Towards In-Context Tone Style Transfer with a Large-Scale Triplet Dataset

Yuhai Deng[1,2*], Huimin She[4*], Wei Shen[4†], Meng Li[4], Ruoxi Wu[4], Lunxi Yuan[4], and Xiang Li[2,1,3✉]

[1] VCIP, School of Computer Science, Nankai University
[2] Nankai International Advanced Research Institute (Shenzhen Futian)
[3] AAIS, Nankai University
[4] OPPO AI Center, OPPO Inc.
yhdeng.me@gmail.com, xiang.li.implus@nankai.edu.cn

Abstract. Tone style transfer for photo retouching aims to adapt the stylistic tone of the reference image to a given content image. However, the lack of high-quality large-scale triplet datasets with stylized ground truth forces existing methods to rely on self-supervised or proxy objectives, which limits model capability. To mitigate this gap, we design a data construction pipeline to build TST100K, a large-scale dataset of 100,000 content-reference-stylized triplets. At the core of this pipeline, we train a tone style scorer to ensure strict stylistic consistency for each triplet. In addition, existing methods typically extract content and reference features independently and then fuse them in a decoder, which may cause semantic loss and lead to inappropriate color transfer and degraded visual aesthetics. Instead, we propose ICTone, a diffusion-based framework that performs tone transfer in an in-context manner by jointly conditioning on both images, leveraging the semantic priors of generative models for semantic-aware transfer. Reward feedback learning using the tone style scorer is further incorporated to improve stylistic fidelity and visual quality. Experiments demonstrate the effectiveness of TST100K, and ICTone achieves state-of-the-art performance on both quantitative metrics and human evaluations. The project page is available online.

Keywords: Tone style transfer · Photo retouching · Tone style transfer benchmark

1 Introduction

Photo retouching and color manipulation are fundamental to computational photography and image editing, allowing images to be adapted to diverse photographic styles. Among these techniques, tone style transfer [18,33] aims to adapt the photographic appearance of a content image according to a reference image. As illustrated in Fig. 1, it transfers tone-related attributes such as color distribution, luminance, contrast, and saturation while preserving the semantic content

*: Equal Contribution. †: Project Leader. ✉ : Corresponding Author.

Fig. 1: Showcases of our method performing tone style transfer across diverse scenarios.

and structural integrity of the content image. Compared with traditional color transfer, which mainly matches global color statistics, and artistic style transfer, which often alters textures and structures, tone style transfer operates at the level of photographic aesthetics and requires semantically aware adaptation across different image regions.

The absence of ground-truth stylized targets for content-reference pairs fundamentally limits the learning capability of existing tone style transfer methods [33]. Consequently, these methods must rely on self-supervised training [18] or proxy style losses [15, 26, 46, 54], which provide only indirect supervision for learning reference-conditioned tone transfer. For example, Neural Preset [18] constructs self-supervised training pairs by applying different presets to the same content image, where one preset-retouched image is used as the reference and the other serves as the pseudo target. Nevertheless, as shown in Tab. 1, training Neural Preset with ground-truth content–reference–stylized triplets significantly improves its performance, demonstrating that self-supervision inherently constrains model capability.

To address the lack of high-quality paired data, we propose a data construction pipeline to build a large-scale dataset of content-reference-stylized triplets. The key challenge in constructing such a dataset is the absence of a reliable metric for tone style similarity. Without such a metric, dataset construction faces an inherent trade-off between quality and scalability: manual retouching produces accurate pairs but is prohibitively expensive, while applying predefined presets is scalable but inevitably introduces noisy and perceptually inconsistent triplets. Specifically, the same preset may produce unreliable tonal correspondences when applied to images with diverse content. To overcome this dilemma, we develop a dedicated tone style scorer through a two-stage training paradigm. Specifically, the scorer is first trained with weakly supervised contrastive learning [19] on

noisy preset-generated pairs to learn general tone style similarity, and subsequently fine-tuned on human-ranked pairs for improved perceptual alignment. By leveraging this robust scorer, we can effectively ensure strict stylistic consistency between the reference and stylized images for each preset-generated triplet. In addition, an aesthetic model [40] is also incorporated to filter out low-quality images and improve the overall visual quality of the dataset.

While large-scale datasets enhance model performance, existing methods still exhibit inappropriate color transfer. This issue fundamentally arises from their architectural design, as most models utilize independent encoders to extract features from the content and reference images separately before fusing them within a decoder. This process inevitably leads to semantic loss, severely restricting the overall semantic perception capability of the model. For instance, as shown in Fig. 6, the retrained CAP-VSTNet incorrectly maps object colors onto human faces, resulting in unnatural skin coloration. To resolve such semantic misalignments, we formulate tone style transfer as an in-context generation task [24, 49, 57] and introduce ICTone. As a diffusion-based framework, ICTone treats the content and reference images as a joint contextual input, leveraging the strong semantic priors of large-scale generative models [6, 9, 10, 39, 47] to achieve accurate, structure-aware tone style transfer. To further improve stylistic fidelity and aesthetic appeal, we incorporate reward feedback learning [52] using the trained tone style scorer as the reward signal.

Our main contributions are as follows:

- We construct **TST100K**, the first large-scale dataset of content-reference-stylized triplets, providing reliable ground truth for training robust tone style transfer models. Furthermore, we introduce **TST2K**, a carefully curated high-quality benchmark of 2,000 triplets that enables precise evaluation.
- We develop a **tone style scorer** to solve a long-standing challenge in the field: how to effectively measure the stylistic tone similarity between two images. This scorer serves a dual role by ensuring strict stylistic and perceptual consistency during data construction, while providing a reward signal during model training to align outputs with perceptual style fidelity.
- We propose **ICTone**, a diffusion-based framework that performs tone transfer as an in-context generation task and incorporates reward feedback learning to enhance stylistic alignment and aesthetic quality. Extensive quantitative and human evaluations demonstrate that ICTone achieves state-of-the-art performance, successfully validating the effectiveness of both our framework and the TST100K dataset.

2 Related Work

2.1 Datasets for Photo Retouching

Existing photo retouching datasets mainly focus on global tone and color enhancement. Representative datasets include the MIT-Adobe FiveK dataset [5] and PPR10K [27], which provide large-scale input–output pairs for supervised

enhancement. However, these datasets are designed for single-image enhancement under fixed expert styles, without modeling content–style correspondences or reference-driven transfer. Recent large-scale triplet datasets for artistic style transfer, such as OmniStyle [44], provide over one million content–style–stylized triplets across diverse artistic styles. However, they focus on artistic transformations that allow texture and structural changes, making them unsuitable for photorealistic tone style transfer, which requires strict preservation of scene geometry and pixel-level details. In contrast, datasets specifically designed for photorealistic style transfer remain limited. DPST [31] provides photograph pairs of content and style images but lacks ground-truth stylized outputs. In contrast, PST50 [13] includes ground-truth results, but contains only 50 pairs, limiting diversity and benchmarking scale for tone style transfer models. To address these limitations, we introduce two large-scale tone style transfer datasets: TST100K, a high-quality dataset that enables supervised training, and TST2K, a carefully curated benchmark for evaluation.

2.2 Style Transfer

Tone style transfer originates from the broader field of image style transfer, which aims to apply the visual style of a reference image to a content image. Style transfer is commonly divided into artistic style transfer [11,30] and photorealistic style transfer [54]. Early artistic style transfer methods [3, 11, 17, 25] represent style using Gram matrices and perform stylization through iterative optimization, successfully reproducing artistic textures but often introducing artifacts and structural distortions. Early approaches [35, 36, 38, 50] to photorealistic style transfer focus on matching low-level color statistics between images, aligning color histograms or per-channel means and variances between the content and reference images in carefully designed color spaces [12, 34], achieving basic color and tone alignment. More recent methods adopt learning-based frameworks to improve transfer quality while preserving structural fidelity [8, 18, 21, 29, 48, 54]. For instance, PhotoWCT [2, 26, 46] performs whitening–coloring transforms followed by edge-preserving smoothing to maintain spatial consistency. More recently, SA-LUT [13] employs spatially adaptive 4D look-up tables to achieve real-time, high-quality photorealistic tone transfer with improved spatial fidelity. However, these methods often exhibit limited semantic understanding, leading to suboptimal performance in cross-scene transfer. To address this limitation, we propose ICTone, which leverages the rich priors of generative diffusion models to enable more accurate and robust semantic-aware transfer.

3 Method

In this section, we first present our dataset construction, including image collection, tone style scorer training, and the overall pipeline, and then introduce ICTone, highlighting its in-context formulation and reward feedback learning.

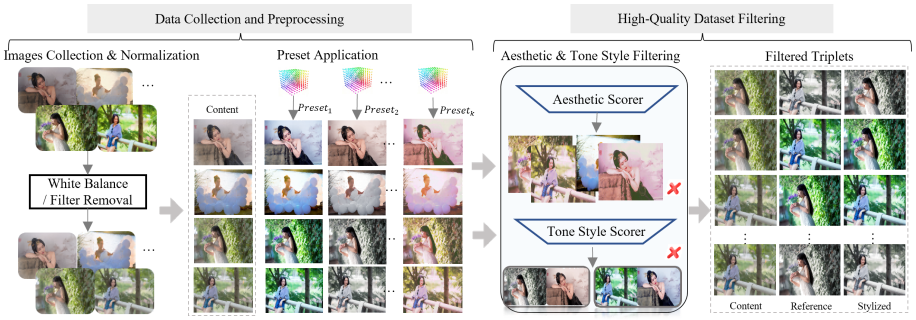

Fig. 2: Overview of the dataset construction pipeline. Left: data collection and preprocessing, where images undergo white balance correction and filter removal, followed by applying diverse tone presets to generate stylized candidates. Right: high-quality filtering using an aesthetic scorer and a tone style scorer to select pairs with high stylistic consistency and visual quality, forming the final content-reference-stylized triplets.

3.1 Dataset Construction

Existing tone style transfer methods often rely on self-supervised training due to the lack of high-quality paired data, limiting their ability to learn accurate tone styles. Paired datasets can be created either by manual retouching, which is accurate but costly, or by applying presets, which is scalable but often inconsistent across diverse content. To address this, we train a tone style scorer to evaluate tone similarity and use it to select high-quality preset-generated pairs, which is the core of our dataset construction pipeline. As illustrated in Fig. 2, our pipeline includes image collection, normalization, and high-quality filtering, producing a large-scale set of content–reference–stylized triplets with consistent tone and high visual quality.

3.1.1 Data Collection and Preprocessing

The training data for the tone style scorer is collected from multiple public datasets, including MIT-Adobe FiveK [5], PPR10K [27], COCO [28], Food-101 [4], and Landscape HQ [41]. These datasets cover a diverse range of visual domains, such as portraits, landscapes, and urban scenes. For constructing TST100K, we primarily utilize images from PPR10K and MIT-Adobe FiveK due to their high image quality. Presets are collected from Adobe Lightroom and professional photography communities, where they are pre-categorized into scenarios such as portrait, landscape, night, lifestyle, and food. For each category, redundant presets are removed using LAB histogram filtering and manual inspection, resulting in approximately 3,000 distinct presets. Before applying presets, all images are normalized to mitigate pre-existing tonal biases. Specifically, automated white balance correction [1] and filter removal [53] are performed. We then apply curated presets to normalized images, creating a substantial pool of different images sharing the same tone style.

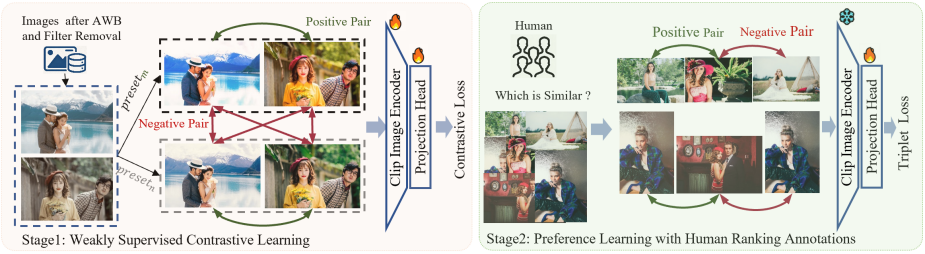

Fig. 3: Overview of the two-stage tone style scorer training pipeline. It combines weakly supervised contrastive learning and preference learning for tone style alignment.

3.1.2 Tone Scorer Training

To ensure tone consistency during dataset construction, we train a tone style scorer to evaluate the tonal similarity between image pairs. We adopt a CLIP architecture [37] for the tone style scorer, leveraging its strong capability in learning discriminative visual embeddings [22, 55, 56]. The image encoder follows the ViT-B/16 backbone, and the projection head maps image features into a normalized tone embedding space. These embeddings are later used to measure tone style similarity when constructing the paired dataset. As illustrated in Fig. 3, we train the tone scorer in two stages to achieve reliable tone similarity estimation. In the first stage, we employ weakly-supervised contrastive learning using preset-generated data, where image pairs produced by the same preset are treated as positive samples, while those from different presets are treated as negative samples. This stage enables the model to capture coarse tone representations from large-scale data. In the second stage, we fine-tune the model with a small set of human-annotated ranking data to refine its perception of tonal similarity and obtain the final tone style scorer for dataset construction.

Weakly-supervised contrastive learning using presets. In this stage, images processed with the same preset are regarded as positives, while those processed with different presets serve as negatives. This formulation enables the model to learn tone similarity based on preset labels without requiring manual annotations. The training objective is formulated as a supervised contrastive loss [19], defined as:

$$\mathcal{L}_{Con} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\mathbf{z}_i\cdot\mathbf{z}_p/\tau)}{\sum_{a\neq i}\exp(\mathbf{z}_i\cdot\mathbf{z}_a/\tau)}, \tag{1}$$

where N denotes the batch size, $P(i)$ is the set of positives sharing the same preset label with sample i, $\mathbf{z}_i \in \mathbb{R}^d$ is the normalized feature embedding of sample i, and τ is a temperature parameter. This formulation guides the model to capture discriminative tone representations that encode global color distributions and illumination patterns. To ensure generalization across diverse tone variations, we train the scorer on the previously collected dataset (Section 3.1.1). Further-

more, we apply image blurring as a data augmentation strategy to encourage the model to capture tone-related characteristics.

Preference learning with human-ranking annotations. While contrastive learning on preset-generated pairs offers discriminative supervision, it may not fully match human perceptual assessments. In practice, a single preset often produces perceptually divergent results across images, leading to noisy positive pairs. We therefore employ multiple discriminators with diverse backbones (VGG, ResNet, and ViT) to identify hard samples with inconsistent predictions, which are then refined by human ranking, resulting in a curated set of 20K human-ranking annotations. Each annotation contains several candidates with similar content but different tone adjustments. Human evaluators rank these candidates based on tone consistency and visual harmony. We adopt a triplet loss to refine the tone style scorer. The loss $\mathcal{L}_{\text{Tri}}$ is defined as follows:

$$\mathcal{L}_{\text{Tri}} = \max\left\{0,\ d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + m\right\}, \tag{2}$$

Here, $d(u, v)$ is the cosine distance, defined as: $d(u, v) = 1 - \frac{\langle u, v\rangle}{\|u\|_2 \, \|v\|_2}$, $\mathbf{z}_a$ denotes the feature embedding of anchor image, $\mathbf{z}_p$ denotes feature embedding of a positive sample ranked higher or perceptually similar, and $\mathbf{z}_n$ denotes feature embedding of a negative sample ranked lower or perceptually dissimilar. The loss encourages the anchor to be closer to the positive than to the negative by at least a margin m. This stage is essential for the scorer to evaluate tone alignment in accordance with human-perceived preferences.

3.1.3 Triplet Dataset Construction

We construct a large-scale dataset of content-reference-transferred triplets, where the content image represents an unedited photo, the reference provides the target tonal appearance, and the transferred image serves as the retouched result.

Reference-stylized pair verification. To ensure high visual quality and reliable tone alignment, we employ a dual-constraint filtering strategy during dataset construction. Specifically, an aesthetic non-degradation constraint and a tone consistency constraint are applied as complementary filtering criteria. For aesthetic quality control, we adopt an aesthetic assessment model [40] to prevent visual degradation introduced by preset-based stylization. Given an original image and its stylized counterpart, we retain only samples whose aesthetic score is no lower than that of the original image prior to preset application. In addition, we introduce a tone style scorer to evaluate tone alignment between the reference and stylized images by measuring cosine similarity in a learned tone embedding space. Reference–stylized pairs with similarity below 0.8 are discarded to enforce tone consistency and ensure reliable training supervision. This dual verification process ensures the stylized image is both visually appealing and tone-consistent with its reference, forming reliable supervision signals for training.

Content diversity augmentation. To improve robustness and better reflect diverse real-world inputs, additional content variations are introduced during training. Specifically, transformations such as color desaturation, exposure

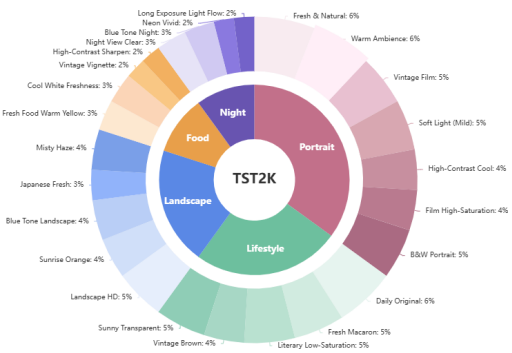

Fig. 4: The distribution of TST2K benchmark.

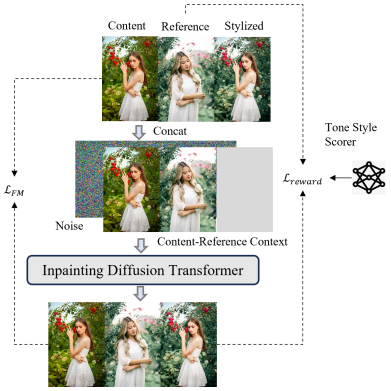

Fig. 5: Overview of the in-context model training pipeline.

reduction, and LUT-based color perturbations are randomly applied to the content image. These operations are performed online and preserve the underlying semantic structure while modifying illumination, contrast, and color distribution. This augmentation strategy expands the appearance diversity of the content domain, enabling the model to maintain stable tone transfer performance across varied input conditions.

Finally, we construct TST100K, a large-scale dataset of 100,000 content-reference-stylized triplets with consistent tone style alignment. Based on TST100K, we further curate a representative benchmark subset termed TST2K. Each triplet is manually reviewed for tone alignment and visual quality. The subset contains 2,000 curated triplets, comprising samples from TST100K and manually retouched image pairs. As illustrated in Fig. 4, TST2K spans diverse real-world scenarios, including portrait, food, landscape, night, and lifestyle scenes, providing a comprehensive evaluation setting.

3.2 In-Context Tone Style Transfer

We formulate tone style transfer as an in-context generation task and propose our ICTone model. As depicted in Fig. 5, it takes both the content and reference images as input and implicitly learns the desired tonal transformation.

3.2.1 In-Context Generation

Our goal is to learn a mapping function that transfers the tone characteristics from a reference image I_r to the content image I_c while preserving its semantic structure. Given a triplet (I_c, I_r, I_t) where I_t denotes the target image with the same tone as the reference image, the model is trained to approximate the conditional distribution $p_\theta(I_t \mid I_c, I_r)$.

Unlike conventional tone transfer methods that explicitly disentangle content and style representations, recent progress in in-context generation [23, 24, 49, 57]

reveals that diffusion models are capable of understanding visual relationships directly from contextual examples. This observation motivates us to formulate tone transfer as a conditional inpainting problem. Specifically, we construct a joint visual context by concatenating the content image and the reference style image (I_c, I_r) along the spatial dimension. A binary mask is then applied to the region corresponding to the stylized output, while the original content and reference regions remain visible. The masked region is initialized as noise and serves as the target area [45]. The diffusion transformer (DiT) is trained to predict and progressively denoise the masked region conditioned on the visible context. The model learns to preserve the structural information from I_c while adapting the tonal characteristics observed in I_r. In this way, tone-consistent stylization emerges implicitly through contextual reasoning, without requiring explicit content-style disentanglement.

3.2.2 Tone Reward Feedback Learning

To further improve tonal fidelity, we introduce a tone reward feedback learning mechanism to constrain the generation process to align outputs with the reference tone distribution. Specifically, we incorporate reward feedback learning inspired by the ReFL framework [52], which optimizes text-to-image synthesis toward reference-conditioned aesthetic alignment. Unlike ReFL, which optimizes text-to-image generation toward reference-conditioned aesthetic alignment, we extend this idea to the content–reference in-context generation setting.

For a triplet $s_i = (I_c^i, I_r^i, I_t^i)$, the generator produces a stylized image $\hat{I}^i = I_\theta(I_c^i, I_r^i)$. A pretrained tone style scorer $\mathcal{M}_{\text{TS-Scorer}}$ then evaluates the alignment between the generated image $\hat{I}^i$ and the reference tone image I_r^i, providing a fine-grained supervision signal. The reward loss is defined as:

$$\mathcal{L}_{\text{tone}} = \mathbb{E}_{s_i \sim \boldsymbol{\mathcal{S}}}\big[\phi\big(\mathcal{M}_{\text{TS-Scorer}}(I_r^i,\, I_\theta(I_c^i, I_r^i))\big)\big], \tag{3}$$

where $\boldsymbol{\mathcal{S}} = \{\boldsymbol{s_i}\}_{i=1}^n$ denotes the set of triplets, ϕ maps scorer outputs to per-sample loss values, and $I_\theta(\cdot)$ is the diffusion model parameterized by θ. This design enforces explicit tone consistency during generation, thereby enhancing both the stability and fidelity of tone style transfer.

3.2.3 Training Objectives

We adopt the flow-matching objective as the primary optimization target [58]. Concretely, the model is trained to predict the target velocity field $\boldsymbol{v}_t$ from a noised latent $\boldsymbol{s}_t$, with the objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\boldsymbol{s}_0, t, \epsilon}\|\boldsymbol{v}_\theta - \boldsymbol{v}_t\|^2, \tag{4}$$

where $v_t = \frac{d\alpha_t}{dt}\,\boldsymbol{s}_0 + \frac{d\sigma_t}{dt}\,\varepsilon$, and $\alpha_t = 1 - t,\ \sigma_t = t$ are continuous-time coefficients with $t \in [0,\, 1]$. $\boldsymbol{v}_\theta$ denotes the predicted velocity. This formulation is consistent with recent flow-based generative paradigms, enabling direct supervision of the continuous transformation dynamics between data and noise distributions.

After the model achieves stable convergence under $\mathcal{L}_{\mathrm{FM}}$, we introduce an additional tone reward feedback loss described in Section 3.2.2 to further encourage tone alignment. The final training objective integrates flow matching with tone reward feedback loss and is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{tone}}\mathcal{L}_{\mathrm{tone}}, \tag{5}$$

where $\lambda_{\mathrm{tone}} = 0$ before training step S and $\lambda_{\mathrm{tone}} = 1$ afterward.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on two datasets: TST2K and PST50 [13]. TST2K is our self-constructed dataset that covers a broader range of real-world scenes and tonal styles, providing a more rigorous benchmark for evaluating transfer quality. PST50 is used as a generalization test, where models are directly evaluated without additional training. It contains 50 paired triplets and is used to assess cross-dataset transfer performance.

Comparison Methods. We compare our method with leading methods from photorealistic style transfer, color transfer, and diffusion-based style transfer, as well as the closed-source image editing model Nano Banana2.

Evaluation Metrics. We evaluate all methods using the following metrics: content preservation (CP), color difference (ΔE), deep color difference (CD), and aesthetic quality (Aes). Among them, ΔE and CD are designed to measure the fidelity and accuracy of tone style transfer. Content Preservation (CP) is measured using the Structural Similarity Index Measure (SSIM) computed on LDC-generated [43] edge maps, evaluating structural similarity between the generated output and the stylized ground truth. Color Difference (ΔE) [32] is a widely adopted international standard for measuring perceptual color differences between the stylized and output images. Deep Color Difference (CD) is a neural network–based metric designed to better align color difference estimation with human perceptual judgments [7]. Aesthetic Quality (Aes) is measured by the mean aesthetic score predicted by an image aesthetic assessment model [40], reflecting the overall visual appeal and tonal harmony of the results.

Implementation Details. The tone style scorer is trained in two stages. For the weakly supervised contrastive learning stage, we train the backbone and the projection layer for 50 epochs with learning rates of 1×10^{-4} and 3×10^{-3}, respectively, and set the temperature parameter in Eq. 1 to $\tau = 0.1$. For the subsequent preference learning stage, the projection layer is fine-tuned for 2 epochs using 20K human-ranked data, with the triplet margin in Eq. 2 set to $m = 0.3$. To better match real-world conditions, our training incorporates degradations such as color desaturation, exposure reduction, blurring, and Gaussian noise. We adopt FLUX.1 Fill as the backbone and fine-tune it on our constructed triplet dataset using the LoRA method [16]. Training is performed on four NVIDIA A100 GPUs with the AdamW optimizer, where the learning rate is set to 1×10^{-4} and the weight decay to 1×10^{-3}. The model is optimized for a total of 50,000 iterations.

Table 1: Quantitative comparisons with state-of-the-art methods on the TST2K and PST50 benchmarks. ICTone achieves superior performance in color difference (CD and ΔE) and aesthetic quality (Aes), while maintaining competitive content preservation (CP). Methods marked with * denote retraining on our triplet dataset, which further improves their performance.

Method	TST2K				PST50			
	CP↑	$\Delta E \downarrow$	CD ↓	Aes↑	CP ↑	$\Delta E \downarrow$	CD ↓	Aes↑
WCT^2 [54]	0.3952	15.13	4.991	0.7136	0.6482	16.34	5.574	0.2381
PhotoNAS [3]	0.6721	19.57	6.183	0.6347	0.6757	32.87	8.566	0.2293
MKL [35]	0.7511	12.67	4.694	0.7363	0.7618	16.12	5.241	0.2742
PhotoWCT [26]	0.5852	21.35	7.021	0.6788	0.6477	20.83	6.487	0.2279
DeepPreset [15]	<u>0.7728</u>	8.205	4.108	0.7567	0.6566	16.90	5.679	0.1147
ModFlows [21]	0.6908	12.43	4.945	0.7500	0.7334	15.52	5.651	0.2774
IPST [29]	0.7364	9.648	4.326	0.7081	0.7275	14.08	5.484	0.2344
RLPixTuner [48]	0.6815	25.61	7.514	0.6354	0.6789	21.51	7.445	0.1946
SA-LUT [13]	0.7082	11.10	4.929	0.6507	<u>0.7697</u>	**11.25**	<u>4.483</u>	0.2048
CAP-VSTNet [46]	0.7013	11.87	4.665	0.7175	0.7384	16.21	5.350	0.2705
CAP-VSTNet* [46]	0.7665	7.359	<u>3.663</u>	0.7801	0.7545	14.45	4.937	0.2779
Neural Preset [18]	0.7480	10.51	4.713	0.7551	0.6502	19.53	6.965	0.1454
Neural Preset* [18]	0.7707	<u>7.886</u>	3.860	0.7808	0.6914	17.81	6.157	0.1714
CSGO [51]	0.3440	19.00	7.166	0.7209	0.3836	26.18	7.747	**0.3607**
Nano Banana2 [14]	0.5968	12.78	4.947	<u>0.7885</u>	0.5901	17.19	5.988	0.2793
ICTone	**0.8644**	**5.776**	**2.634**	**0.7904**	**0.7902**	<u>12.81</u>	**4.448**	<u>0.2820</u>
GT	1.000	0.000	0.000	0.7978	1.000	0.000	0.000	0.3305

Table 2: Comparison of average rankings across different methods.

Method	CAP-VSTNet [46]	Neural Preset [18]	SA-LUT [13]	MKL [35]	ModFlows [21]	IPST [29]	ICTone
Average Ranking ↓	4.35	2.70	4.05	3.60	6.00	5.20	**1.00**

4.2 Experimental Results

4.2.1 Quantitative Evaluation

We conduct a comprehensive quantitative comparison between our ICTone and several SOTA approaches on the TST2K and PST50 benchmarks, where PST50 serves as a direct generalization benchmark without dataset-specific adaptation. The results are presented in Tab. 1.

On TST2K, our ICTone achieves the best performance across all evaluation metrics, demonstrating superior tone alignment with reference styles and content preservation. ICTone achieves the lowest color difference and deep color difference, demonstrating high fidelity in reproducing the target tone style. Note that ICTone achieves the highest aesthetic score 0.7904, closely approaching the ground truth 0.7978, which suggests that the improvements in tone accuracy do not compromise visual appeal. The low CP of Nano Banana2 is mainly caused

by reference foreground leakage, which introduces undesired reference content into the output and degrades content preservation.

We further test on PST50 to evaluate the generalization capability and ICTone achieves the best overall performance. It obtains the highest CP of 0.7902 and the lowest CD of 4.448. Although SA-LUT achieves a lower ΔE on this dataset, its CD and aesthetic score are inferior to those of ICTone. This indicates that ICTone maintains stronger perceptual tone fidelity and overall visual quality under cross-dataset evaluation. The consistent gains across both datasets further demonstrate the robustness and generalization capability of ICTone on data outside the training distribution. Notably, the overall aesthetic scores on PST50 are generally lower, which can be attributed to its relatively darker tone distributions that tend to receive lower scores from the aesthetic assessment model. CSGO is an exception because it is fine-tuned from a style transfer model and tends to produce excessive saturation enhancement. Such unnatural color enhancement can be favored by the aesthetic assessment model, leading to an Aes score even higher than GT, but it does not indicate better retouching fidelity.

4.2.2 Qualitative Evaluation

Fig. 6 presents qualitative comparisons across portraits, urban scenes, and food images. ICTone consistently produces visually superior results with accurate tone style alignment and strong content awareness. In portrait transfer (Fig. 6 (a)), prior methods often incorrectly transfer the warm hue of the railing onto the subject's face, leading to noticeable color contamination. In contrast, our result faithfully reproduces the reference skin tone while preserving natural facial colors. In food image transfer (Fig. 6 (b)), our results closely match reference styles in color hierarchy and contrast, enhancing visual appeal. For urban scenes (Fig. 6 (c)), ICTone effectively captures both chromatic and illumination-level attributes, resulting in outputs with coherent lighting and tone gradation. More qualitative results are provided in the supplementary material.

4.2.3 User Study

To assess the perceptual quality of stylization results, we conduct a user study following the protocols established in [13, 18]. The study compares seven representative methods and uses 20 randomly sampled image pairs from the TST2K dataset. For each method, participants performed pairwise comparisons against the remaining six methods, resulting in 120 comparisons per method.

In each comparison, participants were presented with two stylized outputs and asked to select the one they perceived as superior in overall quality. Based on these selections, we compute the win rate for each method. Tab. 2 summarizes the average ranking across 20 participants.

Our ICTone consistently achieved the top rank across all individual user rankings, reflecting the superior balance our approach achieves among structural fidelity, tone consistency and aesthetic coherence.

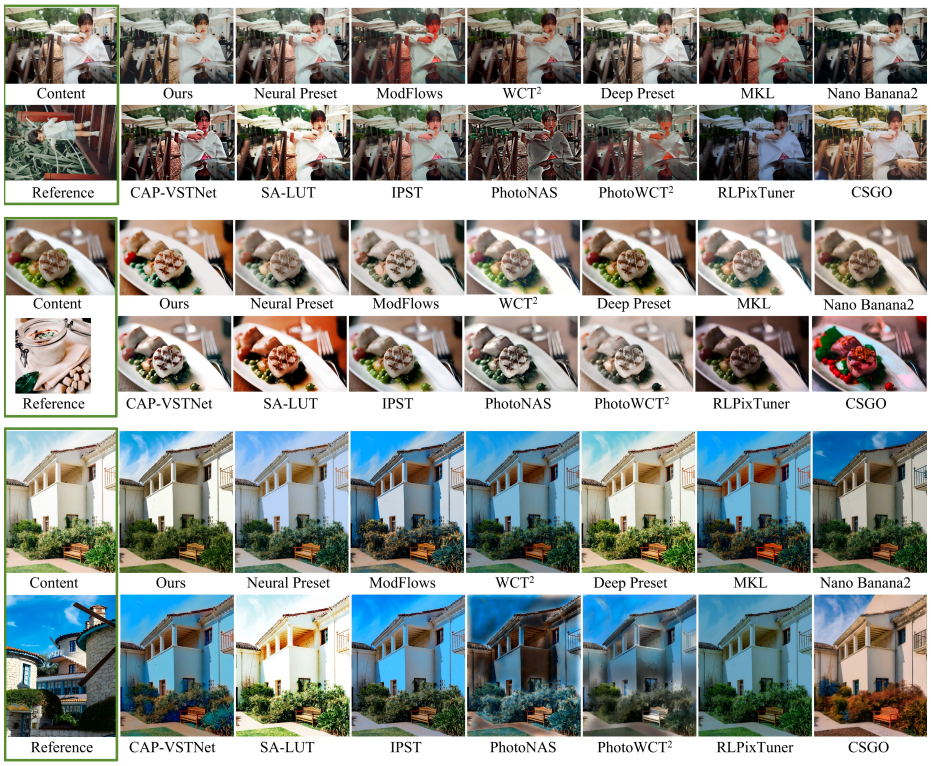

Fig. 6: Qualitative visual comparisons of our method and the baseline methods.

Table 3: Ablation studies of high-quality dataset filtering.

Setup	CP ↑	ΔE ↓	CD ↓	AesScore ↑
w/o filtering	0.7950	9.859	3.674	0.7727
w/ filtering	**0.8567**	**6.035**	**2.737**	**0.7834**

4.3 Ablation Study

4.3.1 Effectiveness of Large-scale Triplet Dataset

To evaluate the effectiveness of our proposed dataset, we retrain two representative tone transfer models, CAP-VSTNet [46] and Neural Preset [18], on TST100K. The retrained variants are denoted as CAP-VSTNet* and Neural Preset*, respectively. Quantitative results are reported in Tab. 1.

Both retrained models exhibit notable performance gains compared to their original counterparts, particularly in color difference and aesthetic quality. This demonstrates that large-scale triplet supervision is crucial for tone style transfer.

In addition, we validate the effectiveness of the high-quality filtering used in the construction of the TST100K dataset. As depicted in Table 3, the results

Table 4: Filter-based image retrieval evaluation on the IFFI dataset. TS-WCL denotes our Tone Scorer trained with weakly supervised contrastive learning, and TS-WCL-PL further incorporates preference learning.

Method	Recall@1 ↑	Recall@2 ↑	Recall@5 ↑	PAcc ↑
VGG Gram [17]	34.00	47.75	67.44	58.21
Neural Disc [18]	34.63	48.19	69.94	65.12
CSD [42]	63.13	73.75	87.50	70.89
TS-WCL	93.19	97.44	99.38	74.59
TS-WCL-PL	**97.50**	**99.56**	**100.0**	**82.67**

Table 5: Ablation studies of Reward Learning. WCL indicates the reward model trained with Weakly-supervised Contrastive Learning only, while PL indicates the reward model further fine-tuned with Preference Learning.

WCL	PL	CP ↑	ΔE ↓	CD ↓	AesScore ↑
-	-	0.8567	6.035	2.737	0.7834
✓	-	0.8572	5.795	2.765	0.7871
✓	✓	**0.8644**	**5.776**	**2.634**	**0.7904**

demonstrate that filtering out inconsistent and aesthetically unpleasing reference–stylized pairs during large-scale data construction significantly enhances model performance. The experiment was conducted without incorporating reward learning. This finding highlights the importance of high-quality data curation for achieving robust tone style transfer.

4.3.2 Effectiveness of the Tone Style Scorer

To evaluate the effectiveness of our tone style scorer in perceiving preset styles, we conduct an image filter retrieval experiment on the IFFI dataset [20], as shown in Tab. 4. We compare the tone style scorer with prior style similarity methods, including VGG Gram [17], Neural Disc [18] and CSD [42]. TS-WCL is our tone style scorer trained using only tone style similarity learning, while TS-WCL-PL is the model further fine-tuned with 20K human-ranked pairs.

We evaluate the tone style scorer from two perspectives: retrieval accuracy and preference alignment. Preference accuracy (PAcc) measures alignment between model-predicted and human-preferred triplet rankings. CSD achieves stronger performance than VGG Gram and Neural Disc, but it still falls behind TS-WCL, especially in Recall@1 and PAcc, indicating that general style similarity is less effective than tone-specific representation learning for capturing fine-grained tonal differences. TS-WCL-PL further improves performance, reaching 100% Recall@5 and 82.67% PAcc. These results validate the effectiveness of our two-stage training strategy and confirm the effectiveness of the tone style scorer in filtering inconsistent reference–stylized pairs.

4.3.3 Effectiveness of Tone Reward Learning

Tab. 5 summarizes the fine-tuning results of ICTone with different tone style scorers. Comparing the first and second rows, we observe a clear improvement in ΔE, indicating that tone similarity learning enables the scorer to better capture global style characteristics and effectively guide ICTone via reward learning. Further, comparing the second and third rows, all metrics show consistent gains, demonstrating that incorporating human preference signals allows the scorer to refine ICTone with improved style consistency and enhanced perceptual quality.

5 Conclusion

This paper presents TST100K and TST2K, the first large-scale datasets providing content-reference-stylized triplets for training and benchmarking tone style transfer. A two-stage tone style scorer is trained to ensure high quality and tone consistency during dataset construction. Leveraging the semantic priors of large generative models, we propose ICTone, a DiT framework that treats content and reference images in an in-context manner. Extensive experiments demonstrate that ICTone achieves perceptually consistent tone transfer when facing significant structural and semantic discrepancies between input images, achieving state-of-the-art in both quantitative metrics and human evaluations.

Limitations and future work: Our dataset construction pipeline is limited by the stylistic bias of existing aesthetic assessment models. Stronger aesthetic models could further improve the filtering of high-quality personalized styles. At the model level, our current strategy concatenates the content and reference images as input, which is not computationally efficient. More efficient conditioning mechanisms for reference-guided tone transfer are also worth exploring.

Acknowledgments

This research was supported by the Fund of the Shenzhen Science and Technology Program (JCYJ20250604184027034, QNXMB20250701090801002), Guangdong Basic and Applied Basic Research Foundation (2026A1515011435), the National Natural Science Foundation of China (Grant No. 62576177), the Fundamental Research Funds for the Central Universities 070-63263247, and the Tianjin Key Laboratory of Visual Computing and Intelligent Perception (VCIP). Computation is supported by the Supercomputing Center of Nankai University (NKSC). "Science and Technology Yongjiang 2035" key technology breakthrough plan project (2025Z053), Chinese government-guided local science and technology development fund projects (scientific and technological achievement transfer and transformation projects) (254Z0102G), National Science Fund of China under Grant No. 62361166670.

References

1. Afifi, M., Brown, M.S.: Deep white-balance editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1397–1406 (2020)
2. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: Artflow: Unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 862–871 (2021)
3. An, J., Xiong, H., Huan, J., Luo, J.: Ultrafast photorealistic style transfer via neural architecture search. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10443–10450 (2020)
4. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision. pp. 446–461. Springer (2014)
5. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: CVPR 2011. pp. 97–104. IEEE (2011)
6. Chang, Z., Duan, Z.P., Zhang, J., Guo, C.L., Liu, S., Chun, H., Park, H., Liu, Z., Li, C.: Pertouch: Vlm-driven agent for personalized and semantic image retouching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 2752–2759 (2026)
7. Chen, H., Wang, Z., Yang, Y., Sun, Q., Ma, K.: Learning a deep color difference metric for photographic images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22242–22251 (2023)
8. Chiu, T.Y., Gurari, D.: Photowct2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2868–2877 (2022)
9. Duan, Z.P., Zhang, J., Lin, Z., Jin, X., Wang, X., Zou, D., Guo, C.L., Li, C.: Diffretouch: Using diffusion to retouch on the shoulder of experts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2825–2833 (2025)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
12. Gevers, T., Gijsenij, A., Van de Weijer, J., Geusebroek, J.M.: Color in computer vision: fundamentals and applications. John Wiley & Sons (2012)
13. Gong, Z., Wu, Z., Tao, Q., Li, Q., Loy, C.C.: Sa-lut: spatial adaptive 4d look-up table for photorealistic style transfer. arXiv preprint arXiv:2506.13465 (2025)
14. Google: Nano Banana 2: Combining pro capabilities with lightning-fast speed. `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/` (2026), accessed: 2026-05-05
15. Ho, M.M., Zhou, J.: Deep preset: Blending and retouching photos with color style transfer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2113–2121 (2021)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

17. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. pp. 694–711. Springer (2016)
18. Ke, Z., Liu, Y., Zhu, L., Zhao, N., Lau, R.W.: Neural preset for color style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14173–14182 (2023)
19. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
20. Kinli, F., Ozcan, B., Kirac, F.: Instagram filter removal on fashionable images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 736–745 (June 2021)
21. Larchenko, M., Lobashev, A., Guskov, D., Palyulin, V.V.: Color transfer with modulated flows. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4464–4472 (2025)
22. Li, D., Wu, T., Lin, B., Chen, Z., Zhang, Y., Li, Y., Cheng, M.M., Li, X.: WOW-seg: A word-free open world segmentation model. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=AyJPSnE1bq
23. Li, X., Liu, W., Li, X., Zhou, F., Li, H., Nie, F.: All-weather multi-modality image fusion: Unified framework and 100k benchmark. Information Fusion p. 104130 (2026)
24. Li, Y., Li, X., Zhang, Z., Bian, Y., Liu, G., Li, X., Xu, J., Hu, W., Liu, Y., Li, L., et al.: Ic-custom: Diverse image customization via in-context learning. arXiv preprint arXiv:2507.01926 (2025)
25. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. Advances in neural information processing systems **30** (2017)
26. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A closed-form solution to photorealistic image stylization. In: European Conference on Computer Vision. pp. 453–468 (2018)
27. Liang, J., Zeng, H., Cui, M., Xie, X., Zhang, L.: Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 653–661 (2021)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
29. Liu, R., Zhao, E., Liu, Z., Feng, A., Easley, S.J.: Universal photorealistic style transfer: A lightweight and adaptive approach. arXiv e-prints pp. arXiv–2309 (2023)
30. Liu, Y., Zheng, H., Yang, S.: Sad: Style-aware diffusion adaptation for few-shot style transfer image generation. Computational Visual Media (2025)
31. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4990–4998 (2017)
32. Luo, M.R., Cui, G., Rigg, B.: The development of the cie 2000 colour-difference formula: Ciede2000. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur **26**(5), 340–350 (2001)

33. Lv, C., Zhang, D., Geng, S., Wu, Z., Huang, H.: Color transfer for images: A survey. ACM Transactions on Multimedia Computing, Communications and Applications **20**(8), 1–29 (2024)
34. Marschner, S., Shirley, P.: Fundamentals of computer graphics. CRC press (2021)
35. Pitié, F., Kokaram, A.: The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In: 4th European conference on visual media production. pp. 1–9. IET (2007)
36. Pitie, F., Kokaram, A.C., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. vol. 2, pp. 1434–1439. IEEE (2005)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
38. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer graphics and applications **21**(5), 34–41 (2002)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
40. Sheng, X., Li, L., Chen, P., Wu, J., Dong, W., Yang, Y., Xu, L., Li, Y., Shi, G.: Aesclip: Multi-attribute contrastive learning for image aesthetics assessment. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1117–1126 (2023)
41. Skorokhodov, I., Sotnikov, G., Elhoseiny, M.: Aligning latent and image spaces to connect the unconnectable. arXiv preprint arXiv:2104.06954 (2021)
42. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Measuring style similarity in diffusion models. arXiv preprint arXiv:2404.01292 (2024)
43. Soria, X., Pomboza-Junez, G., Sappa, A.D.: Ldc: Lightweight dense cnn for edge detection. IEEE Access **10**, 68281–68290 (2022)
44. Wang, Y., Liu, R., Lin, J., Liu, F., Yi, Z., Wang, Y., Ma, R.: Omnistyle: Filtering high quality style transfer data at scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7847–7856 (2025)
45. Wen, H., You, S., Fu, Y.: Lucie: Language-guided local image editing for fashion images. Computational Visual Media **11**(1), 179–194 (2025)
46. Wen, L., Gao, C., Zou, C.: Cap-vstnet: Content affinity preserved versatile style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18300–18309 (2023)
47. Wu, G., Zhang, S., Shi, R., Gao, S., Chen, Z., Wang, L., Chen, Z., Gao, H., Tang, Y., Cheng, M.M., et al.: Representation entanglement for generation: Training diffusion transformers is much easier than you think. Advances in Neural Information Processing Systems **38**, 7714–7743 (2026)
48. Wu, J., Wang, Y., Li, L., Zhang, F., Xue, T.: Goal conditioned reinforcement learning for photo finishing tuning. Advances in Neural Information Processing Systems **37**, 46294–46318 (2024)
49. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. arXiv preprint arXiv:2504.02160 (2025)

50. Xiao, X., Ma, L.: Color transfer in correlated color space. In: Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications. pp. 305–309 (2006)
51. Xing, P., Wang, H., Sun, Y., Wang, Q., Bai, X., Ai, H., Huang, R., Li, Z.: Csgo: Content-style composition in text-to-image generation. arXiv preprint arXiv:2408.16766 (2024)
52. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
53. Yeo, W.H., Oh, W.T., Kang, K.S., Kim, Y.I., Ryu, H.C.: Cair: fast and lightweight multi-scale color attention network for instagram filter removal. In: European Conference on Computer Vision Workshops. pp. 714–728. Springer (2022)
54. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9036–9045 (2019)
55. Zhang, X., Li, D., Dong, X., Wu, T., Yu, H., Wang, J., Li, Q., Li, X.: Unichange: Unifying change detection with multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42169–42179 (2026)
56. Zhang, Y., Li, D., Li, Y., Zhang, X., Xie, T., Cheng, M., Li, X.: Crystal: Spontaneous emergence of visual latents in mllms. arXiv preprint arXiv:2602.20980 (2026)
57. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. Advances in Neural Information Processing Systems **38**, 139195–139227 (2026)
58. Zhao, C., Zhu, C., Feng, X., Hao, A., Zhu, J., Lei, J., Wu, J., Chu, X., Yang, J.: Extending one-step image generation from class labels to text via discriminative text representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 36649–36659 (2026)