# FlowISP: Latent Space ISP Transfer via Conditional Flow Matching

Taehyeon Kim Jinwoo Kim Seon Joo Kim
Department of Artificial Intelligence, Yonsei University
{taehyeon2006, jinwoo-kim, seonjookim}@yonsei.ac.kr

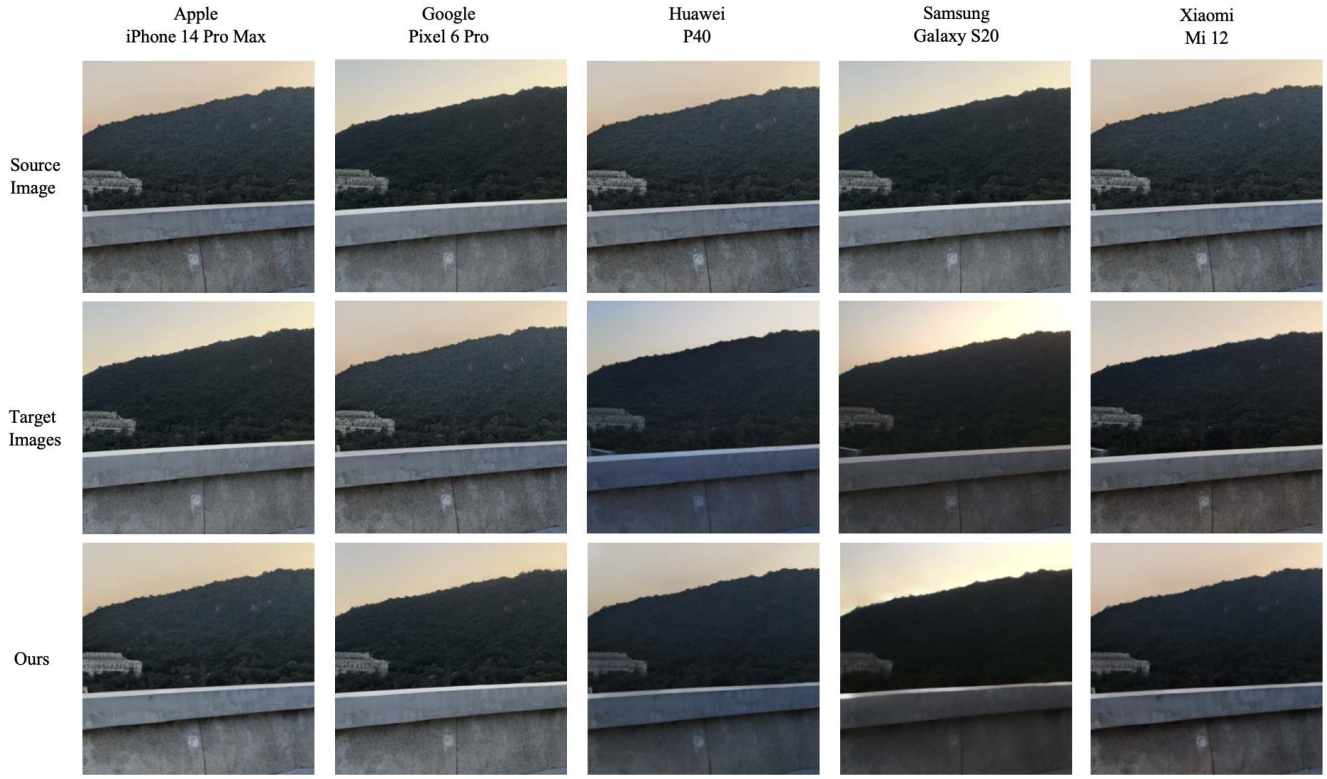



Figure 1. Cross-device ISP transfer results. The top row displays the source images captured by the source ISP. The middle row shows the ground truth images captured by the target ISPs (indicated by the columns). The bottom row shows the results of Ours, where our method transforms the source image to match the color and tone characteristics of each target ISP.

## Abstract

*Modern camera Image Signal Processors (ISP) exhibit a series of complex, non-linear transformation that are highly device-dependent. Therefore, even the same scene appears differently across cameras, creating a barrier to consistent cross-device visual aesthetics. While deep learning methods have shown promise in modeling ISP transfer, existing models suffer from high computational cost due to transfer in pixel-space and rigidly confined to a specific pair of device and domain. In this paper, we propose FlowISP, a framework for ISP style transfer that operates entirely in the latent space of a pretrained latent diffusion model. We furthermore formulate style transfer task using Conditional Flow Matching [7] and introducing specialized Diffusion Transformer (DiT) [10] architecture that explicitly conditions temporal dynamics and device-specific ISP style separately. This design allows precise style manipulation while preserving the structure of the original scene. Experiments on multi-device ISP datasets demonstrate that FlowISP successfully reproduces target ISP aesthetics while strictly preserving scene geometry.* d

# 1. Introduction

The Image Signal Processor (ISP) is the most pivotal part of the modern digital cameras. It transforms RAW images to sRGB images. This process involves a complex series of non-linear operations including demosaicing, white balancing, tone-mapping and color-correction. Consequently, different ISP pipelines exhibit distinct visual styles, tones and color signatures even when capturing the same scene. This device-dependent ISP style is problematic to computational photography. Distinct camera models create isolated domains. A model trained to edit images captured by a specific model often fails catastrophically when applied to other ISP styles as it cannot account for the unique domains inherent to each ISP. Furthermore, these style variances constrain users to the aesthetic signature of their cameras. If users desire specific color tone or nostalgic simulation of a film camera, currently people are forced to acquire that specific hardware.

A natural solution is ISP style transfer [3]. Given an sRGB image processed by a source ISP, generate an sRGB image as if it was processed by a target ISP with a consistent scene. Existing approaches predominantly operate in pixel space. Image-to-image translation architecture like CNN, GAN and diffusion models map directly from one ISP domain to other. However, these pixel space transfer methodologies require heavy computational resources and are sensitive to low-level noise, requiring perfectly pixel-wise aligned image pairs. Meanwhile, latent diffusion models like Stable Diffusion [12] provide a learned latent space including rich image statistics, but the power of this latent space has not been fully exploited in ISP style transfer.

In this work, we propose FlowISP, a novel latent ISP style transfer framework that operates in the latent space of a frozen Stable Diffusion VAE. Rather than employing standard diffusion dynamics, we adopted conditional flow matching [7]. We viewed the style transfer as learning continuous probability path – distribution of source ISP to distribution of target ISP.

We utilized DiT [10] architecture with AdaLN and cross-attention to model the vector field. Time embeddings and style embeddings are separately conditioned. Time embeddings are conditioned with AdaLN layer while, learnable embeddings for each ISP were injected to model through cross-attention. This architecture allows the model to attend to the current state of flow and modulate the style according to the target camera's specific signature, effectively rendering the style onto the existing semantic structure. We trained and evaluated this model on multi-ISP dataset where each scene was captured with multiple camera models. Our results demonstrate smooth style transfer across diverse ISP pairs while maintaining structure and details.

Our contributions are summarized as follows:

**1) Latent Space ISP Transfer**: We formulate ISP style transfer in latent space of frozen Stable Diffusion VAE [12], avoiding high costs of pixel-space training and naturally aligning with existing latent-based generative model, enabling broader reuse and integration.

**2) Flow-Matching Formulation for Style Transfer**: we adapted flow-matching algorithm [7] from typical use in noise-to-data generation to deterministic image-to-image task, learning the continuous path from source ISP latents to target ISP latents.

**3) Disentangled DiT Architecture**: We propose a disentangled DiT-based transfer network [10] that injects temporal information and style information separately, leading to a modular and effective conditioning scheme for diverse ISP transfer.

# 2. Related Work

## 2.1. Neural Camera ISP and RAW Processing

Conventional camera ISP pipelines are composed of series of hand-designed modules. To model these pipelines with learning-based approaches, prior works studied end-to-end photo enhancement and neural ISP replacements. DPED [3] learns a mapping from mobile photos to DSLR-like outputs, while HDRNet [1] predicts fast, locally affine enhancement in a bilateral grid for real-time processing. For RAW-to-RGB ISP modeling, PyNET [4] proposes a multi-scale CNN to replace mobile ISP with a single deep model. More recently, Rawformer [11] formulates unpaired RAW-to-RAW translation to enable reusing pretrained neural ISPs across unseen cameras. While effective, these methods typically operate in pixel space or RAW space, which can be computationally expensive and may require careful dataset construction for robust cross-device transfer.

## 2.2. Reference-Based and Photorealistic Style Transfer

General image style transfer provides relevant baselines for transferring appearance while preserving content. AdaIN [2] aligns feature statistics for arbitrary style transfer, and WCT-based methods explicitly match feature covariance for universal stylization. WCT2 [16] improves photorealism by incorporating wavelet transforms to better preserve high-frequency structures. Neural Preset [5] targets high resolution color style transfer via deterministic neural color mapping and efficient style switching. Although these approaches can reproduce global color and tone characteristics, they are not explicitly designed to capture ISP specific rendering behaviors.

## 2.3. Latent Diffusion Models

Diffusion models have achieved strong results in image synthesis and editing; latent diffusion reduces the cost of pixel-space diffusion by operating in the latent space of a pre-

trained VAE [12]. Diffusion Transformers (DiT) replace U-Net backbones with transformer blocks over latent tokens, offering scalable capacity and flexible conditioning [10]. For reference-image conditioning in pretrained latent diffusion models, IP-Adapter [15] injects image prompt features via lightweight adapters while keeping the base diffusion model frozen. However, ISP transfer is often closer to deterministic distribution transport between two camera renderings of the same scene, where unnecessary stochasticity or heavy iterative sampling can be inefficient.

### 2.4. Flow Matching and Rectified Flow

Flow Matching [7] and Rectified Flow [8] learn continuous-time vector fields (ODEs) to transport between distributions. These frameworks are attractive for deterministic mapping of an image-to-image translation settings. Our work builds on these directions by formulating cross-ISP translation as conditional transport in the latent space of a frozen Stable Diffusion VAE, parameterized by a DiT-style backbone with explicit temporal and ISP style conditioning.

## 3. Method

### 3.1. Overview

FlowISP achieves ISP style transfer not in pixel space, but in the frozen Stable Diffusion VAE latent space. At a high level, the Stable Diffusion VAE is kept frozen to utilize its robust perceptual compression without an additional training. We only design a latent mapping model from source image latent to target image latent, conditioning on target ISP embeddings. We formulate the translation as a Conditional Flow Matching (CFM) problem and utilize a disentangled DiT architecture to model the continuous transition from the source latent distribution to the target latent distribution. Crucially, we train this DiT from scratch due to the fundamental difference in input distributions. While pre-trained diffusion models are optimized to map from a noise distribution to data, our framework learns a trajectory between two structured image latent distributions, rendering pre-trained weights unsuitable. The overall architecture is illustrated in Figure 2.

### 3.2. Latent Space Formulation

We utilized the variational autoencoder from Stable Diffusion [12], consisting of an encoder $\mathcal{E}_{\text{SD}}$ and a decoder $\mathcal{D}_{\text{SD}}$. In our framework, both encoder and decoder are frozen; we never modify VAE parameters. Given a paired dataset of the same scene and different ISP images $(x_{\text{src}}, x_{\text{tgt}})$, we first extract their latent representations:

$$z_{\text{src}} = \mathcal{E}_{\text{SD}}(x_{\text{src}}), \qquad z_{\text{tgt}} = \mathcal{E}_{\text{SD}}(x_{\text{tgt}}). \tag{1}$$

Our core objective is to train a model that transports $z_{\text{src}}$ to




Figure 2. FlowISP Architecture.

$z_{\text{tgt}}$ given a target ISP style embedding $s_{\text{ISP}}$:

$$\hat{z}_{\text{tgt}} = f_\theta(z_{\text{src}}, s_{\text{ISP}}, t). \tag{2}$$

The final stylized image is obtained by decoding predicted latent by the frozen decoder:

$$\hat{x}_{tgt} = \mathcal{D}_{\text{SD}}(\hat{z}_{tgt}). \tag{3}$$

This latent-space formulation enables a significant reduction in computational cost of mapping while benefiting from the Stable Diffusion VAE robust perceptual compression.

### 3.3. ISP Translation via Flow Matching

Flow matching has been mainly explored in generative modeling from simple priors to data distributions; it has been rarely explored in deterministic image-to-image style transfer between structurally aligned domains. In this work, we adopt flow matching for ISP transfer by setting the flow from a source ISP latent to a target ISP latent. Unlike standard diffusion steps that require corrupting data with a noise process, we define a conditional probability path that continuously deforms the source latent into the target latent over a time horizon. We adopt the optimal-transport conditional flow, which is the straight-line interpolation between the domains:

$$z_t = (1-t)z_{\text{src}} + t z_{\text{tgt}}. \tag{4}$$

The ground-truth vector field from source to target can be induced by the conditional probability path:

$$u_t(z_t \mid z_{\text{src}}, z_{\text{tgt}}) = \frac{d}{dt} z_t = z_{\text{tgt}} - z_{\text{src}}. \tag{5}$$

During training, we train a neural network $v_\theta(z_t, t, c)$ to regress the vector field. The model learns to move from source latents to target latents by minimizing the difference between the predicted velocity and the ground truth. During inference, we transfer the source latent by solving the ordinary differential equation:

$$\hat{z}_{\text{tgt}} = z_{\text{src}} + \int_0^1 v_\theta(z_t, t, c_{\text{style}})\, dt. \tag{6}$$

This formulation allows a deterministic and efficient translation, avoiding the stochasticity of noise-based diffusion.

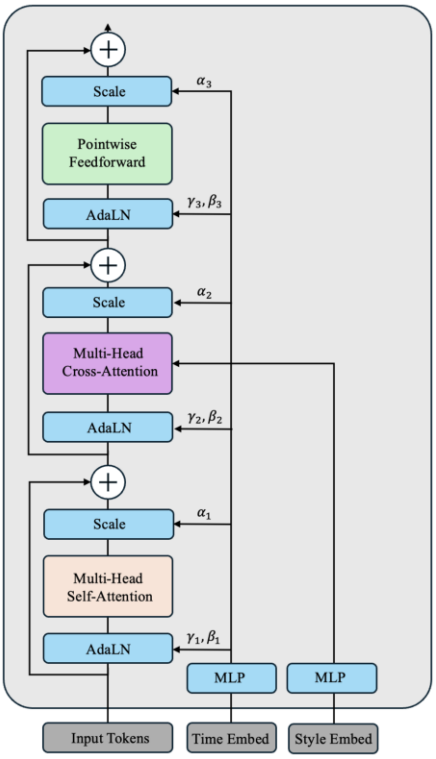



Figure 3. Disentangled DiT-Block. Input latent tokens pass through a standard Transformer stack consisting of self-attention, cross-attention, and feed-forward sublayers, each wrapped by residual connections. The time embedding and style embedding is injected separately.

### 3.4. Disentangled DiT Architecture

To parameterize $v_\theta$, we employ a Diffusion Transformer (DiT) architecture as a backbone. Based on this backbone, we introduce a specialized conditioning strategy to disentangle temporal information for flow from static stylistic information for specific ISP. Figure 3 shows the details of a disentangled DiT block.

**Tokenization and Embeddings.** The input latent $z_{src}$ is reshaped into a sequence of tokens by treating each pixel in the latent feature map as a single token, followed by a linear projection.. Then we added a 2D learnable positional embedding to conserve relative spatial information.

**Disentangled Conditioning Strategy.** Our model explicitly decouples temporal and style conditioning, ensuring precise control over both generative trajectory and aesthetic rendering.

**1) Temporal Conditioning via AdaLN.** Continuous time variable embeddings are exclusively fed by Adaptive Layer Normalization (AdaLN), as the timestep is a global property that dictates the progress of the interpolation. Time embeddings pass through an MLP to regress scaling and shift factors, which normalize each layer:

$$\mathrm{AdaLN}(h, t_{\mathrm{emb}}) = (1+\gamma(t_{\mathrm{emb}}))\cdot\mathrm{LayerNorm}(h)+\beta(t_{\mathrm{emb}}). \tag{7}$$

Furthermore, consistent with AdaLN-Zero [10] design, we employ a **learnable gating scalar** $\boldsymbol{\alpha}$ after each residual branch. This scalar is initialized to zero, ensuring that each block operates as an identity function at the start of training, which significantly stabilizes the learning process.

**2) ISP Conditioning via Cross-Attention.** We hypothesize that ISP effects require style modulation of image content and can be quantified in absolute vectors. Therefore, each ISP style is mapped to a learnable style embedding obtained from an embedding table indexed by an ISP id. These embeddings are converted to style tokens and injected as keys and values to cross-attention layers, while latent features act as queries. This design allows the model to transfer style based on specific spatial features without interfering with the temporal modulation path and provides broader extensibility.

### 3.5. Training Objective

The model is trained using Conditional Flow Matching loss. Given a sampled $t \sim \mathcal{U}(0, 1)$ and paired latents $(z_{\mathrm{src}}, z_{\mathrm{tgt}})$, our objective is to minimize the mean squared error between the predicted velocity and the straight-line vector field:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t, z_{\mathrm{src}}, z_{\mathrm{tgt}}} \left[ \left\| v_\theta(z_t, t, c_{\mathrm{style}}) - (z_{\mathrm{tgt}} - z_{\mathrm{src}}) \right\|^2 \right]. \tag{8}$$

## 4. Experiments

### 4.1. Experimental Setup

We evaluate our method on two datasets that consist of same scene images captured by multiple camera models.

**FiveCam Dataset [6]:** This dataset consists of aligned scenes captured by five distinct camera models – Apple iPhone 14 Pro Max, Google Pixel 6 Pro, Huawei P40, Samsung Galaxy S20, Xiaomi Mi 12.

**HKUST ISP Dataset [14]:** This dataset consists of aligned scenes captured by three distinct camera models – Apple iPhone 6S, Xiaomi Mi 8, Nikon Z6.

Although each scene is captured simultaneously, the images are not perfectly aligned due to small viewpoint changes, lens distortions and cropping. Since our model relies on structurally aligned image pairs, we first perform a geometric alignment step to ensure that structure correspondence across the same scene content. For each scene, we choose one ISP image as the reference and align all other ISP images to this reference using a feature-based method built on LoFTR [13] and homography estimation. Images

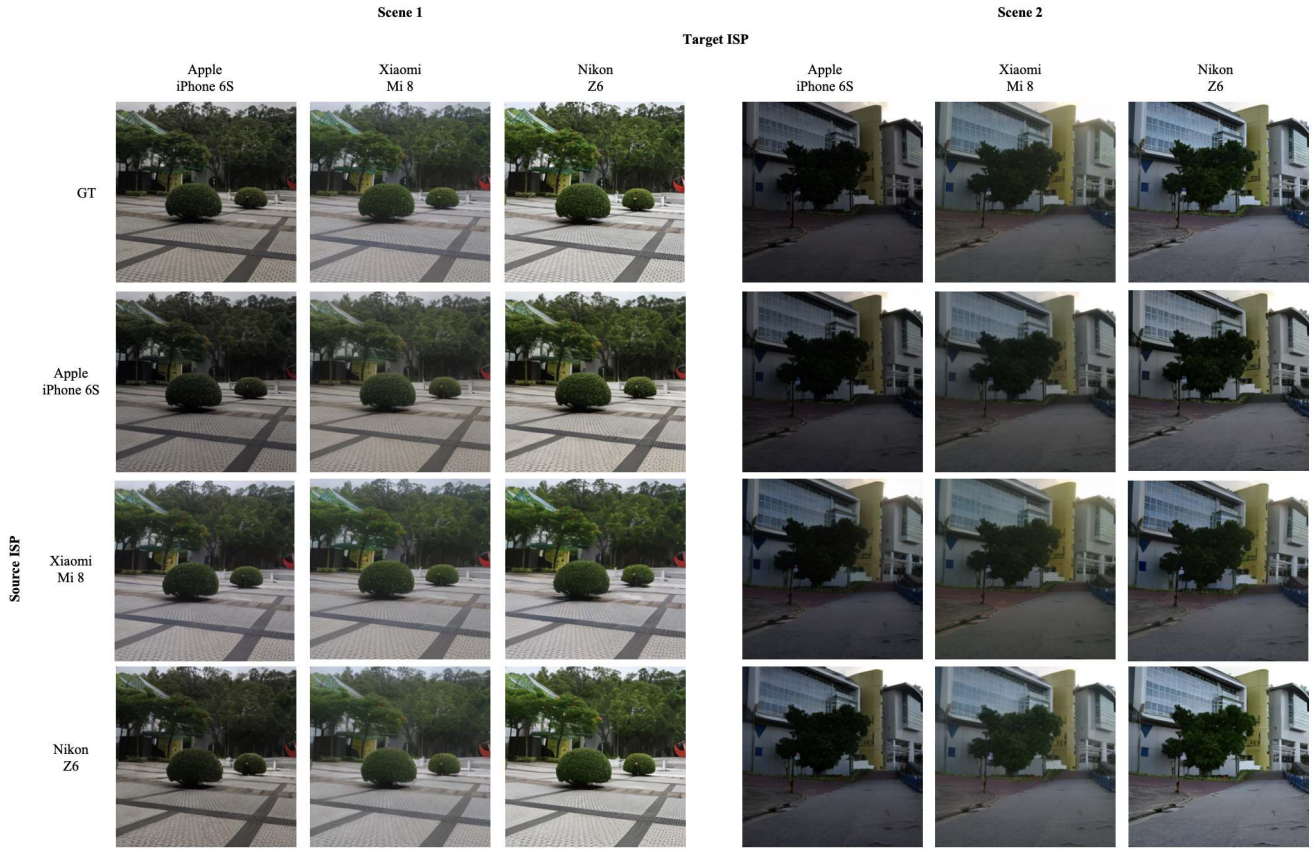



Figure 4. Cross-Device ISP Transfer Results (HKUST ISP Dataset). Visualization of the transfer capabilities of FlowISP across three distinct camera domains: Apple iPhone 6S, Xiaomi Mi 8, and Nikon Z6. Columns represent the Target ISP device, and Rows represent the Source input domain.

with too few LoFTR matches or invalid homography estimates are discarded. The final dataset consists of a total 3335 sRGB images with 896 distinct scenes processed under 8 different ISPs

### 4.2. Implementation Details

We utilize the pretrained VAE from Stable Diffusion [12], which remains frozen during training. For latent-space mapping, we adopt a disentangled DiT-style transformer with a model dimension of 384, 12 layers, and 6 attention heads, which is trained from scratch on our multi-ISP dataset. The model was trained with AdamW [9] with learning rate of 0.0001 with cosine annealing scheduler with 0.05 warmup ratio, using three NVIDIA A5000 GPUs. For data augmentation, we employed a random resized crop with a scale range of [0.8, 1.0] to generate 256×256 images, followed by a random horizontal flip.

### 4.3. Qualitative Results

Figure 4 presents qualitative results of ISP style transfer across three devices - Apple iPhone 6S, Xiaomi Mi 8, and Nikon Z6 - on the HKUST ISP dataset. Columns denote the target ISP devices, while rows denote the source ISP devices. The top row is the ground truth image captured by each target ISP device, while the remaining rows show images transferred by FlowISP. Our method produces outputs that closely match the target ISP characteristics. In particular, the transferred images successfully exhibit device-specific exposure and dynamic range behaviors. As observed in Figure 4, images from Apple iPhone 6S exhibit an overall darker rendering with a more conservative exposure strategy. In contrast, the Xiaomi Mi 8 tends to produce an overall brighter appearance with lifted shadows. Meanwhile, the Nikon Z6 demonstrates a more HDR-like rendition, preserving both highlight and shadow details with enhanced local contrast typical of large-sensor optics. Across all source to target translations, FlowISP consistently reproduces these target-specific signatures, indicating an effective transfer of the ISP's non-linear tone mapping curves while strictly maintaining the semantic content of the scene.

Figure 5 presents qualitative results of ISP style transfer across four smartphone cameras — Apple iPhone 14

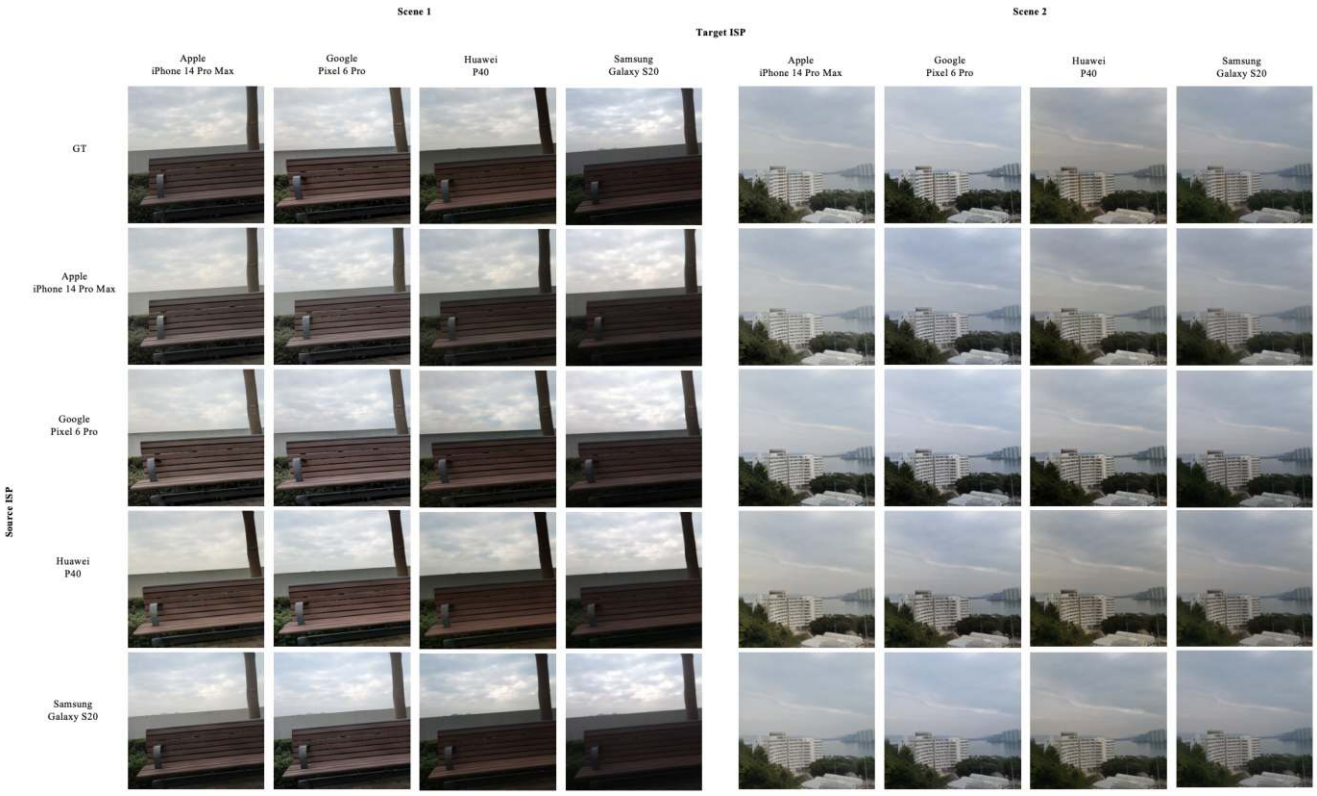



Figure 5. Cross-Device ISP Transfer Results (Fivecam Dataset). Visualization of the transfer capabilities of FlowISP across four distinct camera domains: Apple iPhone 14 Pro Max, Google Pixel 6 Pro, Huawei P40, and Samsung S20. Columns represent the Target ISP device, and Rows represent the Source input domain.

Pro Max, Google Pixel 6 Pro, Huawei P40, and Samsung Galaxy S20 — on the FiveCam dataset, evaluated on two distinct outdoor scenes. Columns denote target ISP devices and rows denote source ISP devices, with the top row showing ground truth images captured by each target device. The transferred images closely match the target ISP characteristics despite substantial stylistic differences among the source devices. In particular, the Apple iPhone 14 Pro Max produces a naturalistic rendering with balanced exposure and moderate color saturation. The Google Pixel 6 Pro tends toward a cooler, more neutral tone with restrained contrast, consistent with its computational photography pipeline. The Huawei P40 exhibits a punchier, higher-contrast rendering with more aggressive shadow compression, while the Samsung Galaxy S20 applies a darker overall exposure with cooler, more muted color rendering, compressing highlights. Across all source-to-target translations, FlowISP reliably reproduces these device-specific tonal and chromatic signatures, demonstrating the generalizability of our approach beyond the HKUST ISP dataset to a broader range of modern smartphone ISPs.

### 4.4. Comparison with Previous Models

As shown in Figure 6, FlowISP shows strong ISP style transfer compared to other reference image-based style transfer models. Baseline models even failed to transfer ISP style when ground truth image was given as a reference like as shown in the top row

The baselines exhibit a common failure mode for ISP transfer. In many cases, WCT2 and Neural Preset produced outputs that are overly biased toward the reference image's global color cast, producing unnatural tints. For IP-Adapter, the output remains too close to the source image. FlowISP, on the other hand, produces results closely align with target ISP image across both examples. Our method reflects the target's global exposure or tone characteristics while maintaining the scene geometry perfectly.

### 4.5. Trajectory Analysis: Visualizing the Flow Path

To validate that our model effectively learns the continuous vector field between ISP domains, rather than merely memorizing source-target pairs, we visualize the intermediate states of the inference process. Figure 7 illustrates a trajectory from Xiaomi Mi 8 to Apple iPhone 6S. We decode the intermediate outputs at $t \in \{0, 0.25, 0.5, 0.75, 1\}$.

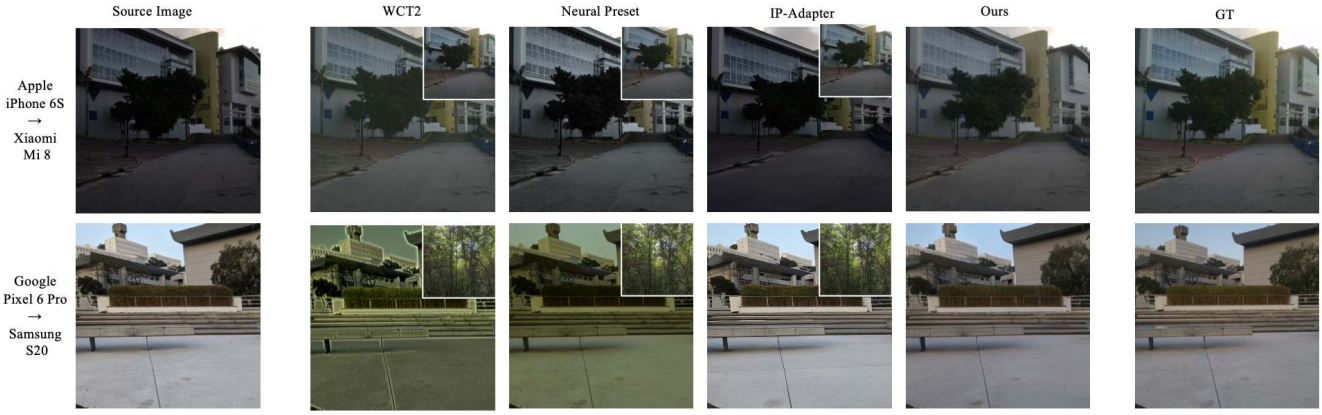



Figure 6. Qualitative comparison with baseline models. Comparison of our method against WCT2, Neural Preset, and IP-Adapter. Top-right corner image is reference style image used for transformation.




Figure 7. Visualization of the Flow Matching Trajectory. The figure shows the continuous style transition from Xiaomi Mi 8 to Apple iPhone 6S at different time steps t. The inset in the top-right corner of the final frame (t=1) displays the ground truth target image.

Two key properties are shown in the trajectory:
**1) Smooth Interpolation.** The global appearance shifts continuously, effectively translating specific ISP signatures. The sequence transitions from the cooler color temperature and flatter tone characteristic of the Xiaomi ISP to the warmer, higher-contrast aesthetic typical of the Apple ISP, verifying that the model follows the optimal transport path.
**2) Structural Invariance.** The geometric structure remains strictly unchanged throughout the transition. High-frequency details—such as the building tiles and fence railings—exhibit no spatial distortion or hallucination. This confirms that the flow operates exclusively on stylistic attributes, leaving the underlying spatial content intact.

## 5. Discussion and Future Work

While FlowISP demonstrates effective ISP style transfer in latent space, two core limitations remain.
**Pixel-wise Alignment Dependency.** Our framework fundamentally relies on structurally aligned image pairs for training. In practice, even simultaneously captured multi-camera datasets suffer from viewpoint discrepancies, lens distortions, and subtle cropping differences. To mitigate this, we applied a LoFTR-based homography alignment as a preprocessing step, but this introduces its own failure cases — scenes with insufficient feature matches or invalid homography estimates were discarded entirely, reducing the effective dataset size. Furthermore, homography-based alignment assumes a planar scene geometry, which breaks down for scenes with significant depth variation or foreground-background parallax. This alignment bottleneck limits both scalability and the diversity of scenes the model can be trained on, and imposes a hard constraint on real-world deployment where paired aligned data is rarely available.
**Generalization to Unseen Cameras.** The current ISP conditioning relies on a learnable embedding table indexed by discrete ISP IDs seen during training. As a consequence, the model cannot transfer to a new, unseen camera without retraining or fine-tuning on new paired data. This is a significant practical constraint, as the value of ISP transfer is greatest precisely when dealing with novel or rare devices.

For future work, we aim to address both limitations. To reduce alignment dependency, we plan to explore weakly

supervised or unpaired training strategies, such as using perceptual consistency losses or cycle-consistency constraints that do not require perfect geometric correspondence. To improve generalization, we intend to replace the discrete ISP ID embedding with a continuous style encoder that extracts ISP characteristics directly from a few reference images of an unseen camera, enabling zero-shot or few-shot transfer to novel devices.

## 6. Conclusion

In this work, we presented FlowISP, a latent-space framework for ISP style transfer. By operating in the frozen Stable Diffusion VAE latent space and formulating the transfer as a Conditional Flow Matching problem, our method avoids the high computational cost of pixel-space training while producing deterministic and structurally consistent style transfer. Furthermore, disentangled DiT architecture demonstrates the effectiveness of separation of temporal and aesthetic condition. Experiments on multi-device ISP datasets confirm that FlowISP effectively reproduces target device rendering characteristics, outperforming existing reference-based style transfer baselines. We hope this work opens a new direction for efficient, latent-space ISP modeling and lays a foundation for future generalization to arbitrary unseen cameras.

## References


[1] Michael Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM TOG*, 2017.

[2] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[3] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017.

[4] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *CVPRW*, 2020.

[5] Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson W. H. Lau. Neural preset for color style transfer. In *CVPR*, 2023.

[6] Lingen Li, Mingde Yao, Xingyu Meng, Muquan Yu, Tianfan Xue, and Jinwei Gu. Uni-isp: Toward unifying the learning of isps from multiple mobile cameras. *IEEE TIP*, 2025.

[7] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

[8] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[10] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[11] Georgy Perevozchikov, Nancy Mehta, Mahmoud Afifi, and Radu Timofte. Rawformer: Unpaired raw-to-raw translation for learnable camera isps. In *ECCV*, 2024.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[13] Jiaming Sun, Zehao Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021.

[14] Yazhou Xing, Changlin Li, Xuaner Zhang, and Qifeng Chen. A well-aligned dataset for learning image signal processing on smartphones from a high-end camera. In *ACM SIGGRAPH 2022 Posters*, 2022.

[15] 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.

[16] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *ICCV*, 2019.