone.

**Adaptation procedure.** Both scores are computed once, before adaptation: $F$ reuses the host's augmentation-consistency scoring objective with squared-gradient accumulation (Eq. (1)), and $G$ is computed on the same samples in the same order (Eqs. (4)–(6)). The scoring passes only accumulate statistics—no parameter, normalization buffer, or EMA state is modified. The masks $\{M_t\}$ are then built per tensor and kept fixed. Each subsequent adaptation iteration is identical to the host pipeline—computing $\mathcal{L}_{\mathrm{adapt}}$ (Eq. (A.2)), backpropagation, and an optimizer step on the same candidate parameter set—followed by the restoration in Eq. (3) with the HGPS mask. The extra cost is a single additional scoring pass (one teacher forward and two group backward passes per scoring sample); the per-iteration adaptation cost is unchanged. The complete procedure is summarized in Algorithm 1 (Appendix B).

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

**Datasets and evaluation protocol.** We evaluate HGPS under two standard all-in-one image restoration protocols, following the configurations of PromptIR (Potlapalli et al., 2023) and IDR (Zhang et al., 2023), respectively. In the *three-task* setting, the backbone is pretrained on BSD400 (Martin et al., 2001) and WED (Ma et al., 2017) with additive Gaussian noise ($\sigma = 15, 25, 50$) for denoising, Rain100L (Yang et al., 2017) for deraining, and the OTS subset of RESIDE (Li et al., 2019) for dehazing. Adaptation and evaluation are performed on target test sets whose distributions deviate from the training data, without any test-time data augmentation: Kodak24 (Franzen, 1999) corrupted by Gaussian noise for denoising, the real-world O-HAZY (Ancuti et al., 2018) and DenseHaze (Ancuti et al., 2019) for dehazing, and Rain100H (Yang et al., 2017) for deraining. The *five-task* setting additionally includes deblurring and low-light enhancement, pretrained on GoPro (Nah et al., 2017) and LOL (Wei et al., 2018), and evaluates on two target domains with severe real-world shift: SDD (Li et al., 2023), a real defocus-blur benchmark, and VILNC-Indoor3 (Shi et al., 2024), the lowest-illumination tier of the real-world VILNC low-light collection. We report PSNR and SSIM (Wang et al., 2004) computed on the RGB channel against the ground-truth images.

**Compared methods.** We compare against representative AiOIR models without test-time adaptation, including AdaIR (Cui et al., 2025), VLU-Net (Zeng et al., 2025), and DATPRL-IR (Dong et al., 2026), and against the current state-of-the-art restoration TTA framework, DCTTA (Tang et al., 2026), instantiated on three backbones: StarIR (Cui et al., 2026), PromptIR (Potlapalli et al., 2023), and DFPIR (Tian et al., 2025). Within each backbone family, the adapted baseline (rows denoted +DCTTA) and our variant (+HGPS) share the entire host pipeline—degradation modeling, adaptation losses, EMA teacher, optimizer, per-tensor update budget $K_t$ (Eq. (2)), and post-step

restoration (Eq. (3))—and differ only in the parameter-selection criterion ($\alpha = 0.5$). Any performance gap within a family is therefore attributable to the membership criterion alone. The design variants of HGPS (HGPS-G, Gen-only, and the hybrid ratio $\alpha$) are studied separately in Section 4.4.

**Implementation details.** All experiments are conducted in PyTorch on a single NVIDIA A100 (40 GB) GPU, building upon the official implementation of the host pipeline with the pretrained backbones. Adaptation uses the Adam optimizer with learning rate $3 \times 10^{-5}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, batch size 1, and 3 adaptation iterations per test image on random $320 \times 320$ crops; the random seed is fixed to 23 for all methods. The EMA teacher uses momentum $0.45$ with teacher weight $5.0$, and the diffusion-based re-degradation generator performs $5$ sampling steps. Following the host pipeline, the Fisher protection ratio is $\rho = 0.6$, so each tensor of $n_t$ elements receives $K_t = n_t - \lfloor \rho n_t \rfloor$ update slots; HGPS splits them with $\alpha = 0.5$ by default, and the hard guard protects the $H_t = \max(n_t - 2K_t, 0)$ highest-Fisher elements (Eq. (7)). Both scores are computed once before adaptation using only the unlabeled degraded inputs; ground-truth images are never accessed. Within each backbone family, the adapted baseline is re-run under the identical learning rate, iteration count, seed, data order, and augmentation as the corresponding +HGPS configuration, ensuring a strictly controlled comparison.

### 4.2 Setting 1: Three-Task Adaptation

Table 1 reports the results under the three-task protocol. Among the non-adapted models, no single backbone dominates across tasks: PromptIR is strongest on Kodak24 denoising, whereas DFPIR and AdaIR are slightly better on the real-world dehazing benchmarks, and all of them remain far below their adapted counterparts on Rain100H, whose rain streaks depart most strongly from the Rain100L training distribution. Test-time adaptation closes much of this gap: the adapted baseline raises the average PSNR over the non-adapted backbones by $2.04$, $1.28$, and $2.15$ dB on StarIR, PromptIR, and DFPIR, respectively. Building on exactly the same pipeline and update budget, HGPS further improves the average on all three backbones and achieves the best within-family result on nearly every target set. Within the strictly controlled PromptIR comparison, the gain covers all four target sets and is the largest on the most shifted domain: Rain100H improves from $19.696$ to $22.041$ dB ($+2.35$ dB). HGPS is equally effective on the other backbones: on DFPIR it reaches the best Rain100H result overall ($22.829$ dB) and improves Kodak24 over the adapted baseline by $2.69$ dB ($19.879 \rightarrow 22.567$ dB). The only case where HGPS does not raise PSNR over its adapted counterpart is StarIR on Rain100H, where it trades $0.54$ dB PSNR for a substantial SSIM gain ($0.707 \rightarrow 0.730$), and the four-set average still improves. Overall, the improvements are consistent across backbones and target sets rather than concentrated on a single task, which supports our claim that the cross-view score $G$ re-allocates the fixed update budget toward elements whose updates actually generalize, instead of merely being safe to update.

### 4.3 Setting 2: Five-Task Adaptation

We further evaluate under the more demanding five-task protocol, which adds deblurring and low-light enhancement and adopts two target domains with severe real-world shift: real defocus blur (SDD) and extremely dark indoor scenes (VILNC-Indoor3). As shown in Table 2, the trends observed in the three-task setting persist: adaptation consistently improves over the non-adapted backbones, and HGPS further raises the average PSNR on all three backbones ($+0.53$, $+0.42$, and $+0.74$ dB on StarIR, PromptIR, and DFPIR, respectively), with DFPIR+HGPS achieving the best overall average ($18.82$ dB / $0.628$ SSIM). Within the PromptIR family, HGPS improves all five tasks, and the margins are the largest precisely where the shift is strongest: $+0.98$ dB on Kodak24 under Gaussian noise and $+0.62$ dB on the unseen real-world defocus domain SDD. This suggests that cross-view agreement is a more reliable selection criterion when the test distribution departs strongly from training—the regime where Fisher importance, estimated from the same shifted inputs, becomes a noisier proxy of parameter criticality. At the same time, VILNC-Indoor3 marks a clear boundary of the approach: the margins of all adapted variants are marginal there ($+0.02$ dB for the PromptIR family), and every adapted configuration remains behind the non-adapted DATPRL-IR, indicating that extremely dark scenes provide only a weak signal for consistency-based pseudo-supervision. We return to this limitation in Section 5.

Table 1: Quantitative comparison under the **three-task** setting (denoising, dehazing, deraining), without test-time data augmentation. PSNR↑ and SSIM↑ are reported. Within each backbone family (backbone, adapted baseline, +HGPS), the best result in each column is in **bold**; the three non-adapted models above are listed for reference. Rows with +HGPS are ours; within each family, +DCTTA and +HGPS share the same pipeline, update budget, and restoration schedule and differ only in the parameter-selection criterion. All rows are evaluated under the same unified protocol.

| Method | Denoising | | Dehazing | | | | Deraining | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Kodak24 | | O-HAZY | | DenseHaze | | Rain100H | | | |
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| AdaIR | 22.716 | 0.428 | 16.077 | 0.633 | 11.272 | 0.433 | 15.704 | 0.482 | 16.442 | 0.494 |
| VLU-Net | 21.623 | 0.373 | 15.975 | 0.623 | 11.696 | 0.433 | 15.465 | 0.474 | 16.190 | 0.476 |
| DATPRL-IR | 19.327 | 0.275 | 15.492 | 0.612 | 11.552 | 0.434 | 16.803 | 0.526 | 15.794 | 0.462 |
| StarIR | 22.857 | 0.502 | 15.871 | 0.621 | 11.577 | 0.435 | 15.914 | 0.510 | 16.555 | 0.517 |
| StarIR+DCTTA | 24.396 | 0.575 | 16.182 | 0.632 | 11.737 | **0.436** | **22.043** | 0.707 | 18.589 | 0.588 |
| StarIR+HGPS | **25.401** | **0.640** | **16.400** | **0.640** | **11.802** | **0.436** | 21.504 | **0.730** | **18.777** | **0.612** |
| PromptIR | 24.857 | 0.595 | 15.806 | 0.623 | 11.310 | 0.428 | 15.550 | 0.487 | 16.881 | 0.533 |
| PromptIR+DCTTA | 25.503 | 0.661 | 15.913 | 0.625 | 11.510 | 0.430 | 19.696 | 0.607 | 18.156 | 0.581 |
| PromptIR+HGPS | **25.585** | **0.683** | **16.010** | **0.628** | **11.577** | **0.431** | **22.041** | **0.652** | **18.803** | **0.596** |
| DFPIR | 16.960 | 0.214 | 16.164 | 0.632 | 11.553 | 0.432 | 16.089 | 0.506 | 15.192 | 0.446 |
| DFPIR+DCTTA | 19.879 | 0.302 | 16.277 | 0.635 | 11.677 | **0.433** | 21.551 | 0.685 | 17.346 | 0.514 |
| DFPIR+HGPS | **22.567** | **0.415** | **16.373** | **0.639** | **11.812** | **0.433** | **22.829** | **0.721** | **18.395** | **0.552** |

Table 2: Quantitative comparison under the **five-task** setting (denoising, dehazing, deraining, deblurring, low-light enhancement). Same conventions as Table 1. SDD and VILNC-Indoor3 are new target domains introduced in this work; all methods are evaluated on them under the same unified protocol.

| Method | Denoising | | Dehazing | | Deraining | | Deblurring | | Low-light | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Kodak24 | | O-HAZY | | Rain100H | | SDD | | VILNC-Indoor3 | | | |
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| AdaIR | 21.707 | 0.376 | 15.694 | 0.620 | 15.482 | 0.486 | 10.916 | 0.473 | 9.723 | 0.588 | 14.704 | 0.509 |
| VLU-Net | 20.125 | 0.309 | 15.431 | 0.611 | 16.218 | 0.507 | 11.396 | 0.475 | 9.984 | 0.562 | 14.631 | 0.493 |
| DATPRL-IR | 19.517 | 0.280 | 15.186 | 0.606 | 17.038 | 0.533 | 8.646 | 0.429 | 13.169 | 0.649 | 14.711 | 0.499 |
| StarIR | 21.297 | 0.374 | 15.760 | 0.620 | 15.649 | 0.501 | 17.384 | 0.571 | 8.170 | 0.460 | 15.652 | 0.505 |
| StarIR+DCTTA | 23.963 | 0.510 | 15.895 | 0.624 | **21.834** | 0.676 | 18.705 | 0.591 | 8.218 | 0.469 | 17.723 | 0.574 |
| StarIR+HGPS | **25.835** | **0.652** | **15.999** | **0.628** | 20.331 | **0.683** | **20.795** | **0.626** | **8.288** | **0.474** | **18.250** | **0.613** |
| PromptIR | 15.576 | 0.182 | 15.372 | 0.607 | 13.224 | 0.394 | 18.491 | 0.591 | 9.503 | 0.547 | 14.433 | 0.464 |
| PromptIR+DCTTA | 16.092 | 0.197 | 15.377 | 0.607 | 13.545 | 0.408 | 18.688 | 0.594 | 9.544 | 0.549 | 14.649 | 0.471 |
| PromptIR+HGPS | **17.073** | **0.227** | **15.396** | **0.608** | **13.998** | **0.424** | **19.304** | **0.605** | **9.568** | **0.550** | **15.068** | **0.483** |
| DFPIR | 20.332 | 0.326 | 15.918 | 0.626 | 16.007 | 0.508 | 21.066 | 0.622 | 9.683 | 0.613 | 16.601 | 0.539 |
| DFPIR+DCTTA | 22.203 | 0.408 | 16.054 | 0.631 | 21.254 | 0.677 | 21.162 | 0.626 | **9.714** | **0.614** | 18.077 | 0.591 |
| DFPIR+HGPS | **24.355** | **0.547** | **16.233** | **0.638** | **22.529** | **0.712** | **21.276** | **0.631** | 9.709 | **0.614** | **18.820** | **0.628** |

### 4.4 Ablation Study

All ablations are conducted on the PromptIR backbone under the three-task protocol and evaluated on Rain100H, the most shifted target domain, with every component of the host pipeline fixed except the selection rule.

**Effect of the hybrid ratio $\alpha$.** Table 3 ablates the fraction of per-tensor update slots assigned to the $G$ branch. Two observations stand out. First, every guarded hybrid configuration with a non-zero $G$ quota ($\alpha > 0$) improves over TIPS by at least 1.39 dB, so the benefit of $G$ does not hinge on a finely tuned $\alpha$. Second, the curve is unimodal: performance rises as the $G$ quota grows from $\alpha = 0.125$ (21.149 dB) to its peak at $\alpha = 0.375$ (22.108 dB), remains essentially tied at $\alpha = 0.5$ (22.041 dB), and degrades as $G$ takes over all slots (HGPS-G, 21.293 dB, though with a comparable SSIM). This confirms that generalization-driven and Fisher-driven criteria are complementary: $G$ promotes parameters whose updates transfer across views, while the low-Fisher branch retains a conservative, stability-oriented allocation for the remaining slots; we use $\alpha = 0.5$ as the default

Table 3: Ablation on the hybrid ratio $\alpha$ and the Fisher guard (three-task setting, PromptIR backbone, evaluated on Rain100H). $\alpha = 0$ recovers the TIPS selection rule and serves as an implementation-consistency check; Gen-only removes the hard guard entirely. The best results are in **bold**.

| Variant | $\alpha$ | PSNR↑ | SSIM↑ |
|---|---|---|---|
| HGPS | 0 | 19.675 | 0.607 |
| HGPS | 0.125 | 21.149 | 0.632 |
| HGPS | 0.25 | 21.986 | 0.647 |
| HGPS | 0.375 | **22.108** | 0.651 |
| HGPS (default) | 0.5 | 22.041 | 0.652 |
| HGPS | 0.625 | 22.005 | **0.656** |
| HGPS | 0.75 | 21.931 | 0.652 |
| HGPS | 0.875 | 21.087 | 0.653 |
| HGPS-G | 1.0 | 21.293 | **0.656** |
| TIPS (Fisher-only) | – | 19.696 | 0.607 |
| Gen-only (no guard) | – | 13.797 | 0.583 |

for its balanced PSNR/SSIM trade-off. Finally, the endpoint $\alpha = 0$ admits a precise interpretation: HGPS reduces to per-tensor low-Fisher selection and is therefore expected to coincide with TIPS. Empirically, the per-tensor update sets overlap by more than $99.99\%$ (up to equal-score ties) and the performance is nearly identical to the TIPS baseline ($19.675$ vs. $19.696$ dB), which serves as an implementation-level consistency check (Section 3.3).

**Necessity of the Fisher guard.** Comparing HGPS-G with Gen-only in Table 3 isolates the effect of the hard guard (Eq. (7)): both fill every update slot by $G$, but only the former forbids the $H_t$ highest-Fisher elements from being updated. Removing the guard is catastrophic: Gen-only collapses to $13.797$ dB, $7.50$ dB below HGPS-G and even $1.75$ dB below the un-adapted PromptIR ($15.550$ dB in Table 1). This shows that $G$, although effective, is not by itself a safe criterion—guarding the most degradation-critical parameters against any update is what keeps adaptation stable under severe distribution shift.

## 5 Conclusion

We revisited the parameter-selection step of Fisher-based test-time adaptation for all-in-one image restoration and identified its inherent one-sidedness: Fisher importance determines which parameters to protect, but not which of the remaining ones are worth updating. We proposed HGPS, a budget-neutral, per-tensor selection principle that complements Fisher stability with a cross-view gradient-agreement score, computed once before adaptation using only unlabeled target images under EMA-teacher pseudo-supervision. Plugged into the state-of-the-art DCTTA pipeline without any change to its optimization or restoration schedule, HGPS consistently improves restoration TTA across three backbones and two all-in-one protocols, by up to $2.69$ dB PSNR on Kodak24, and our ablations show that both the hybrid criterion and the Fisher hard guard are necessary. Since the rule modifies only the membership of per-tensor update sets, it is host-agnostic and can be combined with any Fisher-based selective adaptation scheme.

**Limitations and future work.** HGPS inherits the boundaries of its host pipeline: on extremely dark scenes (VILNC-Indoor3), consistency-based pseudo-supervision provides only a weak adaptation signal, and the gains of all adapted variants remain marginal. Moreover, the mixing ratio $\alpha$ is currently fixed globally, and the cross-view score relies on geometric view groups whose informativeness may vary across degradation types. Making $\alpha$ adaptive per tensor or per degradation, and extending the generalization-aware selection principle beyond Fisher-based hosts to other selective-update TTA frameworks, are promising directions for future work.

### AI use statement

In this work, we used generative AI tools for language editing and polishing of the manuscript. We have not used generative AI tools for research ideation, method design, experimental design, implementation, or the analysis and interpretation of results; these tasks are entirely the authors'

own work. We have reviewed all AI-assisted work and verified that it introduces no plagiarism, fabrication, or misrepresentation. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

### Ethics statement

This work studies test-time adaptation for all-in-one image restoration. It does not involve human subjects, and all experiments are conducted on publicly available benchmarks used in accordance with their licenses. The proposed method modifies only which parameters of a restoration model are updated at test time; we do not foresee direct harmful applications beyond those generally associated with image enhancement techniques, and we encourage downstream users to comply with applicable regulations on image manipulation.

### Reproducibility statement

Our method is fully specified in Section 3 and summarized in Algorithm 1 (Appendix B); all datasets, backbones, hyperparameters, and evaluation protocols are described in Section 4.1, and all ablation configurations are reported in Section 4.4. Our implementation builds upon the official DCTTA codebase, uses a fixed random seed for all compared methods, and changes only the mask-construction step of the host pipeline.

## References

Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 754–762, 2018.

Codruta O. Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-Haze: A benchmark for image dehazing with dense-haze and haze-free images. In *Proceedings of the IEEE International Conference on Image Processing*, pp. 1014–1018, 2019.

Ziyang Chen, Yiwen Ye, Yongsheng Pan, and Yong Xia. Gradient alignment improves test-time adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2429–2437, 2025.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Khan. AdaIR: Adaptive all-in-one image restoration via frequency mining and modulation. In *International Conference on Learning Representations*, 2025.

Yuning Cui, Syed Waqas Zamir, Ming-Hsuan Yang, Alois Knoll, Fahad Shahbaz Khan, and Salman Khan. StarIR: Convolutional image restoration with spatial-frequency fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:8216–8233, 2026.

Guanglu Dong, Chunlei Li, Chao Ren, Jingliang Hu, Yilei Shi, Xiao Xiang Zhu, and Lichao Mou. Learning domain-aware task prompt representations for multi-domain all-in-one image restoration. In *International Conference on Learning Representations*, 2026.

Rich Franzen. Kodak lossless true color image suite. `http://r0k.us/graphics/kodak/`, 1999.

Yuanbiao Gou, Haiyu Zhao, Boyun Li, Xinyan Xiao, and Xi Peng. Test-time degradation adaptation for open-set image restoration. In *Forty-first International Conference on Machine Learning*, 2024.

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, 2019.

Yu Li, Dongwei Ren, Xinya Shu, and Wangmeng Zuo. Learning single image defocus deblurring with misaligned training pairs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1495–1503, 2023.

Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

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, 2017.

David Martin, Charless 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 IEEE International Conference on Computer Vision*, volume 2, pp. 416–423, 2001.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one blind image restoration. In *Advances in Neural Information Processing Systems*, 2023.

Yiqi Shi, Duo Liu, Liguo Zhang, Ye Tian, Xuezhi Xia, and Xiaojing Fu. Zero-IG: Zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3015–3024, 2024.

Ni Tang, Shenghao Nie, Xiaotong Luo, Yuan Xie, and Yanyun Qu. Degradation-consistent test-time adaptation for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15476–15485, 2026.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28165–28175, 2025.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

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.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.

Wenhan Yang, Robby T. Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1357–1366, 2017.

Yijun Yang, Hongtao Wu, Angelica I. Aviles-Rivero, Yulun Zhang, Jing Qin, and Lei Zhu. Genuine knowledge from practice: Diffusion test-time adaptation for video adverse weather removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25606–25616, 2024.

Haijin Zeng, Xiangming Wang, Yongyong Chen, Jingyong Su, and Jie Liu. Vision-language gradient descent-driven all-in-one deep unfolding networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7524–7533, 2025.

Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5825–5835, 2023.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, 2022.

## A Host Pipeline Details

For completeness, we restate the per-iteration adaptation objective of the host pipeline (DCTTA), which HGPS leaves strictly unchanged.

The host per-iteration objective (Tang et al., 2026) couples an EMA teacher $f_{\bar{\boldsymbol{\theta}}}$ with test-time re-degradation. Given the dihedral view group $\mathcal{T} = \{\tau_m\}_{m=1}^{8}$, the teacher restores every augmented view of an input $x$; the restorations are mapped back to canonical coordinates and, over $R$ independent rounds, averaged into a pseudo-label $\bar{y}(x)$ with a per-pixel confidence map $c(x)$:

$$\bar{y}(x) \;=\; \frac{1}{R}\sum_{r=1}^{R} \bar{y}^{(r)}(x), \qquad c(x) \;=\; \tfrac{3}{2} - \sigma\big(\mathrm{Var}_r\,[\bar{y}^{(r)}(x)] \,/\, 4{\times}10^{-4}\big), \tag{A.1}$$

where $\bar{y}^{(r)}(x) = \frac{1}{|\mathcal{T}|}\sum_{m=1}^{8} \tau_m^{-1}\big(f_{\bar{\boldsymbol{\theta}}}(\tau_m(x))\big)$ is the round-$r$ average of the teacher's restorations in canonical coordinates, $\sigma(\cdot)$ is the sigmoid function, and the variance is computed per pixel across the $R$ rounds. A lightweight diffusion-based re-degradation generator $D_\psi$ is updated on the fly with the pair $(\bar{y}(x), x)$ and produces a re-degraded sample $\tilde{x}$ that shares the target degradation of $x$. The student is adapted on $\tilde{x}$ with a degradation-consistency objective,

$$\mathcal{L}_{\mathrm{adapt}}(x;\boldsymbol{\theta}) \;=\; \mathcal{L}_{\mathrm{dc}}\big(f_{\boldsymbol{\theta}}(\tilde{x}),\, \mathrm{sg}[f_{\boldsymbol{\theta}}(x)];\, c(x)\big) \;+\; \lambda_{\mathrm{tea}}\,\mathcal{L}_{\mathrm{dc}}\big(f_{\boldsymbol{\theta}}(\tilde{x}),\, \mathrm{sg}[\bar{y}(x)];\, c(x)\big), \tag{A.2}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator: the first term enforces consistency with the student's own restoration of the original input, and the second term distills the teacher pseudo-label with weight $\lambda_{\mathrm{tea}} = 5$. The per-term loss combines a confidence-weighted pixel loss, a perceptual loss $\mathcal{L}_{\mathrm{per}}$, and an $L_1$ loss on the low-pass subband $\mathcal{W}_{LL}(\cdot)$ of a one-level wavelet transform,

$$\mathcal{L}_{\mathrm{dc}}(a, b;\, c) \;=\; \big\|c \odot a \;-\; c \odot b\big\|_1 \;+\; \lambda_{\mathrm{per}}\,\mathcal{L}_{\mathrm{per}}(a, b) \;+\; \lambda_{\mathrm{wav}}\,\big\|\mathcal{W}_{LL}(a) - \mathcal{W}_{LL}(b)\big\|_1, \tag{A.3}$$

with $\lambda_{\mathrm{per}} = 0.01$ and $\lambda_{\mathrm{wav}} = 0.1$. After each optimizer step on $\mathcal{L}_{\mathrm{adapt}}$, the teacher is updated by exponential moving average, $\bar{\boldsymbol{\theta}} \leftarrow m\,\bar{\boldsymbol{\theta}} + (1-m)\,\boldsymbol{\theta}$. All components of Eqs. (A.1)–(A.3) are identical to the host pipeline; HGPS only redefines the restore mask $M_t$ (Eq. (3) in the main text).

## B Pseudocode of HGPS

Algorithm 1 summarizes the complete HGPS procedure described in Section 3: the one-time scoring pass, the per-tensor construction of the update sets and the restore mask, and the adaptation loop with post-step restoration.

**Algorithm 1** HGPS: per-tensor hybrid generalization-aware parameter selection

**Require:** Pretrained model $f_{\boldsymbol{\theta}^0}$, EMA teacher $f_{\bar{\boldsymbol{\theta}}}$, unlabeled target stream $\{x_s\}$, protection ratio $\rho$, mixing ratio $\alpha$

**Scoring (once, no model update):** compute Fisher scores $F$ via Eq. (1) and cross-view utility scores $G$ via Eqs. (4)–(6) on the same target samples
**for** each candidate parameter tensor $t$ with $n_t$ elements **do**
  $K_t \leftarrow n_t - \lfloor \rho n_t \rfloor$;  $H_t \leftarrow \max(n_t - 2K_t, 0)$
  $\mathcal{H}_t \leftarrow$ top-$H_t$ elements of tensor $t$ by $F$ {hard guard}
  $\mathcal{U}_{G,t} \leftarrow$ top-round$(\alpha K_t)$ elements of $t \setminus \mathcal{H}_t$ by $G$
  $\mathcal{U}_{F,t} \leftarrow$ bottom-$(K_t - |\mathcal{U}_{G,t}|)$ elements of $t \setminus (\mathcal{H}_t \cup \mathcal{U}_{G,t})$ by $F$
  $M_t \leftarrow \mathbb{1}[\cdot \notin \mathcal{U}_{G,t} \cup \mathcal{U}_{F,t}]$ {1: restore, 0: keep update}
**end for**
**for** each adaptation iteration (host update on the target stream) **do**
  compute $\mathcal{L}_{\text{adapt}}$ (Eq. (A.2)), backpropagate, and perform an optimizer step
  restore protected positions via Eq. (3) with masks $\{M_t\}$
**end for**

## C Additional Visual Comparisons

Figure A.1 shows a visual comparison on a degraded test image under the three-task setting (Section 4.2).




Figure A.1: Visual comparison under the three-task setting. From left to right: the degraded input, the adapted baseline (+DCTTA) and our variant (+HGPS) on the PromptIR, StarIR, and DFPIR backbones, and the ground truth (GT). Within each backbone family, +DCTTA and +HGPS share the same pipeline, update budget, and restoration schedule and differ only in the parameter-selection criterion; +HGPS recovers details closer to the GT.

### C.1 Cross-Modality Generalization to MRI Denoising

To further assess the robustness of HGPS under a cross-modality shift, we evaluate the three pretrained backbones, without any re-training, on 20 T1-weighted MRI scans from the MR-T1 test set of DDcGAN (Ma et al., 2020), corrupted by additive Gaussian noise ($\sigma = 80$). All adaptation hyperparameters are identical to those in Section 4.1 (learning rate $3 \times 10^{-5}$, 3 adaptation iterations per image, $\alpha = 0.5$). Table A.1 reports the quantitative results and Figure A.2 shows visual comparisons.

HGPS improves both PSNR and SSIM over the non-adapted backbone and the +DCTTA baseline on PromptIR and DFPIR, and DFPIR+HGPS achieves the best overall result (16.386 dB / 0.2556 SSIM). On StarIR, HGPS trades PSNR for a substantial SSIM gain (0.5191 vs. 0.2356), consistent with the behavior observed on Rain100H in Table 1. Notably, the absolute scores of all methods are far below their in-distribution counterparts, confirming that MR-T1 constitutes a severe cross-modality shift; even under this regime, the generalization-aware re-selection of HGPS remains beneficial on two of the three backbones.

Table A.1: Cross-modality generalization to T1-weighted MRI denoising (MR-T1 test set of DDc-GAN, 20 images, Gaussian noise $\sigma = 80$). PSNR↑ and SSIM↑ are reported. Within each backbone family (backbone, +DCTTA, +HGPS), the best result in each column is in **bold**. All rows share the same unified protocol and adaptation hyperparameters; +DCTTA and +HGPS differ only in the parameter-selection criterion.

| Method | PSNR↑ | SSIM↑ |
|---|---|---|
| PromptIR | 13.561 | 0.1738 |
| PromptIR+DCTTA | 13.658 | 0.1742 |
| PromptIR+HGPS | **13.728** | **0.1746** |
| StarIR | 14.602 | 0.1994 |
| StarIR+DCTTA | **15.935** | 0.2356 |
| StarIR+HGPS | 14.709 | **0.5191** |
| DFPIR | 15.730 | 0.2186 |
| DFPIR+DCTTA | 16.146 | 0.2354 |
| DFPIR+HGPS | **16.386** | **0.2556** |

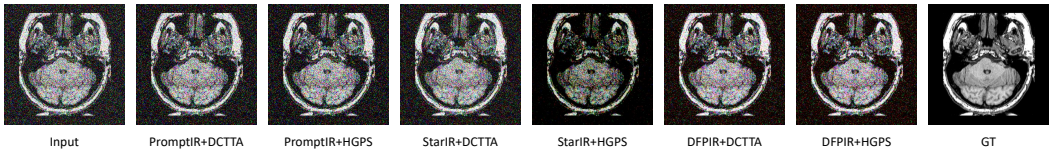


Figure A.2: Visual comparison on a noisy T1-weighted MRI scan (MR-T1, $\sigma = 80$). From left to right: the noisy input, the adapted baseline (+DCTTA) and our variant (+HGPS) on the PromptIR, StarIR, and DFPIR backbones, and the ground truth (GT).