SemISP: Semantic-Consistent Diffusion for RAW-to-sRGB Generation

Zhe Pang[1,2,3] **Juan Wang**[2,3] **Haina Qin**[4] **Ziyu Feng**[5]
Zewen Chen[1,2,3] **Bing Li**[1,2,3*] **Weiming Hu**[1,2,3,6]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Institute of Automation, Chinese Academy of Sciences
[3]Beijing Key Laboratory of Super Intelligent Security of Multi-Modal Information
[4]Kuaishou Technology [5]Beijing Jiaotong University [6]ShanghaiTech University
pangzhe2025@ia.ac.cn bli@nlpr.ia.ac.cn

Abstract

We study RAW-to-sRGB generation from smartphone captures to the reference rendering of a specified DSLR, training a separate model for each fixed camera pair. This task must bridge both cross-camera sensor discrepancies and the representation gap between linear RAW and nonlinearly rendered sRGB images. Existing methods mainly condition on pixel-level supervision and local RAW cues, where content preservation and domain translation remain entangled in low-level measurements. We propose SemISP, a semantic-guided diffusion framework that uses scene semantics as a domain-invariant anchor for cross-domain generation. Specifically, we adopt a frozen DINOv3 as the sRGB semantic backbone and develop a Domain Alignment Adapter to extract sRGB-consistent semantic representations from RAW inputs through parameter-efficient adaptation. Because semantic priors and RAW signals operate at fundamentally different information levels, and the diffusion denoising process follows a coarse-to-fine trajectory, we further design a Time-aware Dual-stream Adaptive Modulation (TDAM) module that separately encodes both condition streams and uses the diffusion timestep to dynamically balance their contributions—emphasizing semantic anchoring at early stages and fine-grained RAW details at later ones. Experiments on two established cross-camera benchmarks show consistent improvements over prior methods in both fidelity and perceptual quality.

1 Introduction

Smartphone cameras capture RAW data that must be processed by an image signal processing (ISP) pipeline to produce displayable sRGB images [30]. Yet because smartphone sensors and optics are fundamentally limited, their outputs still fall short of DSLR cameras in color rendition and fine detail. This gap has motivated learning smartphone RAW-to-target-DSLR sRGB mappings. We train a separate model for each fixed source–target camera pair, using only source-camera RAW at inference. As illustrated in Fig. 1(a), cross-camera mapping must additionally account for differences in optics, sensor response, and ISP tuning. Details captured by the DSLR may be weakly represented in the smartphone RAW, while camera-specific processing introduces further differences in color and tone. End-to-end learning addresses these acquisition and rendering differences through paired source RAW and target sRGB supervision [15, 31].

Early regression-based methods [7, 50, 15, 14] have made steady progress with pixel-wise supervision, but deterministic predictions often average ambiguous mappings, weakening textures and

*Corresponding author: Bing Li (bli@nlpr.ia.ac.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

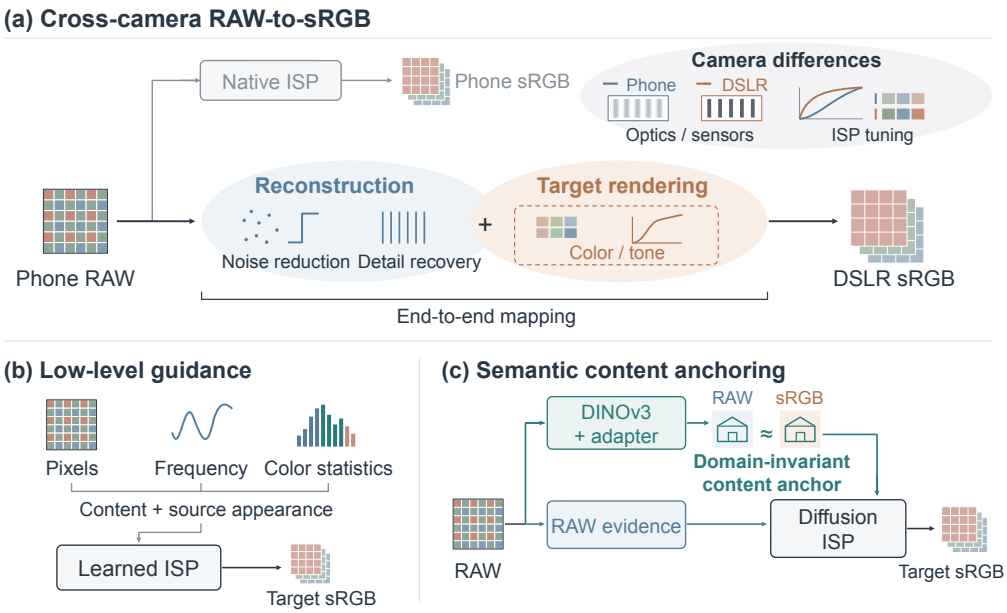

Figure 1: **Motivation for semantic content anchoring in cross-camera ISP.** (a) Cross-camera mapping must bridge differences in optics, sensors, and ISP tuning. (b) Low-level cues encode scene content together with source-camera appearance. (c) SemISP combines a domain-invariant content anchor with RAW evidence to guide target rendering. The RAW/sRGB symbols illustrate shared scene content; inference uses only source RAW.

perceptual realism. Diffusion models [45, 31] offer stronger generative priors, yet existing RAW-to-sRGB methods still mostly rely on domain-dependent low-level cues, such as grayscale maps [31], wavelet/frequency subbands [7, 11], and global color statistics [50]. As illustrated in Fig. 1(b), these low-level cues encode scene content together with source-camera appearance. Content preservation and appearance translation must therefore be learned from cues that themselves vary across domains, complicating the establishment of a stable content reference.

Our key insight is that scene semantics remain largely consistent across RAW and sRGB despite substantial changes in low-level appearance. We therefore use RAW-adapted semantic representations as a domain-invariant content anchor, while retaining RAW observations for photometric and fine spatial evidence (Fig. 1(c)). Together, these conditions guide content-preserving generation toward the target-camera rendering.

Motivated by these observations, we propose SemISP, a semantic-guided diffusion framework for cross-camera RAW-to-sRGB generation. SemISP introduces two key components. First, we adopt a frozen DINOv3 [34] as the sRGB semantic backbone and design a lightweight Domain Alignment Adapter to extract sRGB-consistent semantic priors from RAW inputs, making the semantic anchor reliable despite the RAW-sRGB representation gap. Second, we inject this anchor together with the original RAW observation through a Time-aware Dual-stream Adaptive Modulation (TDAM) module. The semantic branch provides a stable content reference, while the RAW branch retains image measurements and local details. Both jointly guide the generation of the target-camera appearance. Since diffusion denoising naturally follows a coarse-to-fine trajectory [24, 40], TDAM encodes semantic and RAW conditions in separate streams and modulates their contributions according to the timestep, allowing semantic anchoring to dominate early and RAW details to refine texture and edges later.

The main contributions of this paper are summarized as follows:

• We propose SemISP, a semantic-guided diffusion framework for RAW-to-sRGB generation. A Domain Alignment Adapter is introduced to bridge the representation gap between RAW and sRGB domains, enabling reliable semantic prior extraction from RAW inputs via the frozen DINOv3 backbone.

• We design a Time-aware Dual-stream Adaptive Modulation (TDAM) module that decouples semantic priors from RAW detail cues and leverages diffusion timesteps to dynamically balance their contributions, achieving stage-wise control over global appearance and local detail generation.

• Extensive experiments on two established cross-camera benchmarks show that SemISP consistently outperforms prior methods across both fidelity and perceptual quality metrics.

2 Related work

2.1 Learned image signal processing

Cross-camera RAW-to-sRGB reconstruction aims to bridge compact smartphone sensors and DSLR-quality outputs. Early learned ISP methods model this mapping with end-to-end CNNs, including PyNET [15], CycleISP [47], and CameraNet [18]. To handle real RAW–RGB misalignment, AWNet [7] and LiteISPNet [50] introduce wavelet and optical-flow-based alignment, while other works improve mobile efficiency [29] or use frequency priors to decouple structure and style [11]. Despite these advances, regression-based models often produce over-smoothed results. Recent generative and controllable methods, such as DiffRAW [45], ISPDiffuser [31], and ModularISP [1], improve perceptual quality but still largely depend on low-level RAW conditions.

2.2 Diffusion models in low-level vision

Diffusion models [12] provide strong natural-image priors for ill-posed low-level vision problems and have been widely used in super-resolution [33, 8, 17], inpainting [23, 35], and restoration [27, 42, 3]. For RAW-to-sRGB generation, DiffRAW [45] uses color-position preserving conditions to alleviate misalignment, and ISPDiffuser [31] decouples grayscale texture reconstruction from color mapping. However, their conditions remain tied to low-level RAW measurements, where content preservation and domain translation are still coupled. We instead leverage *cross-domain semantic invariance* to provide a domain-invariant anchor while leaving RAW observations to guide texture and edge recovery.

3 Method

3.1 Preliminary: flow matching

Our framework is based on Conditional Flow Matching (CFM) [19, 20], which models RAW-to-sRGB generation as continuous transport from noise to data. Let the target sRGB image be x_0, the noise sample be $x_1 \sim \mathcal{N}(0, I)$, and the linear path be $x_t = (1-t)x_0 + tx_1$ for $t \in [0, 1]$. We learn a conditional velocity field $v_\theta(x_t, t, c)$, where c denotes the RAW input and semantic priors. Since the target velocity is $\dot{x}_t = x_1 - x_0$, CFM minimizes:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\|v_\theta(x_t, t, c) - (x_1 - x_0)\|_2^2\right]. \tag{1}$$

During inference, we start from $x_{t=1} = x_1$ and use the Euler update $x_{t+\Delta t} = x_t + v_\theta(x_t, t, c)\Delta t$ to integrate toward $t = 0$, enabling high-quality generation with few sampling steps.

3.2 Architecture overview

As illustrated in Fig. 2, SemISP first extracts RAW-adapted semantic priors using a frozen DINOv3 backbone and a lightweight Domain Alignment Adapter. These priors are then paired with the original RAW observation and passed to TDAM, which produces scale-wise conditions for the denoising U-Net. In this design, the semantic branch provides a domain-invariant content anchor, the RAW branch supplies local physical evidence, and the U-Net predicts the flow-matching velocity $v_\theta(x_t, t, c)$ under both conditions.

3.3 RAW semantic adaptation

Inspired by ViT-Adapter [5], we use a frozen DINOv3 backbone with an interactive Domain Alignment Adapter. We first apply bilinear demosaicing to obtain an RGB-formatted input for DINOv3.

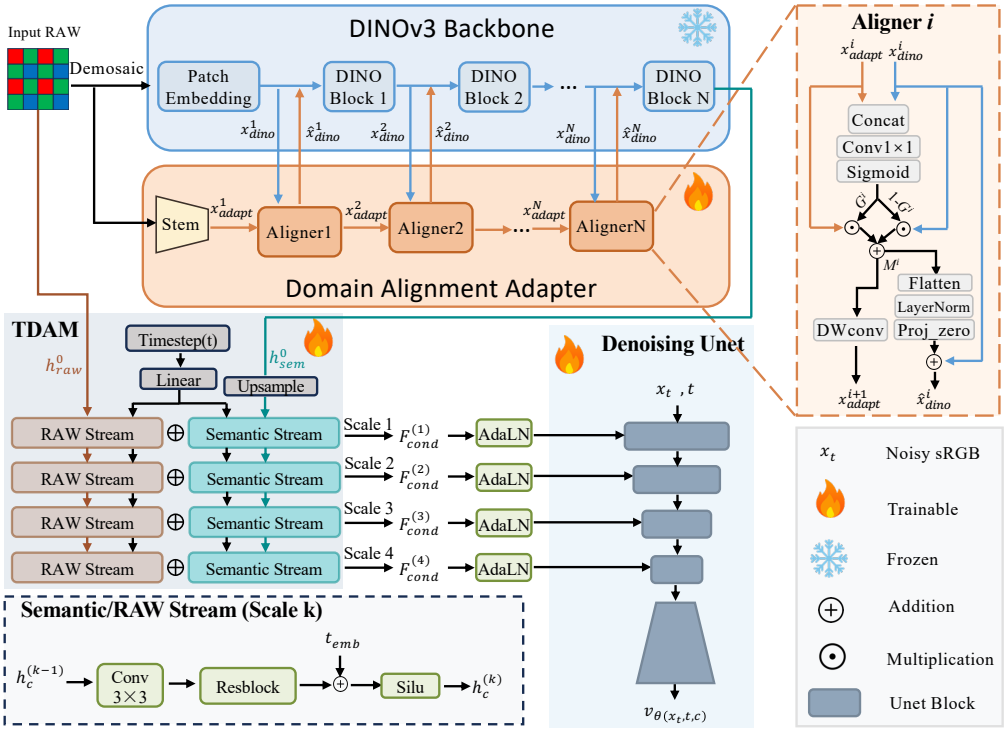

Figure 2: **Overview of the proposed semantic-guided RAW-to-sRGB diffusion framework.** The Domain Alignment Adapter adapts a frozen DINOv3 to RAW inputs through a convolutional stem and gated Aligners, producing sRGB-consistent semantic priors. TDAM first modulates the semantic and RAW condition streams with the timestep at each scale, then sums them and injects the resulting condition into the corresponding U-Net block via AdaLN-Zero. "Stem" denotes multi-layer convolutional downsampling.

In parallel, the adapter processes the RAW-domain signal with a four-stage strided convolutional stem, producing an initial adapter feature x^1_{adapt}. This feature is progressively updated by Aligners, matching the layer-wise semantic flow of DINOv3.

While cross-attention [39] is widely adopted as the standard paradigm for feature interaction [44, 32], directly employing dot-product matching across different domains can be severely compromised by significant discrepancies in embedding distributions. Such domain shifts often lead to unstable or unreliable spatial associations between the raw physical signals and the abstract semantic priors. To circumvent this issue, we replace the global cross-attention with a pixel-level gating mechanism, thereby selectively injecting only the most reliable RAW evidence into the backbone semantic representations. Specifically, for each DINOv3 transformer block $i \in \{1, \ldots, L\}$, we concatenate the backbone semantic representation x^i_{dino} and the input to the i-th Aligner in the adapter x^i_{adapt} along the channel dimension prior to the i-th block, and predict a dynamic gating map G^i via a 1×1 convolution followed by a Sigmoid activation:

$$G^i = \sigma(\text{Conv}_{1\times1}([x^i_{\text{adapt}} || x^i_{\text{dino}}])). \tag{2}$$

Subsequently, we obtain the mixed representation M^i via a convex combination:

$$M^i = G^i \odot x^i_{\text{adapt}} + (1 - G^i) \odot x^i_{\text{dino}}, \tag{3}$$

where $\odot$ denotes element-wise multiplication. This gating mechanism adaptively adjusts the contributions of the two pathways at different spatial locations, balancing the utilization of semantic information and RAW side-branch information.

The mixed feature M^i is routed into two paths. First, a depthwise separable convolution [13] updates the adapter state x^{i+1}_{adapt} for the next Aligner. Second, M^i is flattened, normalized by LayerNorm, and projected through a zero-initialized linear layer. The result is injected into the frozen DINOv3 stream as a residual update [9]:

$$\hat{x}^i_{\text{dino}} = x^i_{\text{dino}} + \text{Proj}_{\text{zero}}(\text{LayerNorm}(\text{Flatten}(M^i))). \tag{4}$$

The zero initialization keeps the original DINOv3 representation unchanged at the beginning of training and gradually learns RAW-domain corrections. After the final block, the adapted DINOv3 feature is used as the initial semantic prior h^0_{sem}.

3.4 Time-aware dual-stream adaptive modulation

TDAM converts the semantic prior and RAW observation into scale-wise conditions for the denoising U-Net. Unlike static or unified condition fusion in many conditional diffusion frameworks [25, 32, 44, 48], TDAM explicitly accounts for the different roles of the two modalities along the denoising trajectory.

TDAM decouples condition encoding into two parallel streams. The semantic stream takes the adapted semantic prior h^0_{sem}, upsamples it by bilinear interpolation to match the RAW resolution, and encodes domain-invariant scene structure. The RAW stream takes the original RAW observation and preserves high-frequency local evidence. Each stream follows a lightweight multi-scale structure similar to T2I-Adapter [25], consisting of convolutional layers and residual blocks, and produces features at four scales aligned with the U-Net hierarchy.

The key to TDAM lies in injecting timestep information deeply into the condition encoding process, rather than merely processing it uniformly after fusion. Specifically, let $h^{(k)}_c$ denote the output feature of stream $c \in \{\text{sem}, \text{raw}\}$ at scale level k, and $\mathcal{F}^{(k)}_c$ denote the corresponding residual transformation. We project the timestep embedding t_{emb} via a linear mapping $\mathcal{W}^{(k)}_c$ to generate a channel-wise bias, which is added to the convolutional features prior to activation:

$$h^{(k)}_c = \text{SiLU}\Big(\mathcal{F}^{(k)}_c(h^{(k-1)}_c) + \mathcal{W}^{(k)}_c(t_{\text{emb}})\Big). \tag{5}$$

This design enables the two streams to adaptively adjust their responses to their respective modalities at different timesteps. Since the dual-streams output features with identical dimensions at each scale level, we perform element-wise addition on the two time-modulated features at scale k to obtain the final multi-scale conditional representation:

$$F^{(k)}_{\text{cond}} = h^{(k)}_{\text{sem}} + h^{(k)}_{\text{raw}}. \tag{6}$$

Following the AdaLN-Zero conditioning paradigm in DiT [26], $F^{(k)}_{\text{cond}}$ is injected into the corresponding U-Net block to guide velocity prediction.

3.5 Training strategy

Our training pipeline is explicitly divided into two sequential stages.

Stage 1: RAW semantic alignment. This stage independently optimizes the Domain Alignment Adapter. We feed paired clean sRGB images into an independent frozen DINOv3 to extract reference token features $F_{\text{srgb}} \in \mathbb{R}^{B\times N\times D}$, and extract predicted token features $F_{\text{sem}} \in \mathbb{R}^{B\times N\times D}$ from the corresponding RAW images using DINOv3 modulated by our adapter. We then minimize the token-wise cosine distance:

$$\mathcal{L}_{\text{align}} = \frac{1}{BN}\sum_{b=1}^{B}\sum_{n=1}^{N}\left(1 - \frac{F^{b,n}_{\text{sem}} \cdot F^{b,n}_{\text{srgb}}}{\|F^{b,n}_{\text{sem}}\|_2 \, \|F^{b,n}_{\text{srgb}}\|_2}\right). \tag{7}$$

Minimizing this loss aligns RAW-derived tokens with the sRGB semantic space of the pre-trained model.

Table 1: Quantitative comparisons on the ZRR and MAI test sets. Best and second-best results are shown in **bold** and underlined. "−" denotes unavailable official code. ISPDiffuser† denotes the paper-reported results with GT-based best-of-10 selection; the unmarked row reports our single-sample evaluation.

Method	ZRR (Original GT)			ZRR (Align GT with RAW)			MAI		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
PyNet [15]	21.19	0.747	0.193	22.73	0.845	0.152	23.81	0.848	0.139
AWNet-R [7]	21.42	0.748	0.198	23.27	0.854	0.151	24.53	0.872	0.136
AWNet-D [7]	21.53	0.749	0.212	23.38	0.850	0.164	24.64	0.866	0.147
MW-ISPNet [14]	21.42	0.754	0.213	23.07	0.848	0.165	25.02	0.885	0.133
LiteISPNet [50]	21.55	0.749	0.187	23.76	0.873	0.133	24.90	0.877	0.123
FourierISP [11]	21.65	0.755	0.182	23.93	0.874	0.124	25.37	0.891	0.072
DiffRAW [45]	21.31	0.743	0.145	–	–	–	–	–	–
ISPDiffuser [31]	21.35	0.744	0.155	23.62	0.881	0.114	25.08	0.893	0.071
ISPDiffuser† [31]	21.77	0.754	0.157	24.09	0.881	0.111	25.64	0.894	0.071
Ours	**22.21**	**0.758**	**0.132**	**24.66**	**0.893**	**0.097**	**26.06**	**0.898**	**0.063**

Stage 2: Semantic-guided generative model training. After Stage 1, we freeze the trained Domain Alignment Adapter and DINOv3 backbone, and jointly optimize TDAM and the denoising U-Net. The total loss consists of the Conditional Flow Matching objective $\mathcal{L}_{\mathrm{CFM}}$ (Eq. 1) and a perceptual loss $\mathcal{L}_{\mathrm{lpips}}$ [49]:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CFM}} + \lambda_{\mathrm{lpips}}\mathcal{L}_{\mathrm{lpips}}, \tag{8}$$

where we empirically set $\lambda_{\mathrm{lpips}} = 0.1$.

4 Experiments

4.1 Experimental settings

Datasets. We conduct experiments on two widely used RAW-to-sRGB benchmarks: the Zurich RAW-to-RGB (ZRR) [15] dataset and the Mobile AI (MAI) [14] dataset. The ZRR dataset provides paired RAW images from a Huawei P20 smartphone and sRGB ground truths from a Canon 5D Mark IV DSLR camera, comprising 46,839 training pairs and 1,204 testing pairs. The MAI dataset maps RAW data from a Sony IMX586 sensor to the color space of a Fujifilm DSLR, comprising a total of 12,080 training pairs. Since the official test set does not provide paired ground-truth images, following the setting of ISPDiffuser [31], we split the official training set into 90% for training and 10% for testing.

Implementation details. Our U-Net model adopts the same architecture as DDPM [12], where the timestep t is embedded into the U-Net blocks via addition. We employ the DINOv3-S [34] (ViT-Small) as our semantic backbone. The model is trained at a resolution of 256×256 and tested at full resolution. We set the batch size to 16 and employ the AdamW optimizer [22] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is annealed from 2×10^{-4} to 10^{-6} using a cosine decay schedule [21]. We adopt a uniform time schedule with 1,000 training timesteps, while using only 2 sampling steps during the inference phase. All experiments are conducted on 4 NVIDIA V100 GPUs for 200K iterations.

Evaluation metrics. We report PSNR, SSIM [41], and LPIPS [49] for full-reference evaluation. To bypass the sensitivity of these metrics to spatial misalignments during cross-camera evaluation, we further adopt no-reference MUSIQ [16] and TOPIQ [2] to assess visual quality.

4.2 Comparison with state-of-the-art methods

Compared methods and evaluation settings. We compare our method with representative end-to-end RAW-to-sRGB methods, including regression-based pipelines (PyNet [15], AWNet-R/D [7], MW-ISPNet [14], LiteISPNet [50]) and strong baselines incorporating explicit priors or generative modeling (FourierISP [11], DiffRAW [45], ISPDiffuser [31]). Given the inevitable registration errors in the ZRR dataset caused by dual-device capturing, we follow the standard protocol [31, 11, 50] to report results under both the *Original GT* and *Align GT with RAW* settings. The aligned setting

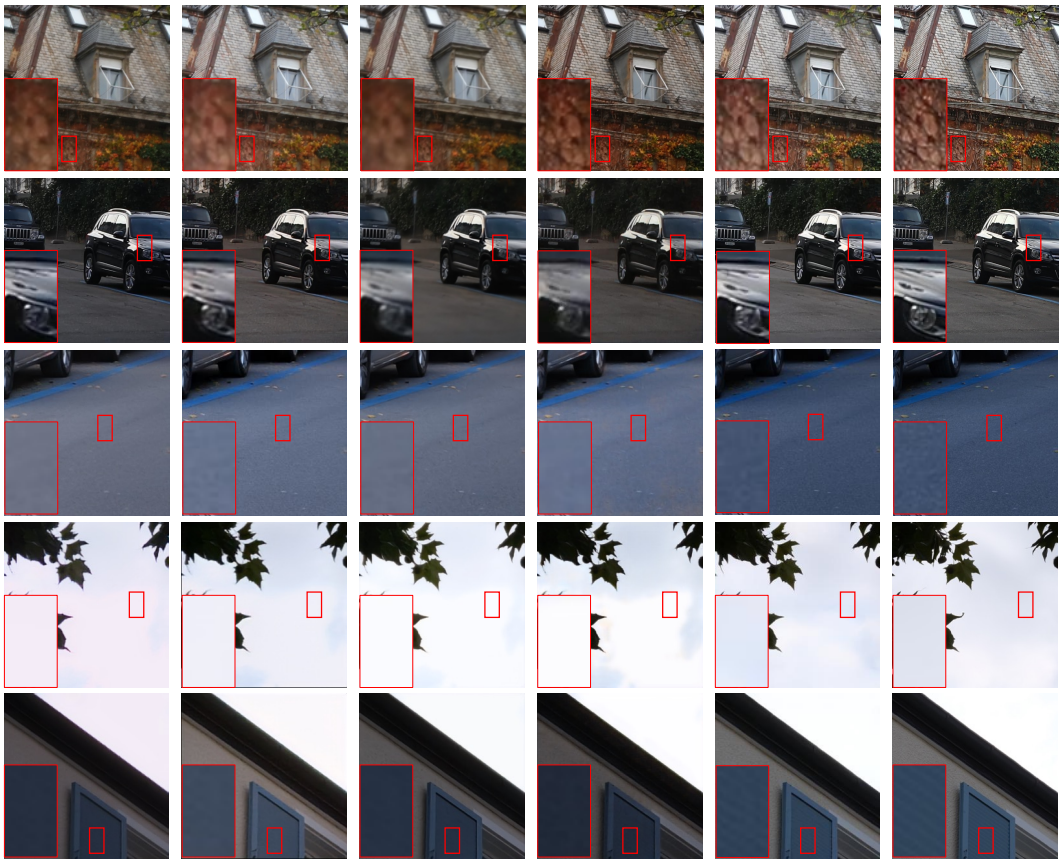

Figure 3: Qualitative comparisons on ZRR and MAI datasets. While baselines suffer from color distortions and blurriness due to the domain gap, our method reconstructs DSLR-like global color distributions while preserving sharp local details. Zoom in for better visualization.

computes the metrics after warping the GT sRGB images to the RAW coordinate system via optical flow [50]. For methods without officially available source code (e.g., DiffRAW), we directly cite the metrics reported in their original papers.

Quantitative comparisons. Table 1 shows that our method achieves the best results on both ZRR and MAI. On ZRR, it improves over FourierISP by +0.56 dB PSNR and −0.050 LPIPS under the Original GT setting, and by +0.73 dB PSNR under the aligned setting. Compared with single-sample ISPDiffuser, our method further reduces LPIPS by 0.017 on aligned ZRR and improves MAI by +0.98 dB PSNR and −0.008 LPIPS. These results indicate consistent gains in both fidelity and perceptual quality on the evaluated test sets.

Non-reference perceptual metrics. From an evaluation perspective independent of paired ground truth, our method achieves the best MUSIQ and TOPIQ scores on both datasets (Table 2). This illustrates that our outputs inherently possess superior perceptual quality and photographic naturalness.

Qualitative comparisons. Fig. 3 shows representative results on ZRR and MAI. Baselines often blur highlight boundaries, shift local colors, or weaken fine textures. In contrast, our method better preserves reflection geometry, edge sharpness, and texture details, producing results closer to the ground truth.

4.3 Ablation study

Effectiveness of the Domain Alignment Adapter. We first verify whether the proposed domain alignment adapter can effectively alleviate the representation discrepancy between RAW and sRGB, and evaluate its structural design advantages over existing adapters.

Table 2: No-reference perceptual metrics. Best in **bold**, second in underlined.

Method	ZRR		MAI	
	MUS.↑	TOP.↑	MUS.↑	TOP.↑
PyNet	43.796	0.362	39.823	0.445
AWNet-R	43.441	0.355	40.211	0.441
AWNet-D	45.100	0.362	39.839	0.432
MW-ISPNet	42.448	0.340	40.652	0.449
LiteISPNet	47.310	0.370	40.365	0.445
FourierISP	44.534	0.369	47.614	0.535
ISPDiffuser	50.117	0.392	48.032	0.535
Ours	**51.452**	**0.412**	**50.663**	**0.573**

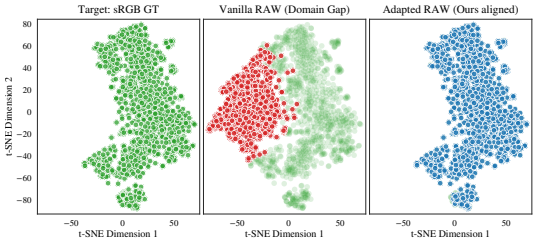

Figure 4: t-SNE visualization of semantic alignment. Our adapter effectively aligns RAW features with the sRGB distribution.

1) Verification of semantic alignment capability. We evaluate RAW-to-sRGB semantic alignment via a linear probe on ADE20K [51] (Table 3). Following InvISP [43], we convert sRGB images to RAW as controlled inputs. We first train a linear segmentation head on frozen DINOv3 features extracted from sRGB images, and then keep both DINOv3 and the head fixed to test (i) unadapted RAW inputs and (ii) RAW inputs with the adapter enabled. The results show that the adapter substantially improves semantic alignment, mapping RAW inputs closer to the sRGB semantic manifold. We further visualize feature distributions with t-SNE using 1,000 paired samples from the ZRR test set. As shown in Fig. 4, unadapted RAW features are separated from sRGB features, while adapted RAW features overlap markedly with the sRGB cluster, supporting the effectiveness of our domain alignment.

2) Comparison of adapter architectures. We further replace our adapter with two representative visual adapters: a cross-attention-based adapter (ViT-Adapter [5]) and a RAW-oriented adapter (RAW-Adapter [6]) (Table 4). The cross-attention baseline (ViT-Adapter) struggles with cross-domain distribution mismatches, while RAW-Adapter injects overly strong low-level signals, biasing the semantic branch toward textures and noise. In contrast, our method selectively filters RAW cues and preserves stable semantics, achieving the best results.

Table 3: Semantic alignment evaluation on the ADE20K validation set. The linear segmentation head is trained only on sRGB features (Setting 1) and kept *frozen* when evaluating RAW inputs (Settings 2 and 3).

Setting	Input	Adapter	Head	mIoU / mAcc (%) ↑
1	sRGB	None	Trained	46.7 / 57.5
2	RAW	None	Frozen	40.1 / 51.4
3	RAW	Trained	Frozen	46.4 / 56.9

Table 4: Quantitative comparison of different adapter architectures on ADE20K. We freeze the DINOv3 backbone and train each adapter under the same setting. Best in **bold**.

Adapter Architecture	Params	mIoU (%) ↑	mAcc (%) ↑
w/o Adapter (Baseline)	–	40.1	51.4
ViT-Adapter [5]	5.8M	45.2	56.1
RAW-Adapter [6]	2.6M	43.9	54.0
Ours	**1.4M**	**46.4**	**56.9**

Table 5: Ablation study of the proposed components on the ZRR dataset. "Sem." denotes semantic alignment ("Imp." indicates implicit training constraint). "Dual" indicates the dual-stream architecture, and "t" denotes timestep-aware modulation. Best results are in **bold**.

Variant	Components			Metrics		
	Sem.	Dual	t	PSNR ↑	SSIM ↑	LPIPS ↓
- without Semantic Prior	✗	✗	✗	20.54	0.726	0.195
- replace with Implicit Align.	Imp.	✗	✗	21.44	0.739	0.163
- replace with Single-stream	✓	✗	✗	21.02	0.743	0.168
- without Timestep Mod.	✓	✓	✗	21.65	0.751	0.140
SemISP (Ours)	✓	✓	✓	**22.21**	**0.758**	**0.132**

Effectiveness of core generation components. In this section, we conduct ablation studies on semantic prior, dual-stream decoupling, and timestep-aware modulation to validate the effectiveness

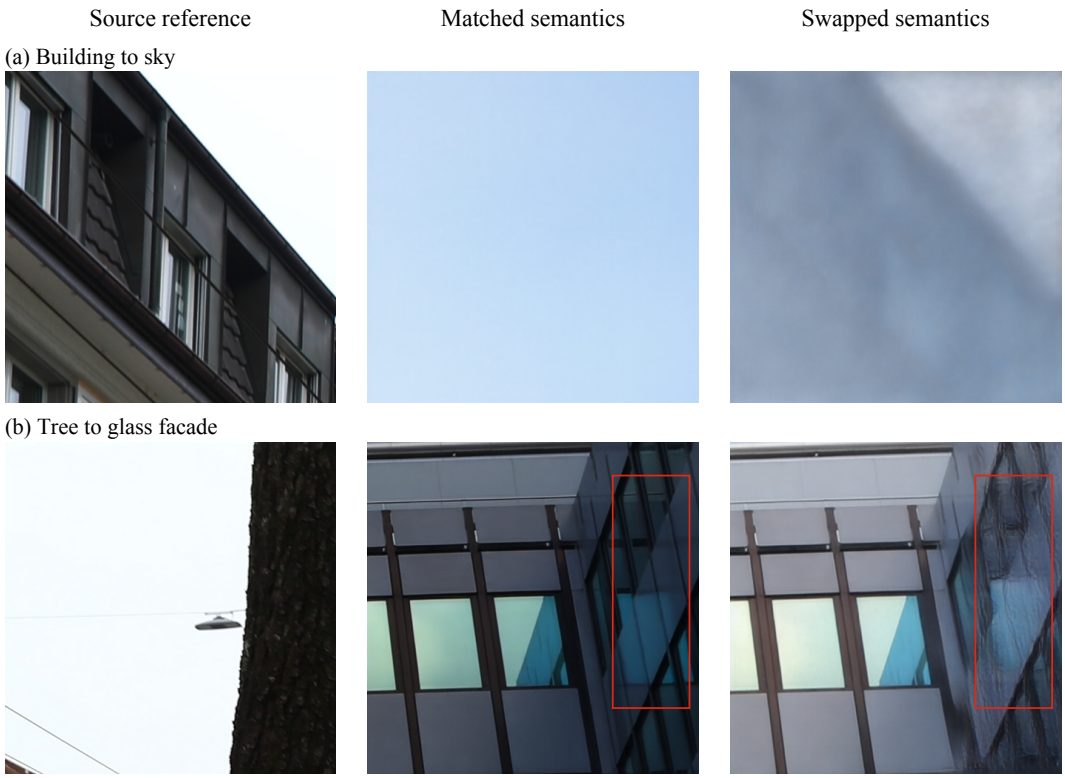

Figure 5: **Content dependence of semantic guidance.** Replacing the semantic source introduces source-related coarse contours (a) and material textures (b), with the target RAW, model weights, and initial noise fixed. Source photographs are shown for reference; conditioning features are extracted from the corresponding RAW inputs through the adapter. Boxes mark corresponding regions.

of the different components in our proposed method. Specifically, - *without Semantic Prior* is used to establish the baseline performance relying solely on RAW inputs. - *replace with Implicit Align.* is used to compare the effects of implicit training constraints against our explicit semantic injection, where the implicit constraint follows [36, 46] by aligning the intermediate features of the U-Net with the sRGB semantic features extracted by DINOv3. - *replace with Single-stream* directly concatenates features to test the superiority of dual-stream decoupling in avoiding mutual interference between high-level semantics and low-level physical cues. - *without Timestep Mod.* is used to evaluate the importance of adaptively coordinating features across different denoising stages. The quantitative results in Table 5 demonstrate that *SemISP (Ours)* performs best in simultaneously improving global color consistency and local detail fidelity.

Content dependence of semantic guidance. We fix the trained model, target RAW input, and initial noise, and replace only the semantic features with those extracted from another RAW image. As shown in Fig. 5, building-derived features introduce a coarse diagonal boundary into the sky, while tree-derived features induce bark-like textures on a glass facade. With the architecture and parameters unchanged, these source-related changes demonstrate that the generator uses content carried by the semantic condition. Additional examples are provided in Appendix E.

Table 6: Ablation on modulation mechanisms.

Modulation	PSNR↑	SSIM↑	LPIPS↓
Concat	22.01	0.751	0.134
AdaLN	21.87	0.744	0.138
Add (Ours)	**22.21**	**0.758**	**0.132**

Table 7: ISP comparison on ZRR (Original GT).

Method	PSNR↑	SSIM↑	LPIPS↓
InvISP [43]	19.85	0.695	0.266
Modular ISP [1]	20.76	0.729	0.386
SemISP (Ours)	**22.21**	**0.758**	**0.132**

Table 8: Comparison of different visual backbones for semantic prior extraction on the ZRR dataset. We freeze each backbone and train the adapter under the same setting.

Semantic Backbone	Params	Metrics		
		PSNR ↑	SSIM ↑	LPIPS ↓
DeiT-S [37]	22M	21.53	0.742	0.156
MoCov3-S [4]	21M	21.65	0.744	0.147
MAE-B [10]	86M	21.37	0.735	0.158
CLIP-B [28]	86M	21.84	0.741	0.137
SigLIP 2-B [38]	86M	21.95	0.754	0.134
DINOv3-S (Ours) [34]	21M	<u>22.21</u>	<u>0.758</u>	<u>0.132</u>
DINOv3-B [34]	86M	**22.38**	**0.760**	**0.128**

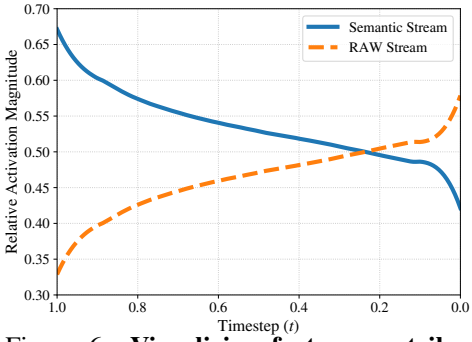

Figure 6: **Visualizing feature contribution dynamics.** Semantic features dominate early, whereas RAW features dominate late.

Modulation mechanism. We ablate different ways of injecting the timestep t into the two condition streams of TDAM. As shown in Table 6, Add performs best across all metrics. By directly providing a timestep-dependent bias to each residual block, Add avoids the extra projection required by Concat and the normalization-induced response changes of AdaLN, preserving spatial evidence and the absolute strength of RAW and semantic cues. This lightweight modulation is therefore well suited to balancing the two streams without disturbing their modality-specific information.

Challenges of cross-camera ISP. Unlike same-camera ISP, smartphone-to-DSLR mapping must bridge differences in sensor characteristics and camera rendering, while ZRR additionally involves spatial misalignment. These factors make reconstruction more challenging and help explain the lower absolute scores compared with same-camera benchmarks. Table 7 compares SemISP with the same-camera ISP method InvISP and Modular ISP on ZRR. SemISP outperforms both, supporting the effectiveness of our end-to-end modeling for this cross-camera task.

Comparison of semantic backbones. We compare several pre-trained visual encoders [37, 4, 10, 38, 28, 34] for semantic prior extraction under the same adapter setting, using the smallest officially available variant of each model; for encoders whose minimum release is Base, we additionally report DINOv3-S/B for completeness. Table 8 shows that DINOv3 performs best across scales, likely due to its stronger patch-level token features that provide finer spatial conditioning for dense RAW-to-sRGB generation. Scaling the target encoder yields only marginal gains, so we use DINOv3-S by default to balance quality and efficiency.

Visualizing feature contribution dynamics. To unveil the working mechanism of TDAM, we quantify the relative activation magnitudes of the Semantic and RAW streams across different timesteps ($t \in [1.0, 0.0]$) by calculating the mean absolute values of their feature maps. As shown in Fig. 6, this dynamic alternation aligns with the coarse-to-fine generation paradigm, demonstrating that TDAM successfully achieves a staged conditional division of labor: relying on semantic priors for global structures in early stages, while dominating RAW features for high-frequency detail restoration later.

5 Conclusion

This work studies how domain-invariant semantics can be reliably established from RAW inputs and used to anchor RAW-to-sRGB generation. SemISP first adapts a frozen DINOv3 backbone with a Domain Alignment Adapter, reducing the representation gap between linear RAW observations and the sRGB semantic space while retaining the backbone's semantic capacity. It then uses TDAM to keep semantic anchoring and RAW detail evidence in separate condition streams and coordinate them with diffusion timesteps, so global appearance and local fidelity are emphasized at appropriate stages. Experiments confirm strong quantitative and visual performance. Looking ahead, we plan to extend the framework to more challenging RAW scenarios, such as extreme low-light enhancement, and further optimize the architecture for efficient on-device deployment.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. U24A20331, 62536002, 62625311, 62473363, 62672491, 62192782, and 62532015), the Beijing Natural Science Foundation (Grant Nos. L251005 and L223003), and the Xinjiang Leading Talents Introduction Program (Grant No. XJRC-2025-KJ-YJ-CXPT-182). The authors declare no competing interests.

References

[1] Mahmoud Afifi, Zhongling Wang, Ran Zhang, and Michael S Brown. Modular neural image signal processing. *arXiv preprint arXiv:2512.08564*, 2025.

[2] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

[3] I Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsuan Yang, et al. UniRestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17979, 2025.

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[5] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations*, 2023.

[6] Ziteng Cui and Tatsuya Harada. RAW-Adapter: Adapting pre-trained visual model to camera RAW images. In *European Conference on Computer Vision*, pages 37–56. Springer, 2024.

[7] Linhui Dai, Xiaohong Liu, Chengqi Li, and Jun Chen. AWNet: Attentive wavelet network for image ISP. In *European Conference on Computer Vision*, pages 185–201. Springer, 2020.

[8] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[11] Xuanhua He, Tao Hu, Guoli Wang, Zejin Wang, Run Wang, Qian Zhang, Keyu Yan, Ziyi Chen, Rui Li, Chengjun Xie, et al. Enhancing RAW-to-sRGB with decoupled style structure in Fourier domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2130–2138, 2024.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[13] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[14] Andrey Ignatov, Radu Timofte, Zhilu Zhang, Ming Liu, Haolin Wang, Wangmeng Zuo, Jiawei Zhang, Ruimao Zhang, Zhanglin Peng, Sijie Ren, et al. AIM 2020 challenge on learned image signal processing pipeline. In *European Conference on Computer Vision*, pages 152–170. Springer, 2020.

[15] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera ISP with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 536–537, 2020.

[16] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021.

[17] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[18] Zhetong Liang, Jianrui Cai, Zisheng Cao, and Lei Zhang. CameraNet: A two-stage framework for effective camera ISP learning. *IEEE Transactions on Image Processing*, 30:2248–2262, 2021.

[19] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[20] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

[21] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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

[24] Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. DeCo: Frequency-decoupled pixel diffusion for end-to-end image generation. *arXiv preprint arXiv:2511.19365*, 2025.

[25] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[27] Haina Qin, Wenyang Luo, Libin Wang, Dandan Zheng, Jingdong Chen, Ming Yang, Bing Li, and Weiming Hu. Reversing flow for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7545–7558, 2025.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[29] Daniel Wirzberger Raimundo, Andrey Ignatov, and Radu Timofte. LAN: Lightweight attention-based network for RAW-to-RGB smartphone image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 808–816, 2022.

[30] Rajeev Ramanath, Wesley E Snyder, Youngjun Yoo, and Mark S Drew. Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1):34–43, 2005.

[31] Yang Ren, Hai Jiang, Menglong Yang, Wei Li, and Shuaicheng Liu. ISPDiffuser: Learning RAW-to-sRGB mappings with texture-aware diffusion models and histogram-guided color consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6722–6730, 2025.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[33] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[34] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[35] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with Fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022.

[36] Yuchuan Tian, Hanting Chen, Mengyu Zheng, Yuchen Liang, Chao Xu, and Yunhe Wang. U-REPA: Aligning diffusion U-Nets to ViTs. *arXiv preprint arXiv:2503.18414*, 2025.

[37] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[38] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding. *arXiv preprint arXiv:2502.14786*, 2025.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[40] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025.

[41] 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.

[42] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. DiffIR: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023.

[43] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6287–6296, 2021.

[44] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[45] Mingxin Yi, Kai Zhang, Pei Liu, Tanli Zuo, and Jingduo Tian. DiffRAW: Leveraging diffusion model to generate DSLR-comparable perceptual quality sRGB from smartphone RAW images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6711–6719, 2024.

[46] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

[47] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. CycleISP: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2696–2705, 2020.

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[50] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning RAW-to-sRGB mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4348–4358, 2021.

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017.

Supplementary Material

A Architecture and Capacity of the Domain Alignment Adapter

Adapter Architecture. Table A.1 summarizes the layer-wise design of our Domain Alignment Adapter, consisting of a 4-stage strided convolutional stem and L Aligners aligned with the L frozen DINOv3 blocks.

Adapter Capacity. Table A.2 reports an ablation on the base dimension C using ADE20K linear probing. We adopt C=128 by default, as increasing C brings marginal gains while significantly increasing parameters, and reducing C degrades alignment.

Table A.1: Architecture details of the proposed Domain Alignment Adapter. We assume the input RAW image has a spatial resolution of $H \times W$. The Base Dimension C is set to 128 to ensure high efficiency, while the frozen DINOv3-S embedding dimension D is 384. L denotes the total number of DINOv3 transformer blocks.

Stage	Output Size	Layer Operations	Channels
Stem (Outputs x^1_{adapt})	$\frac{H}{2} \times \frac{W}{2}$	Conv2d (3×3, stride 2), BN, GELU	$C/8$ (16)
	$\frac{H}{4} \times \frac{W}{4}$	Conv2d (3×3, stride 2), BN, GELU	$C/4$ (32)
	$\frac{H}{8} \times \frac{W}{8}$	Conv2d (3×3, stride 2), BN, GELU	$C/2$ (64)
	$\frac{H}{16} \times \frac{W}{16}$	Conv2d (3×3, stride 2), BN, GELU	C (128)
Aligner i $i \in \{1, \ldots, L\}$	$\frac{H}{16} \times \frac{W}{16}$	DINO feature projection: Linear($384 \to 128$), LayerNorm Reshape to 2D spatial map $\to x^i_{\text{dino}}$	C (128)
	$\frac{H}{16} \times \frac{W}{16}$	Dynamic Gating Map: $G^i = \sigma(\text{Conv}_{1\times1}([x^i_{\text{adapt}} \| x^i_{\text{dino}}]))$	C (128)
	$\frac{H}{16} \times \frac{W}{16}$	Convex Combination (Mixed Feature): $M^i = G^i \odot x^i_{\text{adapt}} + (1 - G^i) \odot x^i_{\text{dino}}$	C (128)
	$\frac{H}{16} \times \frac{W}{16}$	Adapter State Update via Depthwise Separable Convs: $x^{i+1}_{\text{adapt}} = \text{Pointwise-Conv}(\text{GELU}(\text{DW-Conv}(M^i)))$	C (128)
	$N \times 384$	Residual injection to DINOv3: $\hat{x}^i_{\text{dino}} = x^i_{\text{dino}} + \text{Proj}_{\text{zero}}(\text{LayerNorm}(\text{Flatten}(M^i)))$ Forward $\hat{x}^i_{\text{dino}}$ to the next frozen DINOv3 block	D (384)

Table A.2: Ablation study on the Base Dimension (C) of the Domain Alignment Adapter. We evaluate the semantic alignment capability using linear probing on the ADE20K validation set. $C = 128$ offers the best trade-off between semantic representation and computational efficiency.

Base Dimension (C)	Params (M) ↓	mIoU (%) ↑	mAcc (%) ↑
64	0.3	45.9	56.3
128 (**Default**)	1.4	46.4	56.9
256	4.5	46.4	57.0
384	9.0	46.5	57.1

B Characteristics of Evaluated Semantic Backbones

Table B.1 summarizes why different pre-training objectives lead to different dense-token properties for RAW-to-sRGB conditioning. The goal is not to rank foundation models universally, but to explain why DINOv3 is better aligned with our adapter setting and pixel-aligned generation task.

Table B.1: Taxonomy of evaluated semantic backbones. We focus on the spatial token properties that matter for dense RAW-to-sRGB generation. The empirical trend refers to the backbone comparison in Table 8.

Backbone	Pre-training objective	Dense-token property	Implication for RAW-to-sRGB conditioning	Trend
DeiT-S [37]	Image-level supervision	Strong category discrimination, but spatial tokens tend to emphasize discriminative regions.	Useful for object recognition, less ideal for full-frame rendering where background and local layout must remain stable.	Lower
MoCo v3-S [4]	Global contrastive learning	Robust global instance representation with weaker direct constraints on dense tokens.	Provides scene-level cues, but spatial conditioning can be less precise for pixel-aligned restoration.	Lower
MAE-B [10]	Masked image modeling	Encodes local structure through reconstruction, but semantics are learned implicitly.	Captures low-level layout, yet may not sufficiently bridge RAW and sRGB semantic domains.	Lower
CLIP-B [28] / SigLIP 2-B [38]	Vision-language alignment	Strong open-vocabulary image representation; dense tokens are secondary to global alignment.	Global semantics are helpful, but local token consistency can be fragmented for strict ISP supervision.	Medium
DINOv3-S/B [34]	Dense self-distillation	Spatial tokens preserve part-level and layout-aware semantics under local/global views.	Better matches our need for domain-invariant, pixel-aligned semantic anchors.	Best

This analysis is consistent with Table 8: DINOv3-S already provides a strong quality-efficiency trade-off, while scaling to DINOv3-B yields only moderate gains. We therefore use DINOv3-S by default to retain dense semantic guidance without making the semantic branch unnecessarily expensive.

C Pixel-Space Generative Modeling for RAW-to-sRGB

RAW-to-sRGB translation is a pixel-aligned dense prediction problem, but the mapping from noisy smartphone RAW measurements to DSLR-style sRGB images can be one-to-many in texture, color rendition, and local detail. SemISP therefore keeps generation in pixel space and directly optimizes the target image distribution while preserving access to pixel-level supervision:

$$x_t = (1-t)x_0 + tx_1, \quad v_\theta(x_t, t, c) \approx x_1 - x_0, \quad c = \{h^0_{\text{sem}}, \text{RAW}\}. \tag{C.1}$$

Table C.1: Why pixel-space generation is suitable for RAW-to-sRGB. The task requires both perceptual plausibility and strict spatial correspondence to paired supervision.

Modeling choice	Strength	Risk for this task	Role in SemISP
Deterministic regression ISP	Efficient and directly supervised in pixel space.	Pixel losses may average plausible outputs, causing oversmoothed textures under severe information loss.	Serves as strong efficiency baseline.
Pixel-space diffusion/flow	Models a conditional image distribution while keeping pixel-level alignment and losses.	More expensive than lightweight regression networks.	Chosen trade-off for high-fidelity RAW-to-sRGB generation.
Latent diffusion	Efficient generation in a compressed representation and strong semantic prior when pre-trained at scale.	Compression can weaken fine RAW evidence and strict pixel-level supervision.	Useful reference paradigm, but less directly matched to dense ISP supervision.

This design lets the semantic prior guide global structure and appearance while the RAW condition continues to constrain local evidence such as edges, texture boundaries, and color transitions.

D U-Net Design for Multi-Scale Conditioning

SemISP uses a U-Net whose spatial hierarchy provides an interface for conditioning at multiple resolutions. Cross-camera RAW-to-sRGB generation requires consistent color and tone over large regions together with faithful local structure. The encoder–decoder hierarchy aggregates context at coarser resolutions and reconstructs spatial detail at finer resolutions, while skip connections pass features between corresponding levels. This organization supports the joint use of semantic content and RAW evidence throughout generation.

Table D.1: How the U-Net and TDAM support RAW-to-sRGB generation. Each design component addresses a spatial or temporal requirement of conditional rendering.

Task requirement	Architectural mechanism	Role in SemISP
Global rendering consistency	The encoder–decoder hierarchy aggregates context across spatial resolutions.	Semantic conditioning provides a content reference at the corresponding U-Net scales.
Fine spatial fidelity	High-resolution processing and skip connections retain spatial information.	RAW-conditioned features supply local evidence for edge and texture reconstruction.
Guidance across denoising stages	TDAM processes semantic and RAW conditions in parallel and modulates both streams with the timestep.	The contributions of the two conditions can vary throughout generation.

At each scale, TDAM separately encodes the semantic features and RAW observations before combining their outputs for injection into the U-Net through AdaLN-Zero. The spatial hierarchy determines where the conditions enter the generator, while timestep modulation determines how their responses change during denoising. Together, these mechanisms connect the semantic content anchor and RAW image evidence to the reconstruction process at multiple resolutions.

E Additional Semantic Content Interventions

Figure E.1 presents all four semantic-source replacement examples on ZRR. Each pair of matched and swapped outputs uses the same target RAW, trained EMA weights, initial noise (seed 2221101), and two-step sampler. Only the semantic features are replaced; they are extracted from the source RAW through the Domain Alignment Adapter, without feature rescaling or retraining. Reference photographs are shown only to identify the source content and are not used as conditioning inputs. In (d), source 655 is horizontally flipped, and the displayed reference shows the same transformation.

The sky examples show source-related coarse boundaries and window-shaped shading. On structured architectural targets, tree-derived features induce bark-like surface textures while the main window and beam layout remains recognizable. These examples extend the content-dependence observations in Fig. 5 to both smooth and structured targets.



Figure E.1: **Semantic-source replacement on ZRR.** Each row shows a source reference, the output with matched semantics, and the output with swapped semantics. Rows (a)–(d) use target/source pairs 135/1143, 161/460, 431/871, and 846/655, respectively; source 655 is horizontally flipped. The target RAW and sampling noise are fixed within each row. Boxes mark corresponding regions.



Figure F.1: Additional qualitative comparisons on the **ZRR dataset**.

F More Qualitative Results

In this section, we provide additional qualitative comparisons on the ZRR and MAI datasets (Fig. F.1 and Fig. F.2). Compared to baselines, SemISP produces more consistent colors and preserves finer details across challenging scenes.

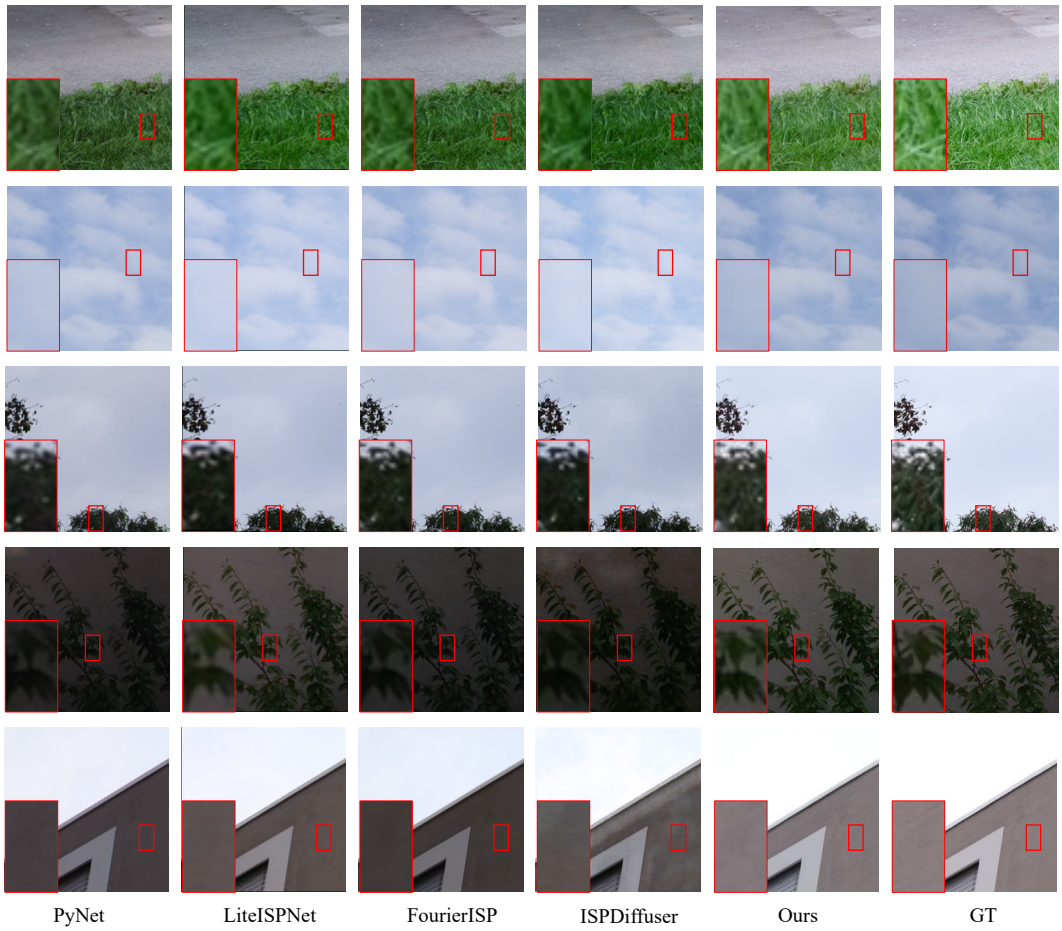

Figure F.2: Additional qualitative comparisons on the **MAI dataset**.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three contributions: the semantic-guided SemISP framework with a Domain Alignment Adapter, TDAM, and extensive experimental validation. These claims are supported by the method description in Section 3 and experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 identifies extensions to more challenging RAW scenarios and further optimization for on-device deployment.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper is primarily empirical and does not present formal theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 provides the benchmark datasets, data split protocol, model resolution, backbone choice, optimizer, learning-rate schedule, training timesteps, sampling steps, hardware, and training iterations. Additional adapter architecture details are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The datasets used (ZRR and MAI) are publicly available benchmarks. This version does not provide a public code-repository link.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies dataset splits, optimizer settings, batch size, learning-rate schedule, training and inference timesteps, image resolution, hardware, and evaluation metrics. Appendix A further reports the adapter capacity choice.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following standard practice in the RAW-to-sRGB literature, we report single-run results.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Section 4.1 reports training on 4 NVIDIA V100 GPUs for 200K iterations. GPU memory and wall-clock training time are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research focuses on image processing and does not involve human subjects, personal data, or dual-use concerns beyond standard image generation considerations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The work may improve smartphone camera quality and make high-fidelity photography more accessible. Since the method is a task-specific RAW-to-sRGB converter conditioned on captured RAW inputs, its misuse risk is lower than that of general-purpose image generators, though generated image enhancement can still affect visual authenticity.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The model performs task-specific RAW-to-sRGB conversion within a constrained pipeline and poses minimal risk for misuse compared to general-purpose image generators.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the ZRR and MAI RAW-to-sRGB benchmarks, ADE20K for semantic-alignment evaluation, and the pre-trained visual backbones used in the experiments, including DINOv3. These are publicly available research assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper does not release new datasets or pre-trained models at submission time.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This research does not involve crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This research does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.