GenColor: Generative and Expressive Color Enhancement with Pixel-Perfect Texture Preservation

Yi Dong[1†], **Yuxi Wang**[1†], **Xianhui Lin**[2], **Wenqi Ouyang**[1], **Zhiqi Shen**[1*], **Peiran Ren**[2*], **Ruoxi Fan**[1], **Rynson W. H. Lau**[3*]
[1]Nanyang Technological University, [2]Alibaba Group, [3]City University of Hong Kong
Project Page: https://yidong.pro/projects/gencolor

Figure 1: GenColor achieves expressive color enhancement with superior texture preservation. While unsupervised methods (Exposure [13], Distort-and-Recover [28]) lack expressiveness, supervised approaches (3D-LUT [35], RSFNet [26]) are constrained by global adjustments in their training data. Generative methods (*e.g.*, Midjourney), on the other hand, create dramatic transformations but alter textures, compromising authenticity. The difference maps (2nd row) show GenColor making selective region-specific adjustments similar to Midjourney, but with better results than Expert C (the best retoucher in Adobe5K [2]), which is limited to global filter adjustments.

Abstract

Color enhancement is a crucial yet challenging task in digital photography. It demands methods that are (i) expressive enough for fine-grained adjustments, (ii) adaptable to diverse inputs, and (iii) able to preserve texture. Existing approaches typically fall short in at least one of these aspects, yielding unsatisfactory results. We propose GenColor, a novel diffusion-based framework for sophisticated, texture-preserving color enhancement. GenColor reframes the task as conditional image generation. Leveraging ControlNet and a tailored training scheme, it learns advanced color transformations that adapt to diverse lighting and content. We train GenColor on ARTISAN, our newly collected large-scale dataset of 1.2M high-quality photographs specifically curated for enhancement tasks. To overcome texture preservation limitations inherent in diffusion models, we introduce a color transfer network with a novel degradation scheme that simulates texture–color relationships. This network achieves pixel-perfect texture preservation while en-

*Corresponding authors, †Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

abling fine-grained color matching with the diffusion-generated reference images. Extensive experiments show that GenColor produces visually compelling results comparable to those of expert colorists and surpasses state-of-the-art methods in both subjective and objective evaluations. We have released the code and dataset.

1 Introduction

Color enhancement turns ordinary photographs into compelling visual narratives through the fine-grained adjustments professional colorists apply to evoke emotion and aesthetic quality. Despite considerable progress, automated algorithms still struggle to match human expertise. As shown in Figure 1, three fundamental challenges impede progress in this domain: (1) achieving fine-grained expressiveness comparable to human colorists; (2) adapting to diverse input conditions involving varied lighting and content; and (3) preserving the textural fidelity of the input images while making color adjustments.

Existing approaches face significant limitations in addressing these challenges. Unsupervised methods like Exposure [13] and Distort-and-Recover [28] preserve texture but lack expressiveness, particularly in difficult lighting conditions. Supervised methods like DeepLPF [23] and RSFNet [26] rely on datasets, such as Adobe FiveK [2] and PPR10K [20] that contain only global adjustments, fail to capture the complex local modifications characteristic of expert retouching. These approaches produce less vibrant results, as visible in Figure 1. Meanwhile, generative methods like Midjourney create dramatic transformations but significantly alter textures, distorting details and introducing unnatural patterns that compromise the authenticity of the original scene. These collective limitations underscore the need for a color enhancement approach that combines expressive adjustments with texture preservation capabilities across diverse real-world conditions.

In this paper, we present GenColor, a novel framework reinterpreting color enhancement as a texture-preserving conditional image generation task. Our key insight is to leverage the expressive generation capabilities of diffusion models while decoupling texture preservation to a specialized transfer network, ensuring fine-grained adjustments without compromising texture fidelity.

For expressive color transformations, we curate ARTISAN—believed to be the largest dataset specifically designed for image enhancement with 1.2 million high-quality photographs—and develop a diffusion-based color generation approach with three distinctive technical advances. A central observation in our work is that diffusion models, traditionally known for generating new content, can be repurposed for color enhancement through targeted conditioning and careful training regimes. First, we reframe color enhancement as conditional generation by leveraging ControlNet [36] conditioned directly on input images, unlocking superior color expressiveness compared to traditional methods that rely on carefully designed filters with limited creative range. Second, we adapt distort-and-recover principles to diffusion models through a self-supervised training strategy with a wider range of color adjustments, enhancing compatibility with challenging lighting conditions and diverse content. Third, we discover that strategic blending of weights from different training stages yields a good balance between creative expressiveness and artifact control. These three technical advances enable highly expressive color generation that adapts to diverse input conditions, achieving sophisticated aesthetic enhancements that traditional methods cannot attain.

Although diffusion models excel at generating aesthetically pleasing colors, their intrinsic iterative denoising process inevitably alters textural elements—a fundamental limitation that can be alleviated but never fully eliminated. To address this limitation, we develop a texture-preserving network that achieves three critical objectives simultaneously: (1) transferring semantically consistent color styles from the diffusion-generated reference images to the original input, (2) restoring artifacts that typically appear as random noise or irregular color strokes in the diffusion-generated color reference image, and (3) performing implicit super-resolution to handle the resolution discrepancy between diffusion models (typically 512×512 pixels) and high-resolution input images. While this task resembles traditional color matching or transfer, a key technical observation guided our approach—the diffusion model's output and the input image are near-duplicates at the texture level, with relevant pixels for color transfer spatially placed in nearby neighborhoods. To leverage these characteristics, we design a self-supervised learning approach with a novel degradation scheme that simulates all three objectives: circular shift to model the nearby spatial misalignment, random stroke degradation to simulate or restore artifacts, and random resolution rescaling to simulate the

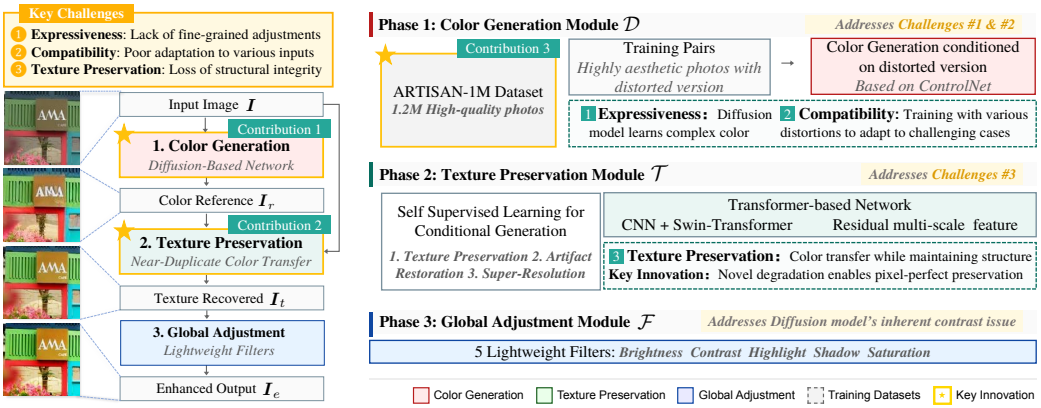


Figure 2: GenColor framework overview: (left) the overall pipeline and (right) summary/highlights of each module. The approach addresses three key challenges through a three-phase process: (1) A diffusion-based Color Generation module creates expressive color enhancements conditioned on the input image, trained on the ARTISAN-1M dataset; (2) A Texture Preservation module maintains texture integrity while transferring colors from the color reference $\boldsymbol{I}_r$ to the input image $\boldsymbol{I}$; (3) A Global Adjustment module applies five lightweight filters to enhance contrast and saturation.

super-resolution challenge. This design, based on accurate observations of underlying data patterns, achieves superior texture preservation and significantly better color matching performance than state-of-the-art methods.

Extensive experiments demonstrate that GenColor outperforms existing methods in both subjective evaluations and objective metrics for color enhancement and texture-preserving color transfer. As illustrated in Figure 1, our approach outperforms existing methods while maintaining texture fidelity that generative methods sacrifice. This example image showcases GenColor's ability to create beautiful contrast between the golden building tones and clear blue sky, intelligently lightening shadowed areas while preserving architectural details. It demonstrates how our approach achieves more expressive results than Human Expert C—a limitation inherent to the supervised learning methods, which have relied on Adobe5K due to the absence of fine-grained paired datasets, for color enhancement—particularly in scenes with complex lighting and content that require different color treatments for specific elements. Our key contributions include:

1. A diffusion-based color enhancement framework capable of expert-level fine-grained adjustments, supported by a training scheme that enables sophisticated color generation with selective manipulations across varying lighting and content conditions.
2. A texture-preserving color transfer network that maintains pixel-perfect textural fidelity while achieving precise color matching with the diffusion-generated reference images, resulting in outputs that are both visually compelling and faithful to the original content.
3. ARTISAN, a large-scale dataset of 1.2 million high-aesthetic-quality photos specifically curated for color enhancement, enabling robust training of advanced models.

2 Related Works

2.1 Image Color Enhancement Methods

Traditional color enhancement relies on global adjustments using predefined filters or lookup tables (LUTs), which lack expressiveness for fine-grained local adjustments. To address this limitation, color enhancement methods that operate at a fine-grained level have been proposed. Supervised learning approaches such as DeepLPF [23], 3D-LUT [35], and RSFNet [26] learn enhancement mappings from paired datasets such as Adobe FiveK [2] and PPR10K [20] but are constrained by these datasets' focus on global adjustments rather than sophisticated local edits. Recent work like ICELUT [34] improves interpretability while still limited by training data expressiveness. Unsupervised methods, including Exposure [1], EnhanceGAN [4], and Distort-and-Recover [27], overcome paired data limitations but struggle to achieve fine-grained control comparable to professional retouching. Our

(a) ARTISAN-100 K

(b) ARTISAN-1 M

Dataset size	$R_{\Delta H} \downarrow$
100 K	11.9 %
1.2 M	1.3 %

(c) Strong hue change

Figure 3: Impact of dataset size on color enhancement quality. (a) Model trained on ARTISAN-100K produces artifacts. (b) Model trained on ARTISAN-1M generates high-quality results. (c) $R_{\Delta \mathrm{H}}$ measures the percentage of output images with significant areas where the hue has changed substantially compared to the input, indicating undesirable color shifts. ARTISAN-1M reduces this metric, enabling effective enhancement learning.

approach addresses these limitations through diffusion models' generative capabilities combined with texture preservation mechanisms.

2.2 Diffusion Models for Image Color Editing

Diffusion models [11] have revolutionized image synthesis, with latent diffusion models [29] improving efficiency through compressed latent space operations. ControlNet [36] enables structure-guided generation through conditional signals, while ControlColor [21] successfully applied this approach to image colorization. In a related vein, DiffRetouch [7] employs Stable Diffusion for multi-style photo retouching. Nevertheless, its focus remains on user-adjustable global attributes, whereas our work enables expert-level fine-grained color generation with adaptability across diverse lighting and content conditions.

2.3 Texture Preservation in Image Enhancement

Diffusion-based methods struggle with texture preservation due to latent space compression (1/8 or 1/16 downsampling). While DiffRetouch [7] mitigates distortion using Affine Bilateral Grid for pixel-wise adjustments, two fundamental limitations remain: it operates in latent space where high-frequency details are already lost during downsampling, and its grid-based approach introduces interpolation errors in complex scenes like text. Several established approaches for texture preservation also face challenges in our setting. Existing approaches include: (1) hint-based color propagation methods (ControlColor [21], UniColor [14]) that preserve lightness but are unsuitable for enhancements requiring luminance adjustments; (2) style transfer methods (CAP-VSTNet [31]) that sacrifice texture fidelity; and (3) color matching/transfer techniques (Color Matcher [9], NeuralPreset [16]) that achieve pixel-perfect texture preservation but aren't optimized for scenarios where relevant color pixels for transfer exist in nearby spatial neighborhoods between input and reference images.

Our approach uses a CNN-transformer hybrid operating directly in image space, leveraging the observation that diffusion outputs and inputs are texture-level near-duplicates with correspondences in nearby spatial neighborhoods. This insight enables self-supervised learning with a novel degradation scheme, significantly outperforming existing methods in accuracy and texture preservation.

3 ARTISAN-1M Dataset

Deep learning requires large, high-quality datasets with diverse content and styles. Existing datasets fall short - Adobe5K [2] and PPR10K [20] are too small, while LAION-Aesthetics V2 [30] lacks real-world texture details as the majority of its aesthetically high subset consists of non-photographic artworks. We introduce ARTISAN-1M, containing 1.2 million carefully curated photographs from online platforms (primarily Flickr), selected based on visual quality (Q-Align score [32]), content diversity, and color representation. It includes 1.0M daytime and 240.9K nighttime images, with balanced distribution between images with people (617.0K) and without (653.0K).

ARTISAN-1M enables our color generation module to learn complex enhancement patterns. As shown in Figure 3, models trained on the smaller ARTISAN-100K subset exhibit artifacts with 11.9% of images having undesirable strong hue shifts, while the full dataset reduces this to just 1.3%. When strong hue changes occur with the full dataset, changing the generator's seed typically resolves the issue. ARTISAN-1M's unprecedented scale and quality provides a solid foundation for GenColor to

Figure 4: Overview of our color generation module. (1) High-quality aesthetic images from our ARTISAN-1M dataset are used as training targets. (2) A wide range of random color adjustments (exposure, contrast, hue, saturation, warmth, *etc*.) are applied to the target images to create inputs with strong to slight deviations as conditioning. (3) The ControlNet architecture, based on Stable Diffusion, is then trained on this large dataset, learning to map the color distorted inputs back to the aesthetic targets, capturing fine-grained color enhancement nuances.

learn expressive yet faithful color enhancement from real-world high-quality photography. Due to space limitations, the details of the datasets are discussed in the appendix.

4 Method

4.1 Method Overview

As shown in Figure 2, we present GenColor, a novel framework that addresses the fundamental challenges of image color enhancement through a three-phase approach. Given an input image $\boldsymbol{I}$, our goal is to produce an aesthetically enhanced output $\boldsymbol{I}_e$ while preserving original texture details. First, our color generation module employs a diffusion model conditioned on the input image to create a color reference $\boldsymbol{I}_r = \mathcal{D}(\boldsymbol{I})$. This module, trained on our ARTISAN-1M dataset, leverages ControlNet [36] to learn sophisticated color transformations while carefully balancing expressiveness and artifact control through strategic blending of weights from different training stages. To address the inherent texture limitations of diffusion models, our texture preservation module $\boldsymbol{I}_t = \mathcal{T}(\boldsymbol{I}, \boldsymbol{I}_r)$ transfers color characteristics from the reference while maintaining the structural details of the input image. Finally, to counteract diffusion models' tendency to produce images with reduced contrast and vibrancy [22], a global adjustment module $\boldsymbol{I}_e = \mathcal{F}(\boldsymbol{I}_t)$ applies five essential filters (brightness, contrast, highlight, shadow, and saturation) to enhance contrast and vibrancy while preserving the sophisticated color transformations.

Our complete enhancement pipeline can be expressed as: $\boldsymbol{I}_e = \mathcal{F}(\mathcal{T}(\boldsymbol{I}, \mathcal{D}(\boldsymbol{I})))$. This formulation effectively combines the expressive power of diffusion models with precise texture preservation, enabling high-quality color enhancement that rivals professional human retouching.

4.2 Color Generation

4.2.1 ControlNet-based Color Generation

As shown in Figure 4, we formulate color enhancement as conditional generation using the ControlNet architecture [36] that . This approach shows superior color expressiveness compared to traditional methods that rely on carefully designed filters with limited creative range. ControlNet adds an auxiliary network branch that conditions on a control signal derived from a color-distorted version of the input image $\mathcal{A}(\boldsymbol{I})$, where $\mathcal{A}$ is a random color adjustment function. We train ControlNet on our high-quality ARTISAN-1M dataset, using the original images $\boldsymbol{I}$ as targets and color-distorted versions as conditioning inputs. The distortions range from strong deviations (extreme lighting conditions, restoration/colorization scenarios) to slight deviations (suboptimal color, fine color edits), creating a comprehensive training regime. This approach enables our method to tackle challenging inputs with high contrast lighting, difficult shadows, and many other complex scenarios that traditional methods struggle with. Although each image is adjusted globally, training over our large dataset enables the model to capture fine-grained color enhancement nuances that apply to different image regions contextually. In addition, we have found a weight blending strategy that can improve the performance of the color generation module, please refer to the appendix §B for details. We utilize a

Task	Degradation
Texture Preserving Color Transfer	①②
Artifact Restoration (Small Color Strokes)	③
Super-Resolution (Low-Res $\boldsymbol{I}_r$)	④

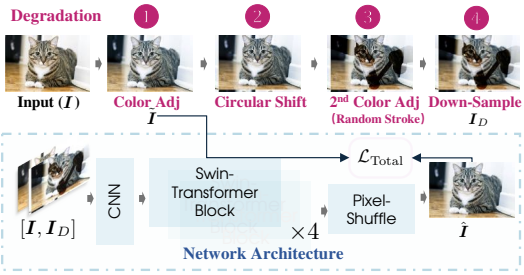


Figure 5: Texture preservation module and its training data generation process. The network aims to transfer the color characteristics from the reference image $\boldsymbol{I}_r$ while preserving its structural details.

null text prompt by default, as textual information provides a negligible contribution to this task, a finding validated in Appendix §B.3.

4.3 Texture Preservation Module

4.3.1 Problem Formulation

The texture preservation network $\mathcal{T}$ aims to transfer the color characteristics from the reference image $\boldsymbol{I}_r$ generated by the color generation module $\mathcal{D}$ while preserving the high-frequency structural details from the input image $\boldsymbol{I}$. This can be formulated as a conditional generation problem: $p(\boldsymbol{I}_e|\boldsymbol{I}, \boldsymbol{I}_r) = \mathcal{T}(\boldsymbol{I}, \mathcal{U}(\boldsymbol{I}_r); \theta)$ where $\boldsymbol{I}_e$ is the enhanced output image, $\mathcal{U}$ represents the upsampling operation to match the input resolution, and θ denotes the learnable parameters of $\mathcal{T}$.

4.3.2 Degradation Process for Training Data Generation

In the absence of ground truth pairs, we design a degradation model to generate synthetic training data that mimics the characteristics of our unique problem, including color differences, texture and structure similarities, and artifacts. As shown in Figure 5, the degradation process consists of:

1. A random circular shift operation that maintains similar overall texture and structure between images while avoiding exact pixel-level correspondence.
2. Global color distortion applied randomly across the entire image to simulate the diverse range of color transformations that may occur during enhancement.
3. Localized color deviation strokes added in randomly selected sparse regions to replicate potential artifacts that could emerge during the color generation process.
4. Resolution adjustment through image resizing operations to account for potential differences in resolution between the input and reference images.

4.3.3 Network Architecture and Training

We propose a Transformer-based network (Figure 5) with residually connected feature extraction blocks integrating CNN and Swin-Transformer modules for effective local-global feature capture. The network is trained self-supervisedly using the color distorted image as ground truth and the degraded image paired with the original input as training data, optimized with a combination of Huber, perceptual, and adversarial losses. During inference, the degraded image is replaced by the output from the color generation network $\mathcal{D}$. This unified framework effectively addresses color transfer, artifact reduction, and resolution enhancement while preserving the input image's structural details. Please refer to the appendix for more details.

5 Experiments

We present only key results here; see appendix for comprehensive analyses.

5.1 Datasets

We evaluate on three datasets: Adobe-FiveK [2] (5,000 raw images with expert retouches), PPR10K [20] (11,161 portrait photos with edits and masks), and FreeRaw (142 high-quality raw photos). Results for Adobe-FiveK and PPR10K are in the main paper, with FreeRaw in the appendix.

5.2 Metrics

We evaluate using three categories of metrics: (1) **Color Enhancement** metrics including Q-Align [32] (emulates human judgment), NoR-HDR [8] (assesses dynamic range), and LIQE [37] (blind image quality evaluation); (2) **Texture Preservation** metrics including TD [6] (quantifies texture distortion), DISTS [5] (assesses structural/textural similarity), and GMSD [33] (compares gradient magnitude maps); and (3) **Color Similarity** metrics including W_1 (measures color distribution similarity), Semantic W_1 (captures fine-grained color characteristics), and MS-SWD [10] (latest metric aligned with human judgment).

5.3 Baseline Methods

We compare GenColor with state-of-the-art color enhancement methods: Exposure [13], Distort-and-Recover [28], 3D-LUT [35], DeepLPF [23], RSFNet [26], and ICELUT [34]. For texture-preserving color transfer, we compare with ColorMatcher [9], NLUT [35], Deep Preset [12], and a hint color propagation method from Colorization via Imagination [3], originally proposed in UniColor [14]. We exclude style transfer methods like CAP-VSTNet [31] (poor texture preservation) and NeuralPreset [16] (unavailable code). Appendix §I shows Tinge, an app, which uses NeuralPreset's algorithm, performs slightly better than ColorMatcher but significantly worse than GenColor.

5.4 Image Enhancement Quantitative Results

Table 1 presents a quantitative comparison of color enhancement quality on the Adobe-FiveK and PPR10K datasets. Our GenColor method achieves the best results on most metrics across all datasets, outperforming state-of-the-art methods and approaching or surpassing the performance of a human expert retoucher on the FiveK and PPR10K datasets. GenColor obtains the highest scores on Q-Align and NoR-VDPNet for all datasets, and the best LIQE on FiveK and FreeRaw, with its LIQE on PPR10K being second only to the human expert. Please note that all supervised methods are trained on their respective dataset, while we provide additional results for unsupervised methods trained on our ARTISAN-1M dataset, which indicated by (ARTISAN).

Table 1: Evaluation of GenColor's color-enhancement quality against state-of-the-art methods on various datasets using five perceptual quality metrics: Q-Align [32], LAION [30], LIQE [37], NoR-VDP [8], and C-VAR (↑). Higher values indicate better performance, with best results in **bold**.

Method	**Adobe5K**					**PPR10K**				
	Aesthetic		Quality	Dyn. Range	Express.	Aesthetic		Quality	Dyn. Range	Express.
	Q-Align↑	LAION↑	LIQE↑	NoR-VDP↑	C-VAR↑	Q-Align↑	LAION↑	LIQE↑	NoR-VDP↑	C-VAR↑
3D-LUT [35]	4.22	5.73	3.66	67.08	11.48	4.22	6.03	2.83	69.83	8.09
RSFNet [26]	4.23	5.74	3.65	67.85	10.09	4.29	6.05	2.97	69.69	6.68
DeepLPF [23]	4.16	5.79	3.53	69.08	10.97	3.92	6.08	2.55	71.07	11.33
ICELUT [34]	4.17	5.70	3.55	66.40	10.44	3.97	5.95	2.49	68.18	7.21
D&R [28]	2.39	5.33	1.35	64.43	10.29	2.44	5.30	1.21	67.78	5.68
D&R (ARTISAN)	2.31	5.24	1.31	63.92	9.92	2.34	5.23	1.18	67.21	5.41
Exposure [13]	4.00	5.72	3.29	66.02	13.96	3.90	5.96	2.37	70.73	12.44
Exposure (ARTISAN)	3.93	5.66	3.25	65.69	13.48	3.86	5.88	2.29	70.32	12.00
GenColor	**4.29**	**5.83**	**3.76**	**69.16**	**16.96**	**4.38**	**6.26**	**3.21**	**71.11**	**13.57**

5.5 Analysis of Human Expert and LLM Preference Evaluation

Table 2 compares GenColor against human expert performance and LLM preferences. GenColor consistently outperforms Human Expert C across all five quantitative metrics: Q-Align (4.29 vs

Table 2: Comparison with human expert and LLM preference evaluation.

Method	Q-Align ↑	LAION ↑	LIQE ↑	NoR-VDP ↑	C-VAR ↑
Human Expert C	4.24	5.72	3.65	67.86	10.91
GenColor (Ours)	**4.29**	**5.83**	**3.76**	**69.16**	**16.96**

(a) Head-to-head comparison between GenColor and professional human retoucher (Expert C in Adobe5K [2]). Results show GenColor matches or exceeds human expert.

	Input	Expert C	GenColor
Preference	0	38	**62**

(b) LLM (ChatGPT o1) preference on Adobe5K [2] (first 100 images). GenColor outperforms Expert C, the widely-used benchmark expert.

4.24), LAION (5.83 vs 5.72), LIQE (3.76 vs 3.65), NoR-VDP (69.16 vs 67.86), and C-VAR (16.96 vs 10.91). Notably, GenColor achieves a substantial 55.5% improvement in C-VAR, indicating significantly better color accuracy.

The LLM preference evaluation provides additional validation. When ChatGPT o1 evaluated 100 Adobe5K images, GenColor was preferred in 62% of cases compared to Expert C's 38%. This 24 percentage point preference gap demonstrates that GenColor exceeds professional human retouching quality as perceived by advanced AI systems.

Table 3: Texture preservation evaluation on Adobe5k. GenColor achieves excellent balance between preserving textures and maintaining color accuracy.

Method	Texture Preservation			Color Similarity		
	TD ↓	DISTS ↓	GMSD ↓	Sem W_1 ↓	W_1 ↓	MS-SWD ↓
NLUT [35]	0.95	0.15	**0.12**	13.01	12.76	0.87
Color Matcher [9]	**0.42**	0.17	0.14	31.45	31.34	2.03
UniColor [14]	0.95	0.15	0.14	38.40	38.20	2.18
Deep Preset (w/o PPL) [12]	1.17	0.15	**0.12**	33.09	32.88	2.38
Deep Preset [12]	1.12	0.15	0.13	34.84	34.64	2.51
GenColor (Ours)	0.94	**0.13**	**0.12**	**3.59**	**3.22**	**0.70**

5.6 Texture Preservation Quantitative Results

As shown in Table 3, GenColor achieves strong overall performance, with Color Matcher achieving the lowest TD score (0.42) while GenColor obtains competitive texture preservation metrics including tied for best GMSD (0.12) and best DISTS scores (0.13). GenColor excels particularly in color similarity with the lowest Semantic W_1 (3.59), W_1 (3.22), and MS-SWD (0.70) values. The ablation study reveals the importance of key components in GenColor's texture preservation module training, such as the circular shift, which contribute to its balanced performance. When compared to existing methods like Color Matcher [9] and hint color propagation [14, 3], GenColor demonstrates significant improvements across most metrics, highlighting its effectiveness in texture-preserving color transfer.

5.7 Visual Results

Figure 6 presents visual comparisons of GenColor with other methods on representative images from the Adobe FiveK dataset. The qualitative results align with our quantitative findings, showing that GenColor generates more appealing and natural-looking enhancements compared to other methods. For texture preserving color transfer, Figure 7 shows that our method can better preserve the texture while achieving accurate color transfer. More results can be found in the supplementary website.

5.8 Robustness Evaluation

As shown in Figure 8, our method achieves superior performance on challenging cases, where prior methods often struggle. Additionally, Figure 9 evaluates the robustness of our method across different image resolutions. We have tested on various resolutions and found that GenColor, which incorporates a resizing operation during training data generation, maintains consistently high performance.

Figure 6: Visual comparison on the Adobe-FiveK [2] dataset. GenColor produces more natural and visually pleasing results while preserving image details compared to other state-of-the-art methods.

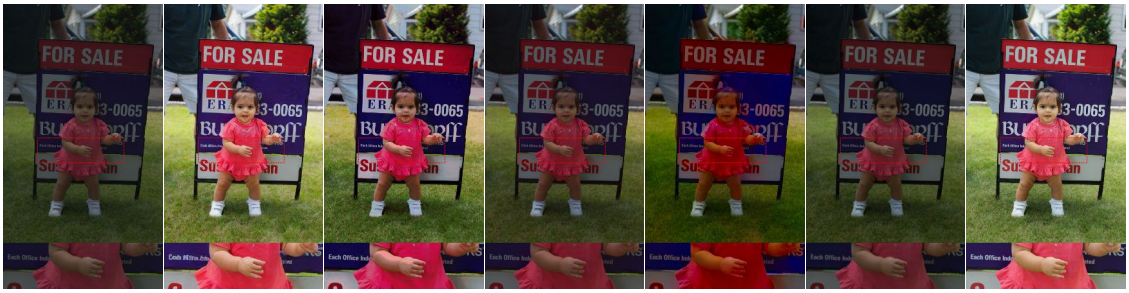


Figure 7: Visual comparison of color transfer methods. Our GenColor method better preserves the texture details while achieving accurate color transfer compared to other approaches.

5.9 User Study

To evaluate the perceptual quality of GenColor, we conducted a comprehensive user study involving 58 participants with varying levels of expertise in image processing and color enhancement, ranging from amateur to professional. Each participant was presented with 5 test images and their corresponding enhanced versions generated by all baseline methods. Participants rated the color enhancement results on a 5-point Likert scale based on color aesthetics. The results are presented in Figure 10. Statistical analysis of the ratings demonstrates that GenColor is the highest-rated method among users, confirming its ability to generate visually appealing and natural-looking enhancements that align with human perception. In addition, we conducted a pairwise comparison study to further validate the superiority of GenColor, which is presented in the appendix §G.

6 Limitations and Future Work

Our work has a few limitations. GenColor's performance is reduced on images with substantial pure black and white regions, and it can occasionally produce unintended hue transformations in extreme lighting or with ambiguous colors. While the latter can often be mitigated by changing the diffusion seed, we propose incorporating a hue preservation mask as future work for more precise control. We also observe that our model, despite being trained on the diverse ARTISAN dataset, converges to a deterministic, high-probability "mean" aesthetic due to the strong input conditioning. This presents a

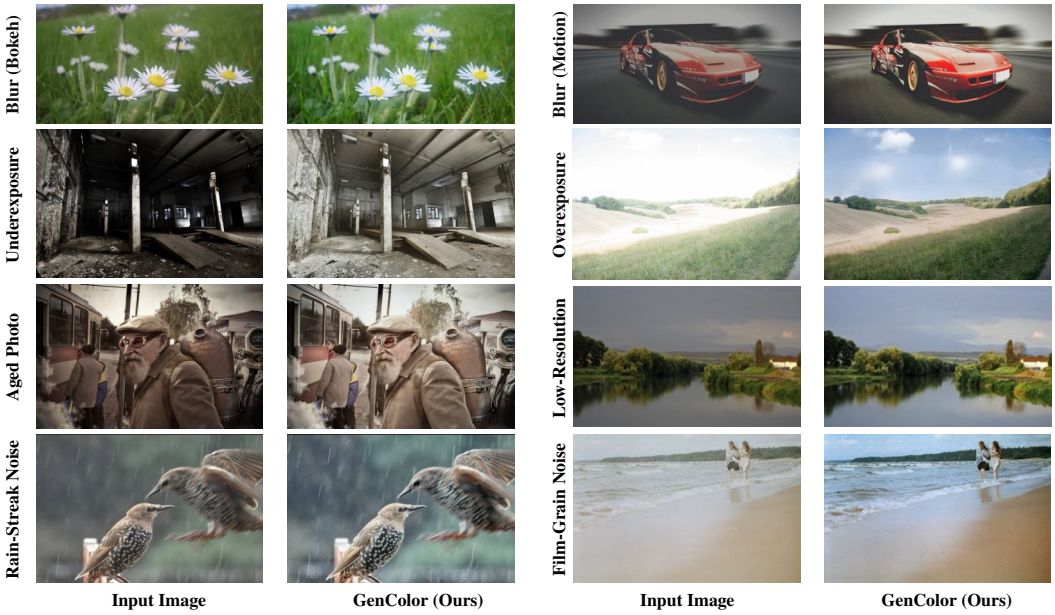
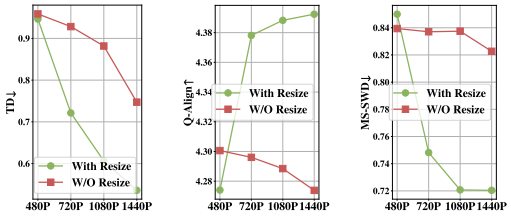


Figure 8: GenColor maintains robust performance on challenging cases.



(a) Texture (b) Aesthetic (c) Color

Figure 9: Robustness on Various Resolutions.

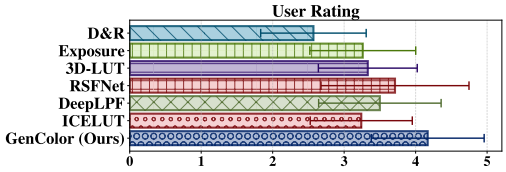


Figure 10: User study results. GenColor receives the highest rating among the compared methods.

significant opportunity for future work to unlock the dataset's full stylistic variety, perhaps through conditional generation with style exemplars or by disentangling latent style codes. Such a foundation would also enable highly data-efficient style specialization (e.g., via LoRA) using smaller, curated datasets. Lastly, while our inference-time weight-blending effectively balances aesthetics and fidelity, an alternative future direction is to design specialized multi-objective loss functions to learn this balance directly during the training process, potentially removing the need for blending.

7 Conclusion

In this paper, we have proposed GenColor, a novel framework for high-quality image color enhancement. Our method leverages a diffusion model to generate expressive color transformations and employs a dedicated texture preservation network to maintain structural integrity while achieving fine-grained color adjustments. Extensive experiments demonstrate that GenColor outperforms state-of-the-art methods on both color enhancement quality and texture preservation metrics, and it generates visually appealing results that align with human perception. The introduction of the large-scale, high-quality ARTISAN-1M dataset further contributes to the advancement of color enhancement research. We will release the code and dataset to the public to facilitate further research. We strongly encourage reviewers to explore our supplementary video and website, which vividly demonstrate the superior performance of our approach in real-world scenarios.

Acknowledgment

This work was supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore. We would like to thank the anonymous reviewers for their insightful comments and suggestions.

References

[1] Mahmoud Afifi and Michael S. Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1535–1544, 2019.

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[3] Xiaoyan Cong, Yue Wu, Qifeng Chen, and Chenyang Lei. Automatic controllable colorization via imagination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2609–2619, June 2024.

[4] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 870–878, New York, NY, USA, 2018. Association for Computing Machinery.

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

[6] Yi Dong, Yuxi Wang, Zheng Fang, Wenqi Ouyang, Xianhui Lin, Zhiqi Shen, Peiran Ren, Xuansong Xie, and Qingming Huang. Movingcolor: Seamless fusion of fine-grained video color enhancement. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 7454–7463, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Zheng-Peng Duan, Jiawei zhang, Zheng Lin, Xin Jin, Dongqing Zou, Chunle Guo, and Chongyi Li. Diffretouch: Using diffusion to retouch on the shoulder of experts, 2024.

[8] Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph.*, 36(4), July 2017.

[9] Christopher Hahne and Amar Aggoun. Plenopticam v1.0: A light-field imaging framework. *IEEE Transactions on Image Processing*, 30:6757–6771, 2021.

[10] Jiaqi He, Zhihua Wang, Leon Wang, Tsein-I Liu, Yuming Fang, Qilin Sun, and Kede Ma. Multiscale sliced Wasserstein distances as perceptual color difference measures. In *European Conference on Computer Vision*, pages 1–18, 2024.

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

[12] Man M. Ho and Jinjia Zhou. Deep preset: Blending and retouching photos with color style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2113–2121, January 2021.

[13] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Trans. Graph.*, 37(2), May 2018.

[14] Zhitong Huang, Nanxuan Zhao, and Jing Liao. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics*, 41(6):205:1–205:16, 2022.

[15] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

[16] Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson W.H. Lau. Neural preset for color style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14173–14182, June 2023.

[17] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson W.H. Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision (ECCV)*, 2022.

[18] LAION. LAION-Aesthetics: Aesthetic subset of LAION-5b. `https://laion.ai/blog/laion-aesthetics/`, 2022. Accessed: 2024-11-19.

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[20] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[21] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Control color: Multimodal diffusion-based interactive image colorization, 2024.

[22] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5392–5399, 2024.

[23] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12826–12835, 2020.

[25] Jiaqi Ouyang, Shijie Li, Shuai Li, Wangmeng Zuo, and Shuhang Gu. Rsfnet: Rich feature attention network for image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–10, 2023.

[26] Wenqi Ouyang, Yi Dong, Xiaoyang Kang, Peiran Ren, Xin Xu, and Xuansong Xie. Rsfnet: A white-box image retouching approach using region-specific color filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12160–12169, October 2023.

[27] Junghyun Park, Seungryong Choi, and Kwanghoon Kim. Distort-and-recover: Color enhancement using deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5928–5936, 2018.

[28] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-Recover: Color Enhancement Using Deep Reinforcement Learning . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5928–5936, Los Alamitos, CA, USA, June 2018. IEEE Computer Society.

[29] 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 (CVPR)*, pages 10684–10695, 2022.

[30] Christoph Schuhmann and Romain Beaumont. LAION-AESTHETICS. `https://laion.ai/blog/laion-aesthetics/`, 2022. Technical report and blog post.

[31] Linfeng Wen, Chengying Gao, and Changqing Zou. Cap-vstnet: Content affinity preserved versatile style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18300–18309, June 2023.

[32] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54015–54029. PMLR, 21–27 Jul 2024.

[33] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014.

[34] Sidi Yang, Binxiao Huang, Mingdeng Cao, Yatai Ji, Hanzhong Guo, Ngai Wong, and Yujiu Yang. Taming lookup tables for efficient image retouching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. Accepted for publication.

[35] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2058–2073, 2022.

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[37] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14071–14081, 2023.

[38] Mingrui Zhu, Xiao He, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. All-to-key attention for arbitrary style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23109–23119, October 2023.

A The ARTISAN-1M Dataset

A.1 Overview of the Dataset Collection and Curation Process

We collected 1.2M high-quality images from multiple online platforms, with Flickr being our primary source. Images were selected based on their visual quality, diversity in content, and color representation. We also manually curated the dataset to ensure its quality and diversity.

A.2 Data Sources and Criteria for Inclusion

The ARTISAN-1M dataset represents a significant advancement in image enhancement datasets, addressing the limitations of existing collections. While datasets like Adobe-FiveK [2] (5,000 images) and PPR10K [20] (11,161 images) have been valuable for research, their limited size constrains the development of more sophisticated enhancement models. Similarly, while LAION-Aesthetics V2 [30] is large, its high-aesthetic subset primarily consists of non-photographic content like illustrations and classic paintings.

A.3 Dataset Statistics

The ARTISAN-1M dataset contains 1.2M images. Table 11 shows the distribution of different attributes in our dataset. As shown in the table, our dataset includes 1.0M daytime images and 240.9K nighttime images. In terms of content, there is a balanced distribution between images containing people (617.0K) and those without people (653.0K).

Attribute	Type	Count
Time of Day	Day	1.0M
	Night	240.9K
Object	People	617.0K
	w/o People	653.0K

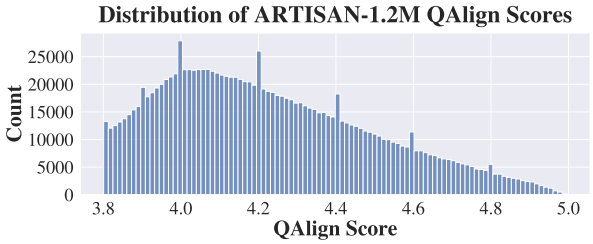


Figure 11: Left: Distribution of attributes in ARTISAN-1M dataset. Right: Distribution of Q-Align scores across the dataset, showing our dataset's high aesthetic quality.

A.4 Quality Control

For robust quality control in our dataset curation process, we leverage Q-Align [32], a state-of-the-art aesthetic assessment framework that demonstrates superior correlation with human perceptual judgment compared to the CLIP+MLP predictor employed in the LAION-Aesthetic Dataset. Through evaluation by a professional colorist, we validate that Q-Align exhibits significantly higher consistency with human aesthetic assessment in photographic enhancement scenarios. Visual examples presented in Figure 12 demonstrate that Q-Align maintains high scoring fidelity for visually compelling photographs, whereas the LAION-Aesthetic predictor shows substantial degradation in score attribution. These findings underscore the enhanced suitability of Q-Align for practical photographic applications compared to existing aesthetic prediction methodologies.

A.5 Comparison with Existing Datasets

Table 4 compares ARTISAN-1M with existing image enhancement datasets. ARTISAN-1M represents advancement in both scale and quality, containing 1.2M professional-grade photographs with diverse artistic styles and lighting conditions. In contrast, Adobe-MIT FiveK [2] provides expert retouches but is limited to 5,000 images with global adjustments only. PPR10K [20], while offering semantic masks and professional retouching, is constrained to portrait photography. LAION-Aesthetics [18], despite its large scale, has its high aesthetic score subset dominated by non-photographic content (illustrations, paintings, digital art) with textures unsuitable for photographic enhancement, making it less applicable for real-world photography tasks.

Q-Align: 4.58/5
LAION: 4.10/10

Q-Align: 4.26/5
LAION: 3.77/10

Q-Align: 4.00/5
LAION: 3.51/10

Q-Align: 4.14/5
LAION: 3.65/10

Figure 12: Comparison between Q-Align and LAION-Aesthetic's CLIP+MLP predictor scores. Q-Align scores (higher is better) are more consistent with human perception of photographic quality compared to LAION-Aesthetic predictor scores.

Table 4: Detailed comparison of image enhancement datasets. ARTISAN-1M provides significantly larger scale and broader diversity compared to existing datasets.

Dataset	**Size**	**Image Type**	**Annotations**	**Key Features**
ARTISAN-1M	1.2M	Web-crawled images	Q-Align ≥ 3.8	- Wide range of artistic styles - Diverse lighting conditions - Real-world photography - All professional-grade photography
Adobe-MIT FiveK [2]	5,000	Raw images	Expert retouches	- Multiple expert retouchers - Global adjustments only - Limited scene diversity - Raw format with metadata
PPR10K [20]	11,161	Portrait photos	Expert edits + masks	- Portrait-focused - Semantic masks - Professional retouching - Limited to portrait genre
LAION-Aesthetics	12M	Web-crawled images	Aesthetic scores ≥ 6	- High aesthetic quality subset - Diverse sources: illustrations, paintings, and digital art - No metadata for non-photographic content

B The Color Generation Module

B.1 Motivation and Clarification on Expressiveness

The paper's claim regarding the proposed method's expressiveness, particularly in comparison to "Human Expert C" is rooted in a specific and fundamental difference in adjustment paradigms. The comparison is not against a human expert with an unrestricted toolset (e.g., Photoshop with masks), but rather against the global-only adjustment paradigm inherent to the Adobe5K dataset, for which "Expert C" serves as a well-known benchmark.

Global vs. Local Adjustment Paradigms The Adobe5K dataset, a cornerstone for supervised enhancement methods, primarily consists of edits made using *global adjustment layers* in Adobe Lightroom. This constrains the expert (e.g., "Expert C") to applying a single set of parameters (e.g., exposure, saturation) uniformly across the entire image. Such a global-only approach cannot, for instance, simultaneously brighten a subject's face while deepening the color of the sky, as one action would invariably counteract the other.

GenColor operates on a different principle. It employs a generative model trained on the large-scale ARTISAN-1M dataset. This dataset is not limited to global edits and contains 1.2 million high-quality images exhibiting diverse and sophisticated local aesthetics (e.g., selective color grading, dodging and burning). By training on this data, GenColor learns to perform *spatially-varying, content-aware, and local adjustments*. It develops a semantic understanding that allows it to treat "sky," "building," and "shadows" as distinct regions deserving different enhancements within the same operation.

Table 5: Color adjustment parameters used in training data generation. We carefully constrain hue rotation to [0, 0.3] and [0.7, 1.0] to avoid drastic shifts while maintaining expressive adjustments. For all other filters, we utilize the full range to improve model expressiveness, balancing enhancement quality and model versatility.

Parameter	Range	Used Range	Full Range	Description
Brightness	-1.0 to 1.0	-1.0 to 1.0	Yes	Additive brightness adjustment
Exposure	-5.0 to 5.0	-5.0 to 5.0	Yes	Multiplicative brightness adjustment
Contrast	-1.0 to 5.0	-1.0 to 5.0	Yes	Expands/compresses tonal range
Warmth	-1.0 to 1.0	-1.0 to 1.0	Yes	Blue-yellow white balance shift
Saturation	-1.0 to 5.0	-1.0 to 5.0	Yes	Global color intensity
Vibrance	-1.0 to 5.0	-1.0 to 5.0	Yes	Selective color intensity boost
Gamma	0 to 10.0	0 to 10.0	Yes	Midtone brightness curve
Hue	0 to 1.0	0 to 0.3 & 0.7 to 1.0	**No**	Rotates colors around color wheel

Overcoming the Paired-Data Bottleneck This generative approach is necessitated by a key challenge in the field: the prohibitive cost and difficulty of collecting large-scale, paired datasets for fine-grained *local* enhancements. No such dataset for supervised local enhancement currently exists. The proposed method cleverly circumvents this data bottleneck by learning from a large, unpaired, high-quality image corpus.

Supporting Evidence The superior expressiveness of this local, content-aware approach is supported by several results:

- **Visual Evidence (Figure 1):** Visual comparisons and difference maps demonstrate this capability. GenColor can concurrently brighten a sunlit facade, deepen the blue of the sky, and lift shadows in the foreground. In contrast, the difference map for Expert C shows a uniform, global change, which is incapable of achieving these distinct, targeted effects simultaneously.
- **Quantitative Metrics (Table 2):** The method's higher expressiveness is quantitatively validated. On the *C-VAR* (Color Variance/Expressiveness) metric, designed to measure fine-grained color transformations, GenColor achieves a score of **16.96**, significantly surpassing Expert C's **10.91**.
- **Perceptual Preference (Table 2b):** In a blind comparison, a vision-language model preferred GenColor's output for "visual beauty" over Expert C's **62%** of the time, suggesting the aesthetic benefits of these expressive, local adjustments.

In summary, the claim of greater expressiveness is a direct reference to GenColor's learned ability to perform local, content-aware enhancements, a capability that is fundamentally more powerful than the global-only adjustment paradigm that defines the "Expert C" benchmark.

B.2 Intensity of Color Adjustment Filters in Training Data Generation

The color generation module's training paradigm uses the original high-quality image as the target and a color-adjusted version as the condition, simulating real-world enhancement scenarios. This approach, using the ARTISAN-1M dataset, enables aesthetically pleasing results even with suboptimal input colors. We carefully control color adjustment intensities in training data generation to balance enhancement quality and model versatility. As shown in Table 5, we constrain hue rotation to [0, 0.3] and [0.7, 1.0] to avoid drastic shifts while maintaining expressive adjustments. For all other filters, we utilize the full range to improve model expressiveness. This represents a trade-off between preserving natural appearances and allowing creative enhancements, especially for objects with ambiguous color properties.

B.3 GenColor Not Sensitive to Text Prompt

GenColor's color generation module demonstrates robustness to text prompts, as evidenced by the visual results in Figure 13. The model consistently produces similar outputs regardless of the prompt's content, including null, semantically incorrect, or color-incorrect prompts. This insensitivity to text

Empty Prompt | BLIP Output: There is a man that is standing on a rocky hill | Color Mismatch: There is a man that is standing on a green rocky hill | Semantic Error: There is a dog that is standing on the sea

Figure 13: GenColor's color generation module demonstrates robustness to various text prompts, consistently producing similar outputs regardless of prompt content (null, semantically incorrect, or color-incorrect). This insensitivity to text input ensures stable and predictable performance across diverse scenarios.

Input Image | Seed Variant 1 | Seed Variant 2 | Seed Variant 3 | Seed Variant 4

Figure 14: Visual demonstration of seed consistency. The model produces consistent outputs across different random seeds, with only minor variations in local details. The ensemble result further stabilizes the output.

input is advantageous for GenColor, as color adjustments are typically challenging to articulate verbally. Consequently, this characteristic ensures that GenColor's performance remains stable and predictable across various input scenarios.

B.4 GenColor Consistency Among Seeds

Our analysis of GenColor's sensitivity to random seeds reveals strong output consistency. Figure 14 demonstrates this visually, showing minimal variations across different seeds and a stable ensemble result. This consistency is crucial for practical applications, ensuring reliable outputs regardless of the random seed used.

B.5 Multi-Stage Weight Blending

We analyze the model's evolution during training, as visualized in Figure 15. Early stage outputs exhibit high saturation, middle stage outputs have the most appealing visual characteristics, and late stage outputs closely match the conditioning image. To balance aesthetic enhancement and texture preservation, we strategically combine model weights from the middle and late training stages during inference. This blending yields a good balance between creative expressiveness and artifact control, as measured by the ratio of regions with strong hue changes $R_{\Delta}\mathrm{H}$, and improves color consistency

Input (a) Early (b) Mid (c) Late (d) Mid+Late

Stage	$R_{\Delta H}\downarrow$	MS-SWD$_S\downarrow$
(a) Early	19.1%	3.33
(b) Middle	3.2%	1.60
(c) Late	2.2%	1.60
(d) M+L	1.3%	0.76

Figure 15: Evolution of model output during training: (a) input image; (b) early stage with unnatural colors; (c) middle stage with visually appealing results; (d) late stage matching conditioning image; (e) our proposed middle+late stage blending (M+L). M+L maintains visual quality while reducing strong hue changes and improving consistency across different generation seeds. The table shows reduced hue change regions ($R_{\Delta H}$) and improved color consistency (MS-SWD$_S$) compared to individual stages.

across different seeds, quantified by the mean pairwise MS-SWD [10] between generated images MS-SWD$_S$.

B.6 Comparing Weight Blending with Single Weight

We conducted an ablation study comparing different model checkpoints and their combinations to validate our weight mixing strategy. This includes the "mid" stage (high expressiveness), the "late" stage (high fidelity), a single checkpoint taken from between these two stages ("mid-late"), and our proposed 50/50 blend.

Table 6: Comparison of single checkpoints versus weight blending strategy. Our 50/50 blend of mid and late checkpoints significantly outperforms any single checkpoint, including the intermediate one, across all perceptual quality metrics while maintaining strong expressiveness.

Mid Weight	Mid-Late Weight	Late Weight	Q-Align↑	LAION↑	LIQE↑	NoR-VDP↑	C-VAR↑
1.0	0.0	0.0	4.08	5.66	3.54	68.96	22.91
0.0	1.0	0.0	4.24	5.77	3.72	68.87	17.87
0.0	0.0	1.0	4.25	5.77	3.73	69.09	16.69
0.5	**0.0**	**0.5**	**4.29**	**5.83**	**3.76**	**69.16**	**16.96**

The results in Table 6 demonstrate two key findings: (1) Simply selecting a single intermediate checkpoint ("Mid-Late Weight") is suboptimal compared to our blending strategy across all key quality metrics (e.g., Q-Align 4.24 vs 4.29), and (2) Our 50/50 blend achieves the highest scores on all four perceptual quality metrics (Q-Align, LAION, LIQE, NoR-VDP) while maintaining strong expressiveness (C-VAR). This makes the 50/50 blend a clear, data-driven choice for balancing aesthetic expressiveness with photorealistic quality, validating our weight mixing approach over single checkpoint selection.

C The Texture Preservation Module

C.1 Detailed Architecture of the Transformer-based Network

We present the detailed architecture of the Swin Transformer-based texture preservation module, designed for preserving textures while transfering the diffusion generated color to the input image.

Table 7: Detailed architecture of the SwinTransformer-based GenColor texture preservation module. B represents batch size. The network processes 1088×728 resolution images through 4 stages of feature extraction blocks (RSTB modules), each containing 6 Swin Transformer Blocks, with careful dimension preservation and feature processing throughout.

Stage	Output Shape	Block	Details
Input Conv	[B, 60, 728, 1088]	Conv2d	Conv2d-1: 6→60 channels, 3×3 kernel, stride=1, padding=1
Patch Embedding	[B, 792064, 60]	PatchEmbed	LayerNorm-2: Normalizes features; PatchEmbed-3: Embeds patches
RSTB Block ×4	[B, 792064, 60]	RSTB	Each RSTB contains 6 SwinTransformerBlocks with residual connections
Patch Unembedding	[B, 60, 728, 1088]	PatchUnEmbed	Linear layers (60→120→60 dims) with GELU activation, followed by LayerNorm and reshaping (pixel shuffling) back to image dimensions
Output Conv	[B, 3, 728, 1088]	Conv2d	Conv2d-2: 60→3 channels, 3×3 kernel, stride=1, padding=1

As illustrated in Table 7, the model begins with an input convolutional layer that expands the channel dimension from 6 to 60 using a 3×3 kernel. This is followed by a Patch Embedding stage, which flattens the spatial dimensions and applies Layer Normalization. The core of the network comprises four Residual Swin Transformer Blocks (RSTB Block ×4), each containing six Swin Transformer Blocks that integrate Window Attention mechanisms with six heads and a window size of 8, alongside Multi-Layer Perceptrons (MLP) with a ratio of 2. These blocks facilitate deep feature extraction while maintaining spatial dimensions. Subsequently, the Patch Unembedding stage reshapes the features back to the original image dimensions, and a final output convolutional layer reduces the channel dimension to 3, producing the restored RGB image. This architecture ensures meticulous dimension preservation and efficient feature processing, achieving robust performance in texture preserving color transfer scenarios.

C.2 Loss Functions and Their Weightings

The model is trained using a combination of Huber loss, perceptual loss, and adversarial loss:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Huber}} + \lambda_2 \mathcal{L}_{\text{Percept}} + \lambda_3 \mathcal{L}_{\text{adv\_G}} \tag{1}$$

where λ_1, λ_2, and λ_3 are the respective loss term weights.

- **Huber Loss**: We employ the Huber loss [15] to measure the discrepancy between the model prediction and the ground truth.
- **Perceptual Loss**: The perceptual loss is computed between the VGG-19 feature maps of the output and ground truth, enabling similarity judgments in the semantic feature space.
- **Adversarial Loss**: We employ an adversarial loss using a discriminator network D that distinguishes between real and fake images. The generator G tries to create realistic outputs to fool D. The objectives are:

$$\mathcal{L}_{\text{adv\_D}} = \mathbb{E}_{x \sim p_{\text{data}}}[-\log D(x)] + \mathbb{E}_{z \sim p(z)}[-\log(1 - D(G(z)))] \tag{2}$$

$$\mathcal{L}_{\text{adv\_G}} = \mathbb{E}_{z \sim p(z)}[-\log D(G(z))] \tag{3}$$

  In implementation, $\mathcal{L}_{\text{adv\_G}}$ uses a weighted binary cross-entropy loss. The total loss $\mathcal{L}_{\text{total}}$ is utilized in backward propagation to update the generator, while the discriminator is optimized separately to enhance its proficiency in distinguishing between real and generated data.

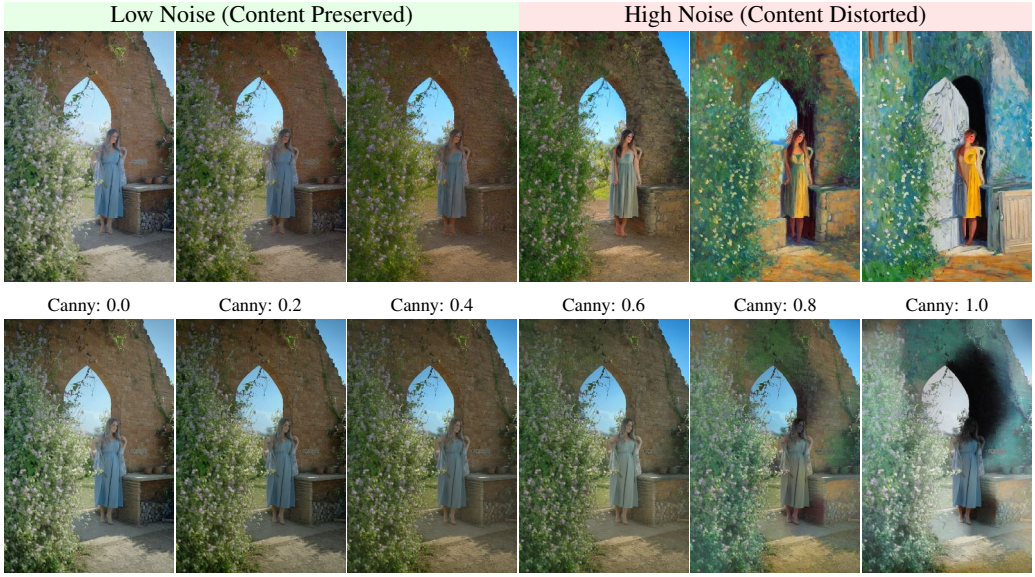


Figure 16: Effect of input-reference content similarity on texture preservation. We use Canny-based controlnet with GenColor's diffusion model at varying noise levels (0.0-1.0). Low noise values (0.0-0.4) maintain content similarity enabling effective texture preservation, while higher values (0.6-1.0) introduce distortions that compromise fusion quality.

C.3 Analyzing Content Affinity Requirements for Effective Texture Preservation

To rigorously evaluate content affinity requirements, we developed a dual-control methodology combining our first-stage color generation with Canny edge-based ControlNet conditioning. This integrated approach allows us to generate reference images while systematically varying the content similarity between input and reference pairs across a continuous range.

Our investigation demonstrates a strong relationship between texture preservation effectiveness and input-reference content similarity. Through systematic variation of Canny edge noise levels from 0.0 to 1.0 (as shown in Figure 16), we found that lower noise ranges (0.0-0.4) successfully maintain content affinity, enabling superior texture preservation in the fused results. Conversely, higher noise levels (0.6-1.0) introduce substantial content deviations that impair the fusion process.

It's important to note that in practical GenColor applications, only scenarios comparable to the first column of Figure 16 are relevant, as our color generation module operates without additional noise factors. This experimental setup effectively demonstrates our algorithm's specialization in high-fidelity color transfer while maintaining robustness across varying conditions.

C.4 Challenging Examples for Texture Preservation

As shown in Figure 17, we present challenging cases for texture preservation. The image features a complex outdoor scene with a tropical resort setting, our texture preservation module successfully maintains these fine details—from the delicate veining in individual leaves to the subtle features of the human subject's appearance and foot. In contrast, the diffusion-only reference exhibits significant degradation, with smudged vegetation textures, loss of definition in the human figure, and compromised detail in both facial features and foot. This example demonstrates GenColor's ability to achieve appealing color enhancement while preserving critical structural and textural elements that are essential for maintaining photorealistic image quality.

Input | Diffusion Ref | Texture Preservation Result

Figure 17: Challenging texture preservation case. Our method successfully preserves fine details in complex scenes while enhancing colors. The scene contains intricate textures from vegetation, architecture, and human subjects that the diffusion reference distorts while producing appealing colors. Our texture preservation module effectively maintains these critical structural details while transferring the enhanced color palette.

D The Global Filter Module

D.1 Detailed Implementation of the Global Filter Module

To enhance texture preservation performance, we incorporate a global filter module inspired by Harmonizer [17]. This lightweight yet effective component integrates seamlessly into the GenColor framework while preserving the sophisticated color transformations and texture details from previous stages.

The module employs five key filters (brightness, contrast, saturation, highlight, and shadow) that perform targeted adjustments to enhance image quality. A lightweight EfficientNet image encoder extracts features from which filter intensities are computed via a simple MLP network trained end-to-end with the rest of the model.

For training, we utilize cleaned versions of the Adobe5K and HDRPlus datasets, ensuring no overlap between training and test data. The global filters are trained for 100k steps with batch size 32, learning rate 1e-4, and Adam optimizer.

The implementation details of each filter are identical to those in Harmonizer, with each filter designed to preserve critical texture information while enhancing specific aspects of the image:

D.1.1 Brightness Filter

The brightness filter operates in HSV color space to modify the value (V) channel while preserving hue and saturation:

$$\text{image}_{\text{RGB}} \rightarrow \text{image}_{\text{HSV}} = (h, s, v)$$

For a brightness parameter $x \in [-1, 1]$, we compute an alpha multiplier:

$$\alpha = \begin{cases} 1/(1 - x + \epsilon), & \text{if } x \geq 0 \\ x + 1, & \text{if } x < 0 \end{cases}$$

The adjusted value channel becomes $v' = v \times \alpha$, and the final image is converted back to RGB:

$$\text{image}_{\text{RGB}} = \text{HSV\_to\_RGB}(h, s, v')$$

D.1.2 Contrast Filter

The contrast filter enhances differences relative to the mean luminance. For a contrast parameter $x \in [-1, 1]$:

$$\text{threshold} = \text{mean}(\text{image})$$
$$x' = \begin{cases} 255/(256 - \lfloor x \times 255 \rfloor) - 1, & \text{if } x > 0 \\ x, & \text{otherwise} \end{cases}$$

The contrast adjustment is then applied as:

$$\text{image}' = \text{image} + (\text{image} - \text{threshold}) \times x'$$

D.1.3 Saturation Filter

The saturation filter modifies color intensity while preserving luminance. First, we compute:

$$\begin{aligned} c_{\text{min}} &= \min(\text{image}_R, \text{image}_G, \text{image}_B) \\ c_{\text{max}} &= \max(\text{image}_R, \text{image}_G, \text{image}_B) \\ \text{var} &= c_{\text{max}} - c_{\text{min}} \\ \text{ran} &= c_{\text{max}} + c_{\text{min}} \\ \text{mean} &= \text{ran}/2 \end{aligned}$$

The saturation factor s is calculated as:

$$s = \begin{cases} \text{var}/(\text{ran} + \epsilon), & \text{if mean} < 0.5 \\ \text{var}/(2 - \text{ran} + \epsilon), & \text{otherwise} \end{cases}$$

For a saturation parameter $x \in [-1, 1]$, we compute the adjustment factor:

$$m = \begin{cases} 1, & \text{if } (x + s) > 1 \\ 0, & \text{otherwise} \end{cases}$$
$$a = \begin{cases} 1/(s \times m + (1 - x) \times (1 - m) + \epsilon) - 1, & \text{if } x \geq 0 \\ 1 + x, & \text{if } x < 0 \end{cases}$$

The final adjustment is:

$$\text{image}' = \begin{cases} \text{image}, & \text{if } x \geq 0 \\ \text{mean} + (\text{image} - \text{mean}) \times a, & \text{otherwise} \end{cases}$$

D.1.4 Highlight and Shadow Filters

These filters adjust the bright and dark regions respectively:

Highlight Filter (for parameter $x \in [-1, 1]$):

$$x' = x + 1$$
$$\text{image}' = \text{invert}(\text{clamp}((\text{invert}(\text{image}))^{x'}, 0, 1))$$

Shadow Filter (for parameter $x \in [-1, 1]$):

$$x' = -x + 1$$
$$\text{image}' = \text{clamp}(\text{image}^{x'}, 0, 1)$$

E Experimental Setup

E.1 Justification for Metric Selection

Given the inherent subjectivity in color enhancement and texture preservation tasks, we adopt a comprehensive suite of both classical and contemporary evaluation metrics.

For image quality assessment, our approach integrates perceptual measures such as **Q-Align** [32], **LIQE** [37], and **LAION-Aesthetics V2** [30], plus **NoR-VDP** [8] to handle no-reference dynamic range evaluation. Furthermore, **C-VAR** is employed to quantify the expressiveness or color variation introduced by an enhancement.

For texture preservation evaluation, we rely on established metrics like **TD** [6] (Texture Difference), **GMSD** [33] (Gradient Magnitude Similarity Deviation), and **DISTS** [5] (Deep Image Structure and Texture Similarity) to ensure structural fidelity. For color transfer assessment, we employ advanced color distribution measures including **Semantic** W_1 (Wasserstein-1 distance for regional color distribution), standard W_1 [9] (Wasserstein-1 distance between global color distributions), and **MS-SWD** [10] (Multi-Scale Sliced Wasserstein Distance) to capture alignment in color space.

By leveraging these well-validated metrics across different aspects of enhancement, our multi-metric strategy offers a robust quantitative evaluation of both enhancement quality and preservation accuracy.

E.2 Detailed Description of Evaluation Metrics

We use a comprehensive set of metrics to evaluate the performance of GenColor and other methods:

- **Color Enhancement Quality Metrics**
  - **Q-Align** [32] (↑): A non-reference perceptual quality metric that evaluates image aesthetic quality without requiring expert-retouched references, considering both color and structural aspects.
  - **LAION** (↑): An aesthetic score from the LAION database, reflecting how well the enhanced result aligns with its dataset's learned aesthetic preferences.
  - **LIQE** [37] (↑): A no-reference image quality assessment framework that employs multi-task learning to jointly optimize scene classification and distortion identification tasks, enabling robust perceptual quality prediction without requiring reference images.
  - **NoR-VDP** [8] (↑): A No-Reference Visual Difference Predictor specifically designed for high dynamic range content that assesses image quality without requiring a reference image, focusing on natural image statistics, dynamic range perception, and human visual system response to luminance variations.
  - **C-VAR** (↑): A measure capturing the "expressiveness" or color variation introduced by the enhancement. Higher values indicate more fine-grained or diverse transformations.
- **Texture Preservation Metrics**
  - **TD** [6] (↓): Texture Difference score that quantifies the preservation of high-frequency details between input and enhanced images.

  - **GMSD** [33] (↓): Gradient Magnitude Similarity Deviation that compares gradient magnitude maps.
  - **DISTS** [5] (↓): Deep Image Structure and Texture Similarity metric that assesses structural and textural similarity.
- **Color Similarity Metrics**
  - **Semantic** W_1 (↓): Wasserstein-1 distance for regional color distribution, measuring color similarity for semantic regions.
  - W_1 (↓) [9]: Standard Wasserstein-1 distance for global color distribution.
  - **MS-SWD** [10] (↓): Multi-Scale Sliced Wasserstein Distance that compares color distributions across different spatial scales.

For all metrics, (↑) indicates higher is better, while (↓) indicates lower is better.

E.3 C-VAR Metric Details

C-VAR (**C**olor-**VAR**iance Ratio) is a superpixel-based metric that estimates the *expressiveness* of an image enhancement by contrasting *how consistently each region is shifted* with *how differently* between *regions those shifts behave*. Higher C-VAR values therefore indicate strong, region-aware color edits that remain locally coherent.

The C-VAR score is obtained through the following procedure:

1. **Superpixel Segmentation**: Segment the *original* RGB image I into $N \approx 200$ perceptually cohesive regions using SLIC (`n_segments=200`, `compactness=10`). The same label mask is reused for its enhanced counterpart I'.
2. **Color-Space Transformation & Difference Map**: Convert both images to CIELAB. For every pixel p compute the Lab shift $\Delta\mathbf{c}_p = \mathrm{Lab}(I'_p) - \mathrm{Lab}(I_p) \in \mathbb{R}^3$.
3. **Per-Superpixel Statistics**: For each superpixel s_i ($i = 1, \ldots, N$) calculate
   - the *mean* color shift $\boldsymbol{\mu}_i = \dfrac{1}{|s_i|} \displaystyle\sum_{p \in s_i} \Delta\mathbf{c}_p$,
   - the *within-region covariance* $\boldsymbol{\Sigma}_i = \mathrm{Var}_{p \in s_i}\big(\Delta\mathbf{c}_p\big)$.
4. **Variance Aggregation**:
   - *Intra-class variance* (local inconsistency)
   $$\sigma_{\text{intra}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{tr}\big(\boldsymbol{\Sigma}_i\big).$$
   - *Inter-class variance* (regional diversity)
   $$\sigma_{\text{inter}} = \mathrm{tr}\Big(\mathrm{Var}_{i=1}^{N}\big(\boldsymbol{\mu}_i\big)\Big).$$
5. **C-VAR Score**:
   $$\text{C-VAR} = \frac{\sigma_{\text{inter}}}{\sqrt{\sigma_{\text{intra}} + \varepsilon}}, \qquad \varepsilon = 10^{-6}.$$

C-VAR captures two critical traits of a color edit:

- **Regional Diversity** (σ_{inter}): large variations between superpixels signal expressive, context-aware adjustments.
- **Local Coherence** (σ_{intra}): small variations within a superpixel ensure the edit is not noisy or spatially erratic.

By normalising inter-class variance with the square-root of intra-class variance, C-VAR becomes scale-invariant and rewards transformations that respect object boundaries while diversifying color themes—making it particularly suitable for evaluating style-transfer, HDR tonemapping, and other generative color-enhancement tasks.

E.4 Large Language Model (LLM) Preference Evaluation

To evaluate the perceptual quality of our color enhancement approach, we conducted a LLM preference evaluation using GPT-o1, a state-of-the-art vision-language model. This evaluation provides an additional perspective on aesthetic quality that complements our quantitative metrics.

For each round, we presented GPT-o1 with three image versions in random sequence:

1. The original unprocessed input image
2. The expert-retouched reference image (Expert C)
3. Our GenColor enhanced result

We prompted the model with the following instruction:

> "You are a professional photo retoucher specializing in color grading. I will provide 3 photos with distinct color treatments. Analyze each photo, clearly explain your reasoning regarding color aesthetics, mood, harmony, and visual appeal, and clearly indicate your choice of which photo is more visually beautiful from a color perspective. You must make a definitive choice."

Due to the budget limit of GPT-o1, we only evaluate the first 100 images in Adobe5K. As shown in Table 8, GenColor is preferred over human expert retouching.

Table 8: LLM preference evaluation results. Higher is better.

Method	LLM Preference↑
Original Image	0
Human Expert C	36
GenColor (Ours)	**64**

E.5 Baseline Methods Implementations

For all baseline methods, we rely on their official implementations and pre-trained weights. When dataset-specific weights are available, we adopt them for the corresponding experiments; otherwise, we default to models trained on the Adobe FiveK dataset. In the case of Distort-and-Recover [28], which does not provide pre-trained weights, we replicate the authors' training pipeline using the Adobe FiveK dataset. Notably, the official Distort-and-Recover implementation only supports low-resolution images (224×224), limiting its applicability to smaller inputs.

E.6 Hyperparameter Settings and Training Procedures

E.6.1 Color Generation Module

Our implementation leverages the official ControlNet [36] with Stable Diffusion 2.1 serving as the base model. Training is conducted at 512x512 resolution with an AdamW optimizer using a learning rate of 1e-5 and a batch size of 8. We train two model checkpoints - one for 450K iterations (middle stage) and another for 700K iterations (late stage). At inference time, we utilize 30 denoising steps and a classifier-free guidance scale of 7.5. The positive prompt is automatically generated by BLIP [19], and the negative prompt is "low quality, bad quality, low contrast, low saturation, dark, black and white, color bleeding, bw, monochrome, grainy, blurry, historical, restored, desaturate". As detailed in our main manuscript, we find that combining predictions from both middle-stage and late-stage checkpoints yields superior results.

E.6.2 Texture Preservation Module

For training the texture preservation module with 256×256 resolution data, we employ a combination of loss functions to ensure both low-level accuracy and high-level perceptual quality. The loss formulation combines a Huber loss ($\lambda = 0.05$) for pixel-wise supervision, a VGG-based perceptual loss ($\lambda = 1.0$) for maintaining semantic consistency, and an adversarial loss term ($\lambda = 0.05$) for enhancing output realism. We optimize the network using AdamW with an initial learning rate of

1e-4 that follows a cosine annealing schedule decreasing to 1e-6. The model trains for 1M iterations with a batch size of 4.

E.7 Hardware and Software Specifications

All experiments were conducted using PyTorch framework on a NVIDIA V100 GPU with 32GB memory. To maximize training efficiency, we employed an optimized data loading pipeline that leverages both CPU prefetching and GPU-accelerated processing.

F Additional Experimental Results

F.1 Additional Results on Color Transfer Compare with Other Methods

Tables 9, 10 and 11 present a comprehensive comparison of GenColor against state-of-the-art color transfer methods across 3 benchmark datasets: Adobe5K, PPR10K and FreeRaw. Our evaluation focuses on two critical aspects:

1. **Texture Preservation**: We measure how well each method maintains the original image structure using three complementary metrics:
   - TD (Texture Difference): Quantifies pixel-level structural changes
   - DISTS (Deep Image Structure and Texture Similarity): Evaluates perceptual texture similarity
   - GMSD (Gradient Magnitude Similarity Deviation): Assesses edge and gradient preservation
2. **Color Similarity**: We evaluate color transfer accuracy using three distribution-based metrics:
   - Semantic W_1: Wasserstein-1 distance for regional color distribution
   - W_1: Standard Wasserstein-1 distance for global color distribution
   - MS-SWD: Multi-Scale Sliced Wasserstein Distance for color space alignment

The results demonstrate that GenColor achieves an exceptional balance between preserving original textures and transferring accurate colors. Notably, GenColor achieves competitive texture preservation scores while dramatically outperforming all baselines in color similarity metrics.

While methods like Color Matcher excel in specific texture metrics, they significantly underperform in color transfer accuracy. Conversely, while other methods achieve similar texture preservation ability with GenColor, they struggle with overall color transfer accuracy. GenColor's balanced performance across all metrics underscores its effectiveness as a comprehensive color enhancement solution.

Table 9: Texture preservation and color similarity analysis on Adobe5K. GenColor achieves excellent balance between preserving textures and maintaining color accuracy. All metrics are lower is better. Bold values indicate best performance.

Method	Texture Preservation			Color Similarity		
	TD ↓	DISTS ↓	GMSD ↓	Sem W_1 ↓	W_1 ↓	MS-SWD ↓
NLUT	0.95	0.15	**0.12**	13.01	12.76	0.87
Color Matcher	**0.42**	0.17	0.14	31.45	31.34	2.03
UniColor	0.95	0.15	0.14	38.40	38.20	2.18
Deep Preset (w/o PPL)	1.17	0.15	**0.12**	33.09	32.88	2.38
Deep Preset	1.12	0.15	0.13	34.84	34.64	2.51
GenColor (Ours)	0.94	**0.13**	**0.12**	**3.59**	**3.22**	**0.70**

F.2 Analyzing Sensitivity to Local Color Fine Details Compared with Color Transfer Methods

GenColor demonstrates superior precision in transferring local color details compared to existing color transfer methods. As shown in Figure 18, our approach accurately preserves and enhances the unique color characteristics of each individual object within complex scenes, maintaining distinct color identities even in areas with intricate patterns and fine boundaries.

Table 10: Texture preservation and color similarity evaluation on PPR10K portrait dataset. GenColor demonstrates superior performance in both texture metrics and color similarity. All metrics are lower is better. Bold values indicate best performance.

Method	Texture Preservation			Color Similarity		
	TD ↓	DISTS ↓	GMSD ↓	Sem W_1 ↓	W_1 ↓	MS-SWD ↓
NLUT	0.46	0.10	0.12	13.81	13.49	0.87
Color Matcher	**0.23**	0.13	0.13	36.23	36.07	2.06
UniColor	0.48	0.13	0.13	44.10	43.86	2.37
Deep Preset (w/o PPL)	0.56	0.14	0.12	40.95	40.73	2.65
Deep Preset	0.53	0.14	0.12	38.77	38.54	2.53
GenColor (Ours)	0.50	**0.09**	**0.11**	**4.20**	**3.91**	**0.65**

Table 11: Performance comparison on high-quality FreeRaw dataset. GenColor achieves the best texture preservation and color similarity scores, demonstrating its effectiveness on high-quality raw images. All metrics are lower is better. Bold values indicate best performance.

Method	Texture Preservation			Color Similarity		
	TD ↓	DISTS ↓	GMSD ↓	Sem W_1 ↓	W_1 ↓	MS-SWD ↓
NLUT	0.69	0.13	0.12	14.21	13.73	0.92
Color Matcher	**0.38**	0.17	0.14	35.35	35.10	2.13
UniColor	0.57	0.15	0.14	43.19	42.68	2.35
Deep Preset (w/o PPL)	0.91	0.15	0.12	37.54	37.05	2.51
Deep Preset	0.90	0.16	0.12	37.68	37.16	2.56
GenColor (Ours)	0.80	**0.12**	**0.11**	**3.73**	**3.31**	**0.70**

Color Matcher applies transformations based primarily on global color statistics, often overlooking local color variations that are crucial for realistic enhancement. This results in homogenized colors that fail to respect object boundaries and fine details. While NLUT and Deep Preset methods attempt to address this limitation through more sophisticated transformations, they still struggle with preserving the distinct color identity of individual objects, particularly in complex scenes with multiple elements.

UniColor operates at a patch-level granularity, which represents an improvement over global methods but remains insufficient for truly fine-grained color transfer. Additionally, its approach of fixing the L channel (luminance) while only modifying the a and b channels in the Lab color space limits its ability to perform comprehensive color enhancements that might require luminance adjustments.

In contrast, GenColor's two-stage approach offers several advantages for preserving local color details:

1. The Color Generation module creates reference images that maintain spatial coherence with the input, ensuring that color transformations respect object boundaries and local contexts.

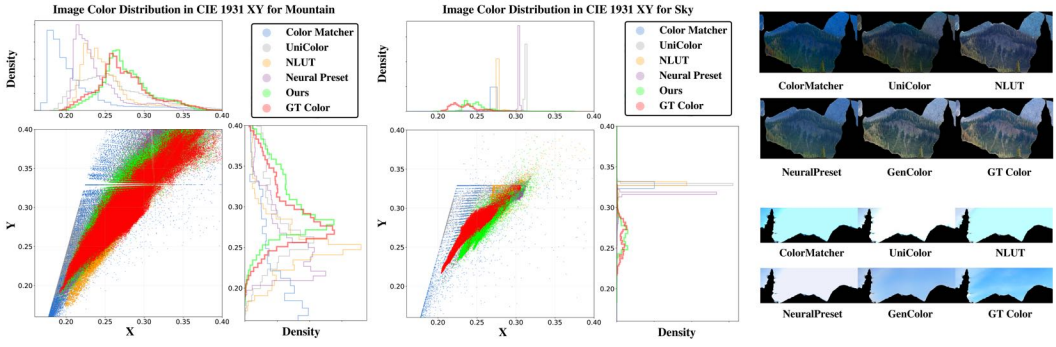


Figure 18: Comparison of color transfer methods on objects with fine color details.

2. Our Texture Preservation module, with its specialized degradation strategy, maintains precise alignment between the input and reference images, preserving fine-grained color details while transferring the enhanced color palette.
3. The transformer-based architecture in our texture preservation network excels at capturing long-range dependencies, allowing it to maintain consistency across similar objects while respecting their unique color characteristics.

This sensitivity to local color details is particularly evident in scenes with multiple objects of varying colors and textures. GenColor successfully preserves these nuances while applying aesthetically pleasing enhancements, resulting in more realistic and visually appealing results that outperform existing approaches.

F.3 Ablation Study of Degradation on Texture Preservation Module

The additional datasets results are shown in Tables 12, 13 and 14. we demonstrate the effectiveness of our texture preservation module. The results are similar to those in the main paper.

Our ablation study in Table 12 reveals the importance of each component in our degradation strategy. As further illustrated in Figure 19, each degradation technique contributes uniquely to the final result. Notably, circular shift plays a crucial role in texture preservation, as removing it significantly increases the TD score from 0.94 to 1.37, indicating substantial texture degradation. This is visually evident in the "w/o Circular Shift" version exhibits noticeable artifacts around the nose area, compromising the overall texture quality.

Meanwhile, stroke noise and random resize primarily influence color similarity metrics, with minimal impact on texture preservation scores. When these components are removed, we observe increased Semantic Wasserstein-1 and Wasserstein-1 distances, suggesting poorer color distribution matching. This is confirmed in where both the "w/o Random Resize" and "w/o Stroke Noise" variants display higher color deviation, particularly in the face and hair regions compared to the full model.

The Diffusion Reference demonstrates why texture preservation is essential - note the misalignment with the input image, particularly in the eyeball position and clothes pocket details. Overall, our full degradation strategy achieves the best balance, with optimal performance in color similarity metrics (Sem W_1, W_1, and MS-SWD) while maintaining competitive texture preservation scores, resulting in the most visually coherent enhanced image.

Figure 19: Visual comparison of different degradation strategies. The Diffusion Ref shows texture misalignment with the input (note eyeball and pocket positions). *w/o Random Resize* and *w/o Stroke Noise* introduce colour drift in the face and hair, while *w/o Circular Shift* yields artefacts around the nose, highlighting the importance of each component.

F.4 Ablation Study on Different Components

In this section, we conduct a comprehensive ablation study to evaluate the contribution of each component in our architecture. Tables 15, 16 and 17 present the results on Adobe5K, PPR10K and FreeRaw datasets, respectively.

Table 12: Texture preservation technique analysis on Adobe5K dataset. Results show how different components affect texture quality and color similarity. All metrics are lower is better. Bold values indicate best performance.

Degradation	Texture Preservation			Color Similarity		
	TD ↓	DISTS ↓	GMSD ↓	Sem W_1 ↓	W_1 ↓	MS-SWD ↓
w/o Random Resize	**0.76**	0.14	0.12	6.21	5.96	0.78
w/o Stroke Noise	0.77	0.14	0.12	5.74	5.44	0.78
w/o Circular Shift	1.37	0.14	0.12	15.49	15.40	1.20
Full	0.94	**0.13**	0.12	**3.59**	**3.22**	**0.70**

Table 13: Analysis of texture preservation techniques on PPR10K portrait dataset. Study reveals impact of different operations on texture and color. All metrics are lower is better. Bold values indicate best performance.

Degradation	Texture Preservation			Color Similarity		
	TD ↓	DISTS ↓	GMSD ↓	Sem W_1 ↓	W_1 ↓	MS-SWD ↓
w/o Random Resize	**0.36**	0.10	0.11	5.98	5.69	0.76
w/o Stroke Noise	0.37	0.10	0.11	5.23	4.87	0.78
w/o Circular Shift	0.76	**0.09**	0.11	12.69	12.57	0.99
Full	0.50	**0.09**	0.11	**4.20**	**3.91**	**0.65**

Our analysis reveals several key insights:

1. The Color Generation module (C) significantly enhances overall color quality and expressiveness, as evidenced by the substantial improvements in C-VAR scores. When this module is included, we observe a 65.8% increase in C-VAR on Adobe5K (from 2.84 to 4.71) and a 10.0% improvement on PPR10K (from 3.10 to 3.41). Additionally, the Color Generation module contributes to better perceptual quality, shown by consistent improvements in Q-Align and LAION scores across both datasets.
2. The Texture Preservation module (T) plays a crucial role in maintaining the original image details while allowing for color transformation. This is demonstrated by the significantly lower TD scores when this module is present. Without compromising on enhancement quality, the texture module ensures that important structural elements of the image remain intact.
3. The Global Filter (G) further refines the overall image quality by providing global adjustments that complement the local operations of the other modules. While its individual contribution may appear subtle in some metrics, its presence in the full model helps achieve a balanced performance across all evaluation criteria.

The full model, which integrates all three components, achieves the most balanced performance across the diverse set of metrics. This confirms that each component addresses a specific aspect of the enhancement process, and their combination yields a robust and versatile enhancement system. As shown in Figure 20, each component contributes uniquely to the final result, with the full model producing the most visually appealing output.

Table 14: Evaluation of texture preservation techniques on high-quality FreeRaw dataset. Results show how different strategies affect performance, with the full model achieving optimal balance. All metrics are lower is better. Bold values indicate best performance.

Degradation	Texture Preservation			Color Similarity		
	TD ↓	DISTS ↓	GMSD ↓	Sem W_1 ↓	W_1 ↓	MS-SWD ↓
w/o Random Resize	**0.57**	0.13	0.12	6.30	5.83	0.77
w/o Stroke Noise	0.59	0.13	0.12	5.70	5.28	0.76
w/o Circular Shift	1.21	**0.11**	**0.11**	14.64	14.55	1.09
Full	0.80	0.12	**0.11**	**3.73**	**3.31**	**0.70**

Input | Global Filter Only | w/o Global Filter | Full

Figure 20: Visual comparison of different module configurations. The Global Filter alone provides basic adjustments but lacks expressiveness and cannot achieve high brightness levels. Without the Global Filter, the image shows improved color but insufficient contrast. The full model combines the strengths of all components, delivering optimal brightness, contrast, and color expressiveness for the most visually appealing result.

Table 15: Ablation study on Adobe5K dataset analyzing each component's contribution. C: Color Generation, T: Texture Preservation, G: Global Filter. TD (lower = better), Q-Align, LAION, LIQE, NoR-VDP, and C-VAR (all higher = better). Best values in **bold**.

Architecture			**Metrics**					
C	**T**	**G**	TD↓	Q-Align↑	LAION↑	LIQE↑	NoR-VDP↑	C-VAR↑
		✓	**0.47**	4.09	5.60	3.47	66.26	10.35
✓		✓	2.75	**4.32**	**5.90**	3.17	62.19	13.24
✓	✓		0.99	4.26	5.77	3.70	**70.21**	12.02
✓	✓	✓	1.15	4.29	5.83	**3.76**	69.16	**16.96**

F.5 Standard Deviation for all Metrics on the Adobe5K Dataset

The standard deviation analysis reveals an important finding about performance consistency. For most metrics (Q-Align, LAION, LIQE), the standard deviations across all top-performing methods, including ours, are highly comparable, indicating similar levels of performance consistency. While some methods like D&R show lower variance on certain metrics, they have significantly lower mean performance. In contrast, GenColor not only achieves the highest mean scores but does so with variance that is on par with, or even lower than (e.g., NoR-VDP), other leading approaches.

F.6 Denoising Steps Analysis

To optimize the computational efficiency of GenColor while maintaining high-quality results, we conducted a comprehensive analysis of the relationship between the number of denoising steps and performance metrics. The full results on the Adobe5K dataset are presented in Table 19.

Table 16: Ablation study on PPR10K portrait dataset analyzing each component's contribution. C: Color Generation, T: Texture Preservation, G: Global Filter. TD (lower = better), Q-Align, LAION, LIQE, NoR-VDP, and C-VAR (all higher = better). Best values in **bold**.

Architecture			**Metrics**					
C	**T**	**G**	TD↓	Q-Align↑	LAION↑	LIQE↑	NoR-VDP↑	C-VAR↑
		✓	**0.20**	4.21	6.13	2.98	68.27	12.92
✓		✓	1.68	**4.45**	**6.35**	2.81	66.23	12.83
✓	✓		0.53	4.36	6.24	3.13	**71.24**	10.79
✓	✓	✓	0.58	4.38	6.26	**3.21**	71.11	**13.57**

Table 17: Ablation study on the high-quality FreeRaw dataset analyzing each component's contribution. C: Color Generation, T: Texture Preservation, G: Global Filter. TD (lower = better), Q-Align, LAION, LIQE, NoR-VDP, and C-VAR (all higher = better). Best values in **bold**.

Architecture			Metrics					
C	**T**	**G**	TD↓	Q-Align↑	LAION↑	LIQE↑	NoR-VDP↑	C-VAR↑
		✓	**0.29**	4.36	5.77	3.94	63.18	9.62
✓		✓	2.42	**4.53**	**6.03**	3.51	60.46	13.97
✓	✓		0.85	4.47	5.89	4.15	**68.19**	10.94
✓	✓	✓	0.96	4.51	5.96	**4.22**	67.03	**14.93**

Table 18: Standard deviation analysis for all objective metrics across the Adobe5K dataset. Lower standard deviation indicates more consistent performance across the dataset. Our method achieves competitive consistency while maintaining superior mean performance.

Method	**Std Q-Align↓**	**Std LAION↓**	**Std LIQE↓**	**Std NoR-VDP↓**	**Std C-VAR↓**
3D-LUT	0.43	0.91	0.91	4.07	6.81
RSFNet	0.43	0.91	0.91	3.84	6.61
DeepLPF	0.45	0.92	0.92	3.72	6.50
ICELUT	0.45	0.92	0.91	4.17	6.64
D&R	**0.31**	**0.83**	**0.45**	6.33	9.24
D&R (ARTISAN)	**0.29**	0.85	**0.42**	6.31	9.07
Exposure	0.45	0.92	0.85	4.58	8.63
Exposure (ARTISAN)	0.44	0.91	0.84	4.67	8.52
GenColor (Ours)	0.43	0.93	0.84	**3.45**	8.18

Key Finding: Our analysis shows that while the performance is low with very few steps (1-5), the quality rapidly converges. We can reduce the number of denoising steps from 30 down to 15—effectively halving the runtime of our most computationally intensive module—with no obvious degradation in quality across all key metrics. Hence, the performance at 15 steps is virtually identical to that at 30 steps. This offers a favorable speed-quality trade-off, making the method more practical for real-world applications.

Future Work and Broader Context: Beyond this immediate optimization, the efficiency of GenColor can be further improved by incorporating techniques from the rapidly advancing field of fast diffusion sampling. Methods such as knowledge distillation (e.g., Progressive Distillation) or consistency models could potentially reduce the step count to as few as 1-4 steps while maintaining high quality, and we consider this as a promising direction for future work.

G Additional User Study: Pairwise Comparison Study

To further validate our findings and provide additional statistical rigor, we conducted an additional, dedicated randomized pairwise comparison (2AFC) experiment.

Methodology: We recruited 25 participants, and each completed 20 comparison trials. In each trial, the system randomly selects two of the seven methods for a head-to-head comparison on a given image. Participants then choose the more visually appealing result. This process produces a total of 500 pairwise judgments.

Results: We analyzed the data using the Bradley-Terry model to derive a preference score for each method. The results show that GenColor achieved the highest preference score, significantly outperforming all baselines. The complete ranking is presented in Table 20.

Crucially, the final ranking of all methods from this new pairwise study is highly consistent with the ranking derived from our original side-by-side rating study. This agreement between two different comparative methodologies provides strong support for our conclusions.

Table 19: Analysis of denoising steps vs. performance on Adobe5K dataset. All metrics show higher is better except where noted. Performance rapidly converges after 15 steps, enabling a favorable speed-quality trade-off.

Denoising Steps	Q-Align↑	LAION↑	LIQE↑	NoR-VDP↑
1	2.75	4.86	1.91	68.75
2	2.84	4.92	2.01	68.95
3	3.88	5.42	3.30	69.09
5	4.17	5.67	3.61	69.20
10	4.28	5.80	3.74	69.19
15	4.29	5.82	3.75	69.18
20	4.29	5.83	3.76	69.15
25	4.29	5.83	3.76	69.12
30	4.29	5.83	3.76	69.16

Table 20: Pairwise comparison results using Bradley-Terry model. GenColor achieves the highest preference score among all methods.

Rank	Method	Bradley-Terry Score↑
1	GenColor (Ours)	100.00
2	RSFNet	72.10
3	DeepLPF	67.07
4	ICELUT	58.00
5	3D-LUT	52.31
6	Exposure	25.92
7	D&R	17.18

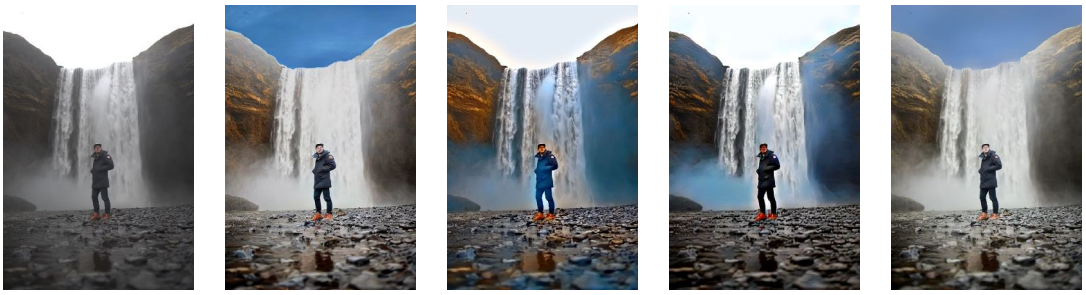
H Additional Comparison with Style Transfer Methods

To provide a comprehensive evaluation of our approach, we visually compare GenColor against state-of-the-art style transfer methods, specifically CAPVST [31] and StyA2K [38]. While these methods differ fundamentally from color matching/color style transfer approaches in that they modify both color and textural elements during style transfer, this comparison helps demonstrate the unique capabilities and advantages of our color-focused enhancement pipeline. As shown in Figure 21, our method achieves more accurate color transfer while preserving the original image content.

I Additional Comparison with Commercial Color Transfer Tools: NeuralPreset and Adobe Photoshop

We further compare our method with NeuralPreset [16], a state-of-the-art color style transfer approach. While the official implementation is not publicly available, we evaluate against Tinge, an application that implements the same methodology. Additionally, we benchmark against Adobe Photoshop's color matching functionality, which serves as an industry standard baseline. Due to the manual nature of generating results with these tools, we conduct this comparison on a curated set of 20 representative images that span diverse scenarios and color distributions.

Table 21 presents a comprehensive quantitative comparison across three key aspects: texture preservation, quality, and color similarity. Our method, GenColor, achieves superior performance in most metrics. Notably, while Adobe PhotoShop shows strong texture preservation (TD: 0.801) and Neural Preset excels in DISTS (0.112), GenColor maintains competitive performance in these aspects while significantly outperforming in quality metrics (Q-Align: 0.023, LIQE Gain: 0.704, NoR-VDP Gain: 8.540). Most importantly, GenColor demonstrates substantially better color similarity scores across all metrics (Semantic W_1: 4.544, W_1: 4.221, MS-SWD: 0.647), indicating more accurate color transfer while preserving image structure.



Input | Diffusion Ref | CAPVST [31] | StyA2K [38] | GenColor (Ours)

Figure 21: Comparison with style transfer methods. While style transfer methods modify both color and texture, our method focuses on color enhancement while preserving the original image content.

Table 21: Comparison results of commercial applications on texture preserving color transfer performance. Metrics are categorized into Texture Preservation and Color Similarity. Best results are **bolded** and second-best are in regular text.

Method	Texture Preservation			Color Similarity		
	TD ↓	DISTS ↓	GMSD ↓	Sem W_1 ↓	W_1 ↓	MS-SWD ↓
Adobe PhotoShop	**0.80**	0.14	0.13	12.19	11.18	0.95
Neural Preset	1.60	**0.11**	0.13	10.97	9.78	0.86
GenColor (Ours)	0.97	0.13	**0.13**	**4.54**	**4.22**	**0.65**

J Limitations and Future Work

J.1 Failure Cases

Despite the overall effectiveness of GenColor, we observe occasional limitations in two specific scenarios. First, as demonstrated in Figure 23, the method exhibits reduced performance when processing images containing substantial regions of pure black and white values. This limitation stems from the inherent difficulty in meaningfully enhancing regions with extreme luminance values while maintaining natural appearance.

Second, as illustrated in Figure 24, the model may produce unintended hue transformations when processing scenes containing extreme illumination conditions or objects with perceptually ambiguous colors. While these outputs maintain aesthetic appeal, they might deviate from the desired color enhancement objectives. We note that this limitation can often be mitigated by utilizing alternative seeds in the Diffusion Generator (*e.g.*, the blue and purple clothes variations in Figure 14 first row). For future iterations, we propose incorporating a hue preservation mask as an additional conditioning signal during both training and inference to provide more precise control over color transformations.

J.2 The Paradox of Deterministic Output from Diverse Data

Your observation is accurate. The nearly deterministic output stems from the optimization dynamics of our current training objective.

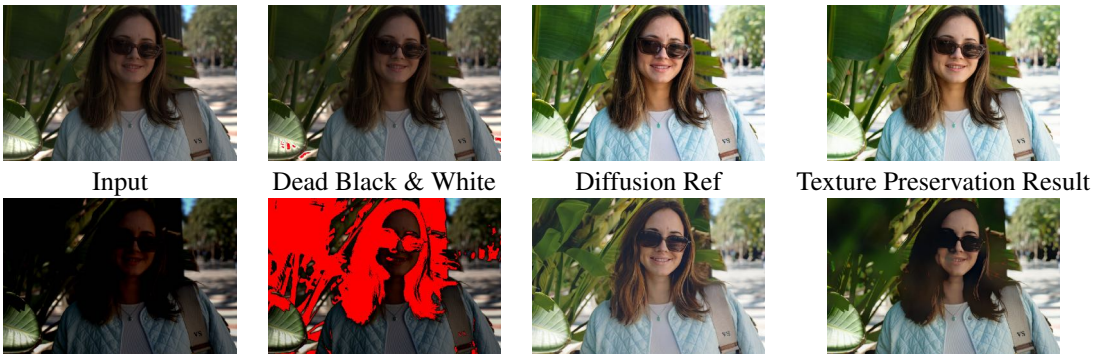
While ARTISAN's data distribution is highly diverse and multi-modal (representing many styles), the model is strongly conditioned on the input image via ControlNet. The training process is optimized to find the most probable, high-quality enhancement. This encourages the model to converge towards a high-probability density region within the distribution—often resembling a "mean" or dominant aesthetic mode. The current objective function prioritizes a robust, broadly appealing automatic solution rather than diversity of output.

J.2.1 Implications and Future Directions

As you rightly suggest, this realization opens up significant avenues for future work to build upon the GenColor foundation.

Input | Diffusion Ref | Adobe PhotoShop | Neural Preset [16] | GenColor (Ours)

Figure 22: Visual comparison with commercial applications. Our method achieves more accurate color transfer.



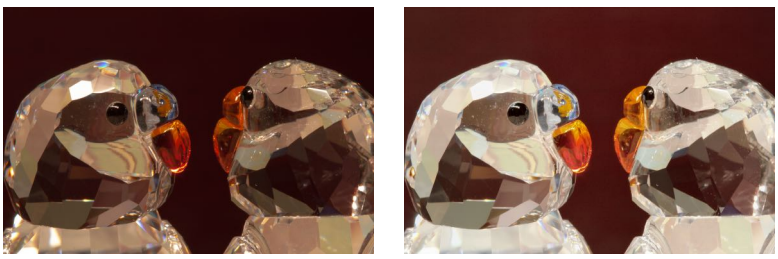
Input | Dead Black & White | Diffusion Ref | Texture Preservation Result

Input | Dead Black & White | Diffusion Ref | Texture Preservation Result

Figure 23: Examples showing how our method cannot handle input with large area of pure black and white regions.

1. Unlocking the Aesthetic Variety within ARTISAN To move beyond a single deterministic output and unlock the variety inherent in ARTISAN, the model needs mechanisms to navigate the different modes of the aesthetic distribution. Future work could explore:

- **Conditional Generation:** The most promising direction is transitioning to conditional generation. By incorporating style embeddings or exemplar images during training, the model could learn to navigate the diverse aesthetic landscape within ARTISAN.
- **Disentanglement of Style Codes:** Introducing explicit or latent style codes during training could disentangle different aesthetic modes. During inference, sampling different codes would yield diverse, yet equally valid, outputs for the same input.



Input | GenColor

Figure 24: Example of undesired hue changes for objects with ambiguous colors. While the enhanced image remains visually pleasing, the hue shift may not be appropriate for some scenarios.

2. Steering the Model Towards Desired Styles The current tight coupling between GenColor and the dominant aesthetic of ARTISAN can be adapted. The strength of our approach is that the model has already learned the complex task of high-quality, content-aware enhancement, providing a robust foundation for specialization.

- **Dataset Requirements for Steering:** To steer the model toward a specific style, another massive dataset is not required. As you hypothesized, a smaller, highly curated dataset exhibiting a consistent target style would be effective for fine-tuning.
- **Data Efficiency and Adaptation:** Because the model is already well-initialized by ARTISAN, we anticipate this adaptation could be highly data-efficient, potentially requiring only a few dozen to a few thousand samples. Techniques like Low-Rank Adaptation (LoRA) could make this style specialization highly practical.

3. Model Behavior: "Snapping" vs. Interpolation The question of whether the model would "snap" into different sub-styles or interpolate between them is excellent. We hypothesize that the behavior would depend on the implementation of the steering mechanism:

- **Interpolation:** Given the continuous nature of the diffusion latent space, if style is controlled via continuous embeddings or adjustable weights (e.g., varying the influence of a LoRA or using classifier-free guidance), the model would likely be capable of smoothly interpolating between styles.
- **Snapping:** If the steering is achieved by switching between distinct fine-tuned models, or if the fine-tuning signal for a new style is very strong and distinct from the ARTISAN mean, the output would likely "snap" to the new style (mode-switching).

J.3 Alternative Approaches of Weight Blending: Loss Function Design

While our weight-blending approach provides a pragmatic and effective solution for balancing aesthetics and fidelity, we acknowledge that alternative approaches could be explored. One promising direction, as suggested by reviewers, would be to design specialized loss functions that directly guide the model to an optimal state during training, rather than blending weights at inference time.

Such an approach could potentially incorporate terms that explicitly penalize strong hue changes while maintaining fidelity to the conditioning image, effectively learning the desired balance during the training process itself. This could involve multi-objective optimization techniques or carefully designed loss components that capture both aesthetic quality and color consistency.

However, exploring and properly tuning such loss functions would require extensive experimentation and a series of lengthy training runs to validate their effectiveness across different scenarios. Given the computational constraints and the proven reliability of our current weight-blending method, we leave this investigation as valuable future work. Our current approach offers the advantage of being robust, simple to implement, and achieving consistent results without altering the complex training dynamics of the diffusion model.

K More Visual Results

Additional qualitative results and comparisons can be found on our supplementary website, which is included in the supplementary materials package.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 clearly outline our contributions in color enhancement and accurately reflect the scope and limitations of our approach.

   Guidelines:

   - The answer NA 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 NA 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: The paper discusses the limitations of the work performed by the authors in the limitations section (§J).

Guidelines:

- The answer NA 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: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA 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: The paper provides all the information needed to reproduce the main paper and appendix.

   Guidelines:

- The answer NA 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: We will release the code and data after the review process.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details necessary to understand the results in the main paper and appendix.

   Guidelines:

   - The answer NA 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: [Yes]

   Justification: The paper reports error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments.

   Guidelines:

   - The answer NA 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: [Yes]

   Justification: The paper provides sufficient information on the computer resources needed to reproduce the experiments.

   Guidelines:

   - The answer NA 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: The paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA 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 paper discusses both potential positive societal impacts and negative societal impacts of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: We will release the model and data after the review process.

Guidelines:

- The answer NA 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 properly credits and mentions the license and terms of use of the assets used.

    Guidelines:

    - The answer NA 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: [Yes]

    Justification: The paper properly credits and mentions the license and terms of use of the assets used.

    Guidelines:

    - The answer NA 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: [NA]

Justification: We follow the NeurIPS Code of Ethics.

Guidelines:

- The answer NA 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: [NA]

Justification: We follow the NeurIPS Code of Ethics.

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use LLMs for writing, editing, and formatting purposes.

Guidelines:

- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.