StatLUT: Statistical Feature-Driven Multimodal 3DLUT Generation for Photorealistic Style Transfer

Yifan Wang, Zhixiang Hao, Yu Wang*, Congchao Zhu
Honor Device Co., Ltd.
wyf141060@163.com, {haozhixiang, wangyu24, zhucongchao}@honor.com

Abstract

Photorealistic Style Transfer (PST) aims to transfer the color and tonal style of a reference to a content image while strictly preserving its structural integrity. However, existing deep learning-based methods inherently suffer from semantic entanglement caused by pre-trained image encoders, leading to unnatural spatial distortions. Moreover, current pixel-level mapping paradigms often ignore color gamut topology, resulting in color banding, while also lacking the multimodal capability for precise, tone-specific text-driven control. To address these bottlenecks, we propose StatLUT, an innovative multimodal framework for 3D LUT generation. First, we bypass traditional encoders and introduce a Lab-Extractor to derive spatially-agnostic statistical features, fundamentally decoupling color distributions from structural semantics to ensure artifact-free rendering. Second, we formulate LUT generation as a Transformer-based Seq2Seq translation task, utilizing a Multi-dimensional Residual Mapper (MR-Mapper) to predict topologically smooth 3D LUTs. Finally, to break the unimodal barrier, we propose the H-Diffuser, a lightweight Diffusion Transformer that directly synthesizes statistical features from tone-descriptive text prompts, enabling controllable and robust color grading. Extensive experiments on standard benchmarks demonstrate that StatLUT significantly outperforms state-of-the-art methods in both visual quality and quantitative metrics, pioneering a highly robust and versatile paradigm for multimodal photorealistic style transfer.

1 Introduction

Photorealistic Style Transfer (PST) aims to faithfully transfer the color and tonal style of a reference image to a content image while strictly preserving the original spatial structure. Unlike artistic style transfer[1–5], PST demands visual realism and structural fidelity, precluding unnatural artifacts or texture distortions during non-linear color mapping.

Although deep learning has advanced PST, adhering to this "artifact-free" baseline remains challenging. Current mainstream methods widely adopt pre-trained encoder-decoder architectures[6–8]. However, these models naturally extract high-level semantic features, creating a fundamental mechanistic mismatch with the PST task, which inherently relies on low-level color distributions. This semantic entanglement allows local structural semantics to interfere with color mapping, leading to spatial distortions. Furthermore, their massive memory consumption constitutes a prohibitive bottleneck for high-resolution (e.g., 4K/8K) deployment[9, 10].

To overcome these barriers, recent research has shifted towards a pixel-level mapping paradigm[9–11], with 3D lookup tables (LUTs) emerging as the standard due to their non-linear capacity and low computational cost[10]. Nevertheless, existing LUT generation mechanisms exhibit significant

*Corresponding author.

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

limitations: 1) traditional blending paradigms[12, 13] restrict representations to a linear subspace, hindering out-of-distribution generalization; 2) recent generative paradigms[14] fail to completely decouple semantics from color, risking texture artifacts; 3) point-wise mapping approaches[9] ignore the topology between adjacent color bins, causing color banding and inter-frame flickering in videos. Moreover, they lack multimodal capabilities for precise, tone-specific text-driven color control.

To address these limitations, we propose StatLUT, a novel multimodal photorealistic style transfer and color generation framework. In summary, our main contributions are as follows:

- Bypassing traditional image encoders, we utilize multi-dimensional low-level statistical distributions to construct explicit color priors. This blocks semantic interference during color mapping, preserving structural fidelity and mitigating spatial distortions.
- By adopting a Transformer-based sequence-to-sequence paradigm, our method globally accounts for the mesh topology of the 3D LUT. This generates smooth color manifolds, which effectively resolves the issues of color banding and temporal flickering.
- We introduce a Diffusion Transformer (DiT) architecture[15] to break the unimodal barrier, enabling color control via tone-descriptive text prompts without relying on reference images, thereby expanding the application scope of PST.

2 Related Work

2.1 Feature-Encoding-based Photorealistic Style Transfer

Unlike artistic style transfer, PST strictly preserves structural integrity. Early optimization-based methods[6, 16] achieved this via photorealistic regularization but suffered from slow iterative processes. Subsequent feed-forward approaches and AdaIN-based pipelines[3] accelerated this process by utilizing pre-trained vision models for feature fusion.

However, this "Encode-Fuse-Decode" paradigm has two fundamental limitations. First, pre-trained encoders inherently extract high-level semantics[17, 18], causing a mechanistic mismatch with PST. Consequently, relying solely on loss functions fails to disentangle semantics from color, inevitably leading to texture artifacts. Second, processing full-resolution feature maps via heavy decoders incurs prohibitive computational costs, making real-time high-resolution rendering unattainable.

2.2 Pixel-Level Mapping and 3D LUT Generation

To bypass decoder bottlenecks, researchers explored pixel-level mapping. Early statistical and matrix-based methods[19, 20] were extremely fast but lacked complex non-linear expressiveness. Bilateral grids[21] introduced local mapping but disrupted inter-frame consistency in videos.

Consequently, 3D LUTs emerged as the optimal solution. However, basis-LUT blending paradigms[12, 13] restrict representations to a linear subspace, resulting in poor out-of-distribution generalization. Recent alternatives also face significant challenges: point-wise mappings ignore color lattice topology, causing banding, and reliance on pre-trained features inevitably introduces semantic interference[9, 22]. Diffusion-based methods require expensive instance-specific retraining[10], while other generative approaches[14] still mix structural semantics into condition vectors, failing to eliminate the risk of semantic entanglement.

2.3 Multimodal and Text-Driven Image Manipulation

While text-guided image manipulation has advanced rapidly[23–27], most works focus on generation[27] or colorization[28]. Colorization fundamentally differs from PST: it acts as a "painter" filling colors into grayscale structures, whereas PST acts as a "colorist" establishing precise color mappings to alter stylistic atmospheres.

In PST, multimodal exploration remains scarce. Existing text-to-image pipelines[24, 25] often overlook the fact that PST intrinsically relies on color priors, not specific semantics (e.g., a "sunset beach" and a "twilight forest" share similar colors despite different semantics). A direct, efficient, and temporally consistent text-to-LUT mechanism remains an unresolved challenge.

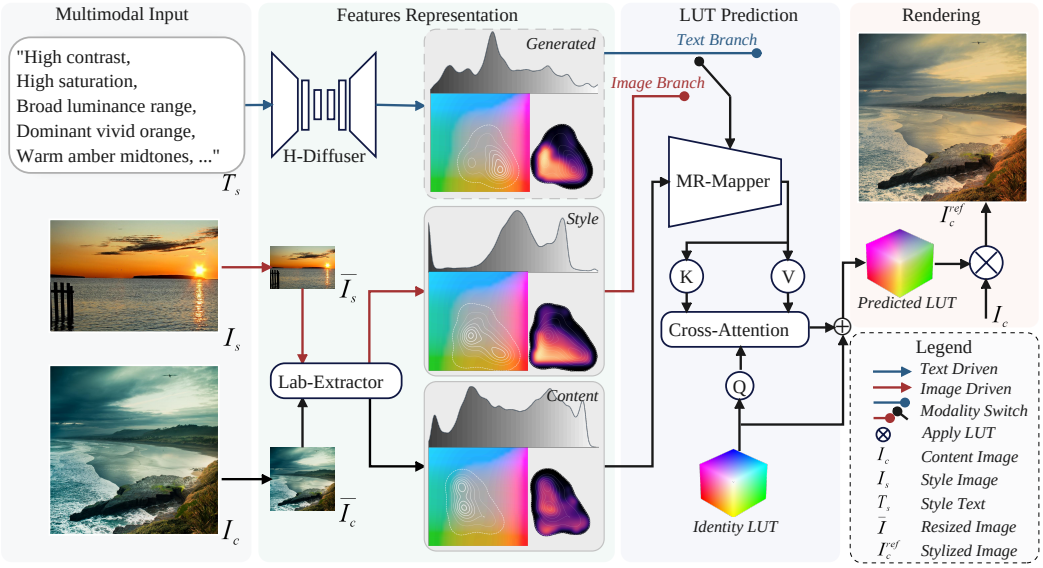

Figure 1: Overview of the proposed StatLUT framework. (1) Feature Extraction: The Lab-Extractor extracts spatially-agnostic statistical features from resized input images. (2) LUT Prediction: The MR-Mapper fuses these features to guide a Cross-Attention module, which transforms an identity LUT into the predicted 3D LUT. (3) Text-Driven Mapping: For text conditions, the H-Diffuser directly generates target statistics. Finally, the predicted LUT is applied to stylize the high-resolution content image.

3 Methodology

Given a high-resolution content image I_c and a style condition (reference image I_s or text prompt T_s), StatLUT predicts a 3D LUT for color stylization, transforming the source colors of I_c to align with the target style distribution (Figure 1). To eliminate semantic entanglement, we extract spatially-agnostic statistical features (Sec. 3.1). These features guide a Seq2Seq cross-attention module to predict a topologically smooth 3D LUT (Sec. 3.2). Finally, a DiT-based module enables text-driven color generation by directly synthesizing target statistics (Sec. 3.3).

3.1 Spatially-Agnostic Feature Extraction via Lab-Extractor

Since the 3D LUT formulation inherently guarantees the structural preservation of the content, the primary challenge in PST is extracting pure color and illumination distributions from the style image while strictly discarding spatial or semantic content. Existing approaches relying on deep encoders [29] or AdaIN [3] inevitably suffer from semantic entanglement and spatial biases. To fundamentally bypass this, we propose a purely spatially-agnostic feature representation using histograms.

Furthermore, to avoid the severe luminance-chrominance coupling inherent in the RGB space [30], we compute these statistical descriptors in the CIE Lab color space. Specifically, the input image is converted to the Lab space, and we extract three spatially-agnostic statistical descriptors: a 1D luminance histogram H_L (capturing independent illumination), a 2D color histogram H_{ab} (capturing independent chrominance), and a color-conditioned mean luminance map $M_{L|ab}$ (capturing luminance-chrominance correlation). The detailed architectural pipeline, mathematical formulations, and robust extraction strategies (e.g., non-linear stretching and soft-binning) of the proposed Lab-Extractor are provided in Appendix A.

3.2 Attention-Guided Residual LUT Generation via MR-Mapper

We formulate 3D LUT prediction as a Transformer-based Sequence-to-Sequence (Seq2Seq) translation task. To effectively inject the heterogeneous statistical features (1D/2D histograms and conditional luminance) into the network, we propose the Multi-dimensional Residual Mapper (MR-

Mapper) to unify them into a standard token sequence (Figure 2). This holistic prediction ensures mapping smoothness. Beyond the inherent temporal consistency of global LUTs ($f(x) = f(x)$), such smoothness suppresses flickering caused by minor pixel fluctuations ($x_1 \approx x_2 \Rightarrow f(x_1) \approx f(x_2)$).

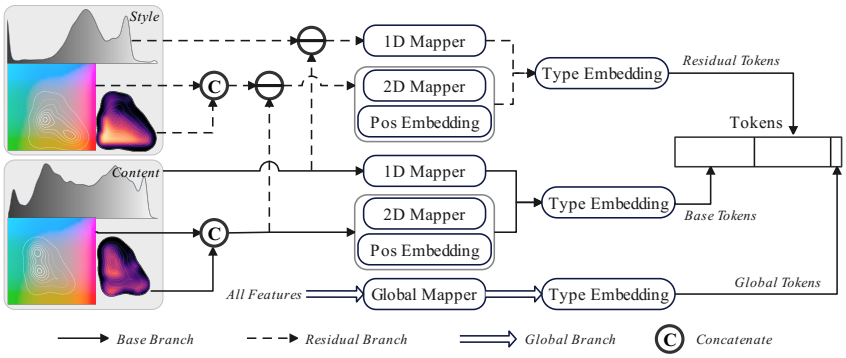

Figure 2: Architecture of the Multi-dimensional Residual Mapper (MR-Mapper), which unifies heterogeneous input features into a standard token sequence via Base, Residual, and Global branches.

The MR-Mapper comprises three distinct branches. The *Base Branch* encodes the content image's physical distribution, injecting 2D positional embeddings to preserve CIE Lab topological coordinates. The *Residual Branch* computes statistical differences ($\ominus$ in Figure 2) between style and content features, explicitly guiding the network to learn color shifts. The *Global Branch* extracts macroscopic statistics (e.g., contrast difference, chroma shift) as global constraints. Finally, branch outputs (F_{base}, F_{res}, F_{global}) are summed with type embeddings (E_{type}) and concatenated into a unified condition memory M:

$$M = \text{Concat}(F_{base}, F_{res}, F_{global}) + E_{type} \tag{1}$$

This memory M serves as the Key (K) and Value (V) in a Transformer decoder, while a 3D Identity LUT grid ($LUT_{id} \in \mathbb{R}^{N\times 3}$, $N = D^3$) acts as the Query (Q). To preserve color lattice topology, learnable 3D positional encodings (PE_R, PE_G, PE_B) are added to Q. The queries, keys, and values are formulated as:

$$Q = W_q(LUT_{id}) + (PE_R \oplus PE_G \oplus PE_B), \quad K = W_k(M), \quad V = W_v(M) \tag{2}$$

where W_q, W_k, W_v are linear projections, and $\oplus$ denotes broadcast addition.

Through cross-attention, each lattice point adaptively aggregates color transformation information to predict a color residual ΔC via a Feed-Forward Network (FFN):

$$\Delta C = \text{FFN}\left(\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V\right) \tag{3}$$

Predicting residuals aligns with the physical nature of color shifting and significantly eases optimization. The final stylized LUT is:

$$LUT_{pred} = \text{Clamp}(LUT_{id} + \Delta C, 0, 1) \tag{4}$$

To ensure training stability, we zero-initialize the final FFN projection layer. This guarantees an initial identity mapping ($\Delta C = 0$), preventing early-stage color distortion.

To eliminate the reliance on expensive paired datasets, we design an efficient self-supervised paradigm. We synthesize ground-truth stylized images I_{gt} by applying random LUTs (LUT_{gt}) to content images I_c. Spatial augmentations (e.g., cropping, scaling) are then applied to I_{gt} to generate the style reference I_s. Crucially, because our Lab-Extractor is strictly spatially-agnostic, it fundamentally prevents the network from taking a "shortcut" by memorizing spatial alignments, enabling robust training without semantic entanglement.

The training objective comprises three components. First, Smooth L1 loss constrains 3D color space:

$$\mathcal{L}_{LUT} = \mathcal{L}_{\text{SmoothL1}}(LUT_{pred}, LUT_{gt}) \tag{5}$$

Second, we ensure image-domain fidelity via pixel-wise L1 and SSIM [31] losses between the stylized output $I_{pred} = LUT_{pred}(I_c)$ and I_{gt}:

$$\mathcal{L}_{img} = \|I_{pred} - I_{gt}\|_1 + \mathcal{L}_{SSIM}(I_{pred}, I_{gt}) \tag{6}$$

Finally, monotonicity ($\mathcal{L}_{mono}$) and total variation ($\mathcal{L}_{tv}$) regularization terms are applied to maintain a stable color manifold. The total loss is:

$$\mathcal{L}_{total} = \lambda_{lut}\mathcal{L}_{LUT} + \lambda_{img}\mathcal{L}_{img} + \lambda_{mono}\mathcal{L}_{mono} + \lambda_{tv}\mathcal{L}_{tv} \tag{7}$$

3.3 Text-Driven Feature Synthesis via H-Diffuser

To bypass the computationally expensive two-stage paradigm (text-to-image generation followed by style transfer) and the under-constrained nature of directly predicting 3D LUTs from text, we propose the H-Diffuser (Histogram Diffuser). It is a lightweight Diffusion Transformer (DiT) designed to directly synthesize spatially-agnostic statistical features from a text prompt T_s. Specifically, we flatten and concatenate the target features (H_L, H_{ab}, $M_{L|ab}$) into a 1D token sequence X_0.

Unlike standard diffusion models that predict added noise, H-Diffuser directly predicts the clean sequence $\hat{X}_0$ conditioned on CLIP[26] text embeddings. This X_0-prediction objective is crucial for enforcing the strict physical boundaries (e.g., non-negativity) of statistical histograms. The predicted tokens are then linearly projected and reshaped back to their original spatial dimensions to reconstruct $\hat{H}_L$, $\hat{H}_{ab}$, and $\hat{M}_{L|ab}$.

A critical challenge in generating color histograms is the extreme sparsity of the 2D chrominance space, where most bins are empty. To prevent the network from predicting overly smoothed distributions dominated by these empty bins, we propose a density-aware loss. We construct a spatial weight mask $W = 1 + \alpha\bar{H}^{gt}_{ab}$, where $\bar{H}^{gt}_{ab} \in [0, 1]$ is the normalized ground-truth chrominance distribution and α is a scaling factor. This mask explicitly penalizes errors more heavily in dense color clusters. The final objective combines a Smooth L1 loss for the 1D luminance and a mask-weighted L1 loss for the sparse 2D sequences:

$$\mathcal{L}_{diff} = \lambda_L\mathcal{L}_{\text{SmoothL1}}(\hat{H}_L, H^{gt}_L) + \left\|W \odot \left(\lambda_{ab}(\hat{H}_{ab} - H^{gt}_{ab}) + \lambda_{abL}(\hat{M}_{L|ab} - M^{gt}_{L|ab})\right)\right\|_1 \tag{8}$$

where $\odot$ denotes element-wise multiplication. Once generated, these text-driven statistical features seamlessly replace the reference-extracted features to generate the stylized 3D LUT.

4 Experiments

We evaluate StatLUT against state-of-the-art methods. We detail the experimental setup, followed by qualitative and quantitative comparisons (Sec. 4.1), a user study (Sec. 4.2), and ablation studies (Sec. 4.3). Implementation details are provided in Appendix B. Video color grading results demonstrating temporal consistency are provided in Appendix C. Experimental results for cross-semantic scenarios are provided in Appendix D.

Datasets. We train on MS COCO [32] for content images and a custom dataset of 10,000 diverse 3D LUTs for self-supervised style conditions. Evaluation is conducted on the PST50(CC-BY 4.0 license) [13] and PhotoNAS(MIT license) [1] benchmarks, alongside high-resolution images for generalization assessment.

Implementation Summary. Images are resized to 224×224. The Lab-Extractor quantizes the color space to yield a 2304-dimensional statistical feature, which modulates a 16^3 3D LUT via a Transformer decoder. For text-driven generation, we employ a frozen CLIP encoder and an 8-block DiT (H-Diffuser) sampled via 200-step DDIM [33].

Quantitative Metrics. Since traditional metrics (e.g., PSNR, LPIPS) often misalign with human perception in photorealistic style transfer, we adopt the specialized protocol from Neural Preset [11]. We report style similarity (via a pre-trained discriminator) and content similarity (SSIM on LDC [34] features), both normalized to $[0, 1]$. Efficiency is measured by learnable parameters (Params) and multiply-accumulate operations (MACs).

4.1 Comparisons

We compare our proposed StatLUT against state-of-the-art deep learning-based color mapping methods: NLUT[12], Neural Preset[11], D-LUT[10], SA-LUT[13], and DLUT-VCG[14]. These methods serve as highly relevant baselines since they all generate explicit parameters for color mapping. For the NAS dataset, we additionally include the PhotoNAS[1] results provided by the original authors. To ensure a fair comparison, all baselines are evaluated using their official pre-trained models or default configurations. Quantitative results are computed comprehensively across the entire set of 90 content-style pairs to guarantee an unbiased evaluation.

4.1.1 Qualitative Results

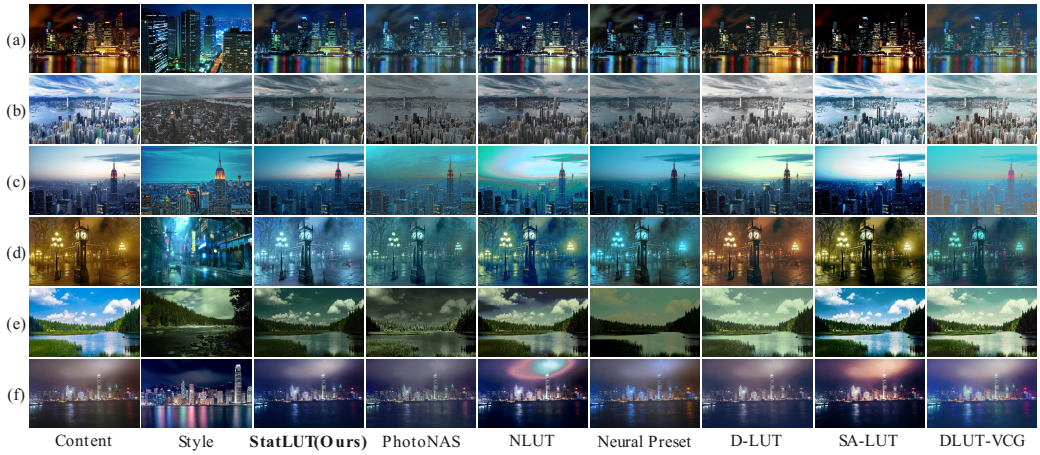

Figure 3: Visual evaluation of different methods on the NAS dataset.

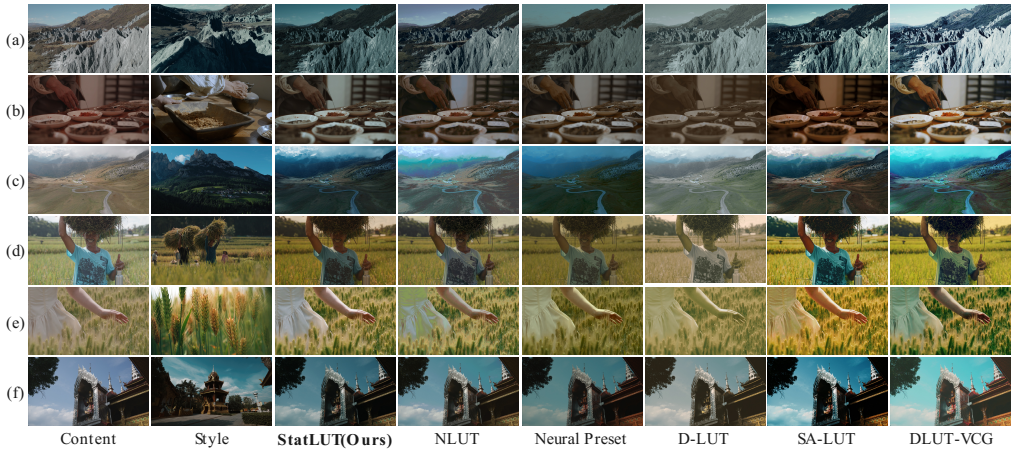

Figure 4: Visual evaluation of different methods on the PST50 dataset.

Figure 3 presents a visual comparison on the NAS dataset. Overall, StatLUT successfully achieves a color distribution that closely matches the reference style while faithfully preserving the structural integrity of the content images. In contrast, baseline methods exhibit various visual artifacts. Specifically, PhotoNAS and DLUT-VCG suffer from severe contrast degradation in certain scenes. NLUT, while effective in some cases, introduces pronounced color banding artifacts, particularly in smooth regions like skies. Neural Preset struggles with insufficient contrast, leading to blurred structural boundaries. Furthermore, D-LUT and SA-LUT demonstrate limited stylization capacity, resulting in outputs that noticeably deviate from the target styles.

Figure 4 illustrates the results on the PST50 dataset, where StatLUT consistently delivers superior and natural color transfer. Among the baselines, NLUT continues to exhibit color banding, whereas Neural Preset yields slightly degraded visual quality due to diminished global contrast. D-LUT shows inadequate stylization and occasionally introduces erroneous color mappings, such as unnatural skin tone renditions. SA-LUT tends to produce overly saturated results that fail to align with the style references. Lastly, DLUT-VCG demonstrates inconsistent performance; although it achieves reasonable local stylization, it is prone to unnatural global color shifts.

Generally, the NAS dataset presents a greater challenge due to larger domain gaps between content and style images, which exposes the limitations of the baseline methods. While the PST50 dataset is less challenging, allowing most methods to achieve basic color transfer, the baselines still suffer from subtle detail degradation. Across both datasets, StatLUT significantly outperforms competing methods by striking an optimal balance between accurate style rendition and robust content texture preservation.

4.1.2 Quantitative Results

Table 1: Quantitative comparisons on the NAS and PST50 datasets. StatLUT (Img) achieves the shortest distance to the ideal point, indicating an optimal balance between content preservation and style transfer. StatLUT (Text) relies solely on textual prompts without reference images. Best and second-best image-driven results are **bolded** and underlined, respectively.

Dataset	Metric	PhotoNAS	NLUT	Neural Preset	D-LUT	SA-LUT	DLUT-VCG	StatLUT (Img)	StatLUT (Text)
NAS	Content Sim. ↑	0.477	0.640	**0.709**	0.700	0.696	0.629	0.707	0.688
	Style Sim. ↑	0.557	**0.787**	0.656	0.562	0.426	0.676	0.746	0.498
	Dist. to Ideal ↓	0.681	0.400	0.448	0.517	0.598	0.491	**0.386**	0.565
PST50	Content Sim. ↑	—	0.746	**0.774**	0.734	0.729	0.713	0.760	0.757
	Style Sim. ↑	—	0.849	0.738	0.467	0.650	0.549	**0.870**	0.536
	Dist. to Ideal ↓	—	0.284	0.345	0.544	0.438	0.514	**0.259**	0.487

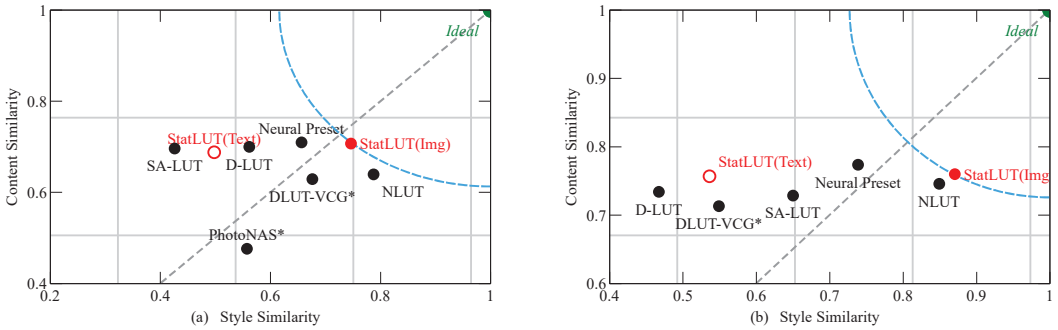

Figure 5: Quantitative evaluation on the (a) NAS and (b) PST50 datasets. The top-right corner $(1, 1)$ represents the ideal point. StatLUT (Img) (solid red dot) achieves the best balance between content and style similarity. Methods marked with (*) were resized during evaluation due to resolution mismatches. The hollow red circle denotes our text-driven variant, StatLUT (Text).

Figure 5 presents the quantitative evaluation. As depicted, StatLUT (Img) consistently positions closest to the ideal top-right corner across both datasets, demonstrating an optimal balance between structural preservation and style transfer. While NLUT achieves comparable style similarity, it suffers a notable drop in content similarity (e.g., 9.48% lower than StatLUT on the NAS dataset). This degradation directly corroborates the color banding artifacts observed in Figure 3, which inevitably corrupt underlying image textures. Conversely, Neural Preset maintains high content similarity but struggles with style transfer (15.17% lower style similarity than StatLUT on the PST50 dataset), aligning with its limited stylization capacity shown in Figure 4.

4.1.3 Cross-modal Text-driven Performance

Existing text-driven editing methods predominantly focus on localized manipulations rather than global color mapping. Due to this task disparity, direct comparisons are infeasible. Therefore,

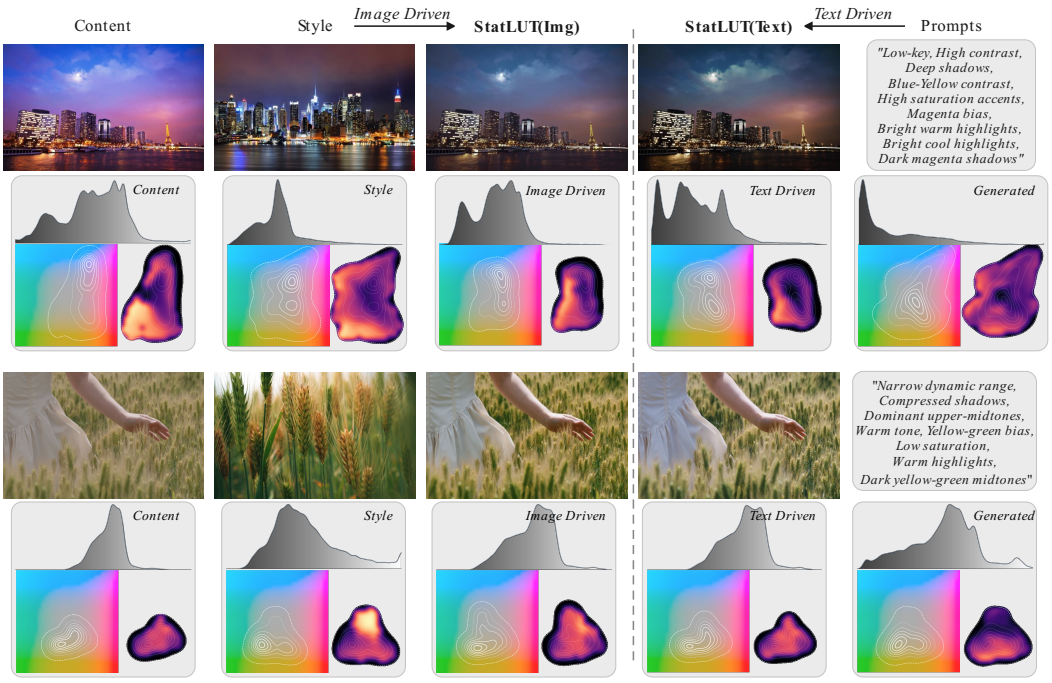

Figure 6: Visual comparison of intermediate features and stylized results between the image-driven (left) and text-driven (right) modes of StatLUT. Visualized features include 1D luminance histograms (top), 2D joint color distributions (bottom left), and mean luminance conditioned on color bins (bottom right).

we evaluate our cross-modal variant, StatLUT (Text), against our image-driven model as an upper-bound oracle.

Qualitatively (Figure 6), despite relying solely on tonal descriptions without visual references, StatLUT (Text) generates intermediate statistical features (Column 5) that closely align with those extracted from actual style images (Column 2). Consequently, the text-driven outputs (Column 4) visually approximate the image-driven results (Column 3), demonstrating highly effective cross-modal feature alignment.

Quantitatively (Table 1 and Figure 5), StatLUT (Text) maintains highly competitive Content Similarity, ensuring robust structural preservation. Although its Style Similarity naturally trails the image-driven oracle due to the inherent cross-modal gap between abstract text and concrete visual references, StatLUT (Text) offers precise and controllable global color grading without compromising structural integrity.

4.1.4 Semantic Decoupling Analysis

As emphasized in our motivation, existing methods often suffer from semantic entanglement, where the spatial structure of the style image inadvertently interferes with the color mapping process. To rigorously validate that StatLUT successfully decouples spatial semantics from color and luminance information, we design a novel style patch-shuffling experiment.

Specifically, we divide the style image into $P \times P$ patches and randomly shuffle them, which completely destroys its structural semantics while perfectly preserving its global color distributions. When applying these shuffled style images to the same content image, StatLUT maintains highly consistent color mapping results with near-zero residuals, significantly outperforming all baseline methods. This confirms that our statistical feature-driven paradigm is fundamentally immune to spatial semantic interference. We strongly encourage readers to refer to Appendix E for detailed visual and quantitative analyses of this compelling experiment.

4.2 User Study

Since quantitative metrics may not perfectly align with human perception, we conducted a user study with 50 participants (yielding 1,000 valid votes; see F for detailed protocols). As reported in Table 2, our StatLUT significantly outperforms all baselines, achieving a dominant Top-1 preference rate of 70% and the best average rank of 1.40 (compared to the runner-up Neural Preset at 15% and 2.75). This subjective evaluation strongly corroborates our quantitative findings, demonstrating that StatLUT consistently produces the most visually pleasing results.

Table 2: Quantitative results of the user study. We report the Top-1 preference rate (↑) and the average rank (↓) across 1,000 valid votes. The best results are highlighted in **bold**.

Metric	NLUT	Neural Preset	D-LUT	SA-LUT	DLUT-VCG	StatLUT (Ours)
Top-1 Preference (%) ↑	10	15	0	5	0	**70**
Average Rank ↓	2.95	2.75	4.70	4.95	4.25	**1.40**

4.3 Ablation Study

To validate our key components, we conduct ablation studies under identical settings.

Table 3: Ablation on feature combinations.

Features	Cont. ↑	Style ↑	Dist. ↓
A: RGB 1D-hist	0.714	**0.874**	0.286
B: Lab L+ab	0.716	0.848	0.304
C: Lab L+ab+abL	**0.726**	0.871	**0.281**

Table 4: Ablation on mapping modules.

Mapper	Cont. ↑	Dist. ↓	Params ↓
Simple MLP	0.723	0.286	428.9M
Bottleneck	0.726	0.281	5.24M
MR-Mapper	**0.760**	**0.259**	**0.38M**

Statistical Feature Combinations. As shown in Table 3, compared to highly coupled RGB histograms (Features A) or purely disentangled Lab histograms (Features B), our complete set (Features C) explicitly re-establishes the binding relationship between specific colors and illumination. It provides the most precise prior, achieving the optimal Dist. to Ideal (0.281).

Condition Mapping Module. Table 4 compares our MR-Mapper against MLP baselines using Features C. Traditional MLPs destroy structural information or introduce massive redundancy. In contrast, our MR-Mapper explicitly models residual differences and global statistics, achieving the highest generation quality (Content Sim. 0.760) with a compact footprint of 0.38M parameters.

LUT Size Selection. We also investigate the 3D LUT size D. Considering the cubic spatial complexity $\mathcal{O}(D^3 \times 3)$, we observe diminishing marginal utility beyond $D = 16$. Thus, we set $D = 16$ to balance fidelity and efficiency (see Appendix G).

5 Limitations and Future Work

Despite achieving state-of-the-art performance, StatLUT has limitations. First, our pipeline relies on absolute style descriptions. Future work could adapt H-Diffuser for relative textual prompts (e.g., "make it warmer") for intuitive editing. Second, the global nature of 3D LUTs precludes semantic-aware local fine-tuning. Exploring 4D LUTs with a semantic channel could address this while maintaining efficiency. Finally, extreme content-style gaps (e.g., daytime snow to nighttime neon) may yield suboptimal results. Integrating traditional statistical matching as a fallback for such outliers remains a practical necessity.

6 Conclusion

In this paper, we present StatLUT, a novel multimodal framework that advances photorealistic style transfer (PST) via statistical feature-driven 3D LUT generation. To resolve semantic entanglement and spatial leakage, we extract spatially-agnostic statistical distributions in the CIE Lab space as explicit color priors, ensuring strict structural preservation. By formulating LUT prediction as a Transformer-based Seq2Seq task with our MR-Mapper, StatLUT globally models color topology

with a highly compact parameter footprint, effectively eliminating color banding. Furthermore, we pioneer text-driven PST via the H-Diffuser, which seamlessly synthesizes statistical features from tone-descriptive prompts. Extensive experiments confirm that StatLUT achieves state-of-the-art performance in color fidelity, structural integrity, and multimodal flexibility. We envision this statistical prior paradigm as a highly promising direction for high-resolution image and video color grading.

Acknowledgments and Disclosure of Funding

This work was supported by Honor Device Co., Ltd. The authors declare no competing interests.

References

[1] Jie An, Haoyi Xiong, Jun Huan, and Jiebo Luo. Ultrafast Photorealistic Style Transfer via Neural Architecture Search. In *AAAI*, 2020.

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image Style Transfer Using Convolutional Neural Networks. In *CVPR*, 2016.

[3] Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *ICCV*, 2017.

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *ECCV*, 2016.

[5] Dae Young Park and Kwang Hee Lee. Arbitrary Style Transfer With Style-Attentional Networks. In *CVPR*, 2019.

[6] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *ECCV*, 2018.

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

[8] Tai-Yin Chiu and Danna Gurari. PhotoWCT^2 : Compact Autoencoder for Photorealistic Style Transfer Resulting from Blockwise Training and Skip Connections of High-Frequency Residuals. In *WACV*, 2022.

[9] Tianwei Lin, Honglin Lin, Fu Li, Dongliang He, Wenhao Wu, Meiling Wang, Xin Li, and Yong Liu. AdaCM: Adaptive ColorMLP for Real-Time Universal Photo-Realistic Style Transfer. In *AAAI*, 2023.

[10] Mujing Li, Guanjie Wang, Xingguang Zhang, Qifeng Liao, and Chenxi Xiao. D-LUT: Photorealistic style transfer via diffusion process. In *WACV*, 2025.

[11] Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson W.H. Lau. Neural Preset for Color Style Transfer. In *CVPR*, 2023.

[12] Yaosen Chen, Han Yang, Yuexin Yang, Yuegen Liu, Wei Wang, Xuming Wen, and Chaoping Xie. NLUT: Neural-based 3D Lookup Tables for Video Photorealistic Style Transfer. *arXiv preprint arXiv:2303.09170*, 2023.

[13] Zerui Gong, Zhonghua Wu, Qingyi Tao, Qinyue Li, and Chen Change Loy. SA-LUT: Spatial adaptive 4D look-up table for photorealistic style transfer. In *ICCV*, 2025.

[14] Seunghyun Shin, Dongmin Shin, Jisu Shin, Hae-Gon Jeon, and Joon-Young Lee. Video color grading via look-up table generation. In *ICCV*, 2025.

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

[16] F. Pitie, A.C. Kokaram, and R. Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *ICCV*, 2005.

[17] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[18] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[19] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color Transfer between Images. *IEEE CG&A*, 2001.

[20] Junjie Kang, Jinsong Wu, and Shiqi Jiang. Photorealistic image style transfer based on explicit affine transformation. In *ICME*, 2024.

[21] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. Joint Bilateral Learning for Real-Time Universal Photorealistic Style Transfer. In *ECCV*, 2020.

[22] Aupendu Kar and Guan-Ming Su. Temporal Consistent Semantic Video Color Transfer from Multiple References. In *CVPR*, 2025.

[23] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding Instruction-based Image Editing via Multimodal Large Language Models. In *ICLR*, 2024.

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-Based Real Image Editing with Diffusion Models. In *CVPR*, 2023.

[25] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*, 2023.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

[27] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*, 2021.

[28] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors. In *NIPS*, 2023.

[29] Rong Liu. Instant Photorealistic Style Transfer: A Lightweight and Adaptive Approach. *arXiv preprint arXiv:2309.10011*, 2023.

[30] Chenlei Lv, Dan Zhang, Shengling Geng, Zhongke Wu, and Hui Huang. Color Transfer for Images: A Survey. *TOMM*, 2024.

[31] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 2004.

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021.

[34] Xavier Soria, Gonzalo Pomboza-Junez, and Angel Domingo Sappa. Ldc: Lightweight dense cnn for edge detection. *IEEE Access*, 2022.

[35] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

[36] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[37] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 2011.

A Detailed Formulation of the Lab-Extractor

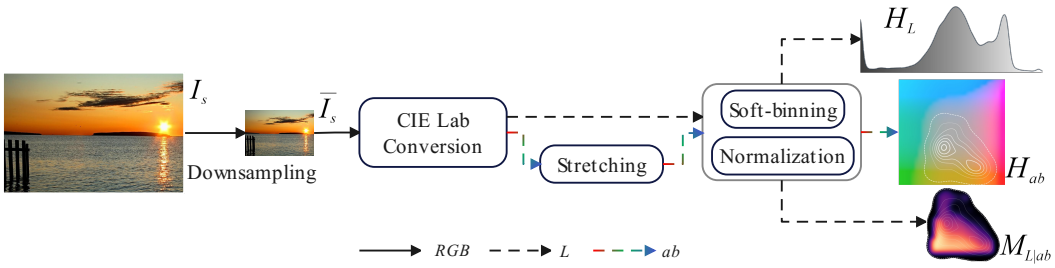

Figure 7: Pipeline of the proposed Lab-Extractor. The downsampled image is converted to the CIE Lab space to decouple luminance and chrominance. Following non-linear stretching of the a and b channels, three spatially-agnostic statistical descriptors (H_L, H_{ab}, and $M_{L|ab}$) are extracted via soft-binning. Contour lines in H_{ab} denote color frequency, where innermost regions indicate dominant colors. Overlaying these contours on $M_{L|ab}$ visually highlights the luminance properties corresponding to these dense color regions.

As discussed in Section 3.1, the input image is first spatially downsampled and converted to the Lab space to ensure efficient computation and strict decoupling of luminance and chrominance. In natural images, chrominance values (a and b) typically concentrate heavily around zero (neutral grays), causing sparse activations in peripheral histogram bins. To address this long-tail distribution and maximize representational capacity, we apply a **non-linear stretching** to the chrominance channels before quantization:

$$\hat{c} = \text{sgn}(c) \cdot \left(\frac{|c|}{S}\right)^{\gamma}$$

where $c \in \{a, b\}$, $\hat{c}$ is the stretched output, $\text{sgn}(\cdot)$ is the sign function, and S is a normalization scaling factor. The hyperparameter $\gamma < 1$ expands the central dense regions while compressing the sparse extremes, yielding a more uniform and informative distribution across color bins.

As illustrated in Figure 7, the Lab-Extractor derives three statistical components from the stretched Lab space: H_L, H_{ab}, and $M_{L|ab}$. The discrete extraction is formulated as:

$$H_L(i) = \frac{1}{HW}\sum_{x=1}^{W}\sum_{y=1}^{H}\mathbb{I}\big(L(x,y) \in \mathcal{B}_i^L\big)$$

$$H_{ab}(j,k) = \frac{1}{HW}\sum_{x=1}^{W}\sum_{y=1}^{H}\mathbb{I}\big(\hat{a}(x,y) \in \mathcal{B}_j^a \wedge \hat{b}(x,y) \in \mathcal{B}_k^b\big)$$

$$M_{L|ab}(j,k) = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H} L(x,y)\cdot\mathbb{I}\big(\hat{a}(x,y) \in \mathcal{B}_j^a \wedge \hat{b}(x,y) \in \mathcal{B}_k^b\big)}{\sum_{x=1}^{W}\sum_{y=1}^{H}\mathbb{I}\big(\hat{a}(x,y) \in \mathcal{B}_j^a \wedge \hat{b}(x,y) \in \mathcal{B}_k^b\big) + \epsilon}$$

where H and W are the image dimensions, (x, y) are spatial coordinates, and $\mathbb{I}(\cdot)$ is the indicator function. $\mathcal{B}_i^L$, $\mathcal{B}_j^a$, and $\mathcal{B}_k^b$ denote the i-th, j-th, and k-th bins for the L, $\hat{a}$, and $\hat{b}$ channels, respectively, and ϵ prevents division by zero.

In practice, to avoid quantization artifacts from the hard indicator $\mathbb{I}(\cdot)$, we employ a **soft-binning strategy**. Specifically, bilinear interpolation is used to distribute each pixel's contribution across adjacent bins, yielding a smoother and more robust descriptor. Finally, to prevent dominant colors (e.g., large background areas) from overwhelming the histogram, we apply a **square-root normalization** to the 2D color histogram: $\hat{H}_{ab}(j,k) = \sqrt{H_{ab}(j,k) + \epsilon}$. This non-linear dampening balances feature magnitudes, ensuring minority colors still contribute effectively to the style representation.

B Detailed Experimental Setups

Detailed Datasets. To establish style conditions for our self-supervised paradigm, our constructed dataset comprises 10,000 3D LUTs, where 4,000 are sourced from professional color grading resources and 6,000 are synthesized via statistical distribution sampling to guarantee stylistic diversity.

Image-Driven Pipeline Details. During training, the proposed Lab-Extractor quantizes the luminance H_L into 256 bins, and both H_{ab} and $M_{L|ab}$ into 32×32 grids, resulting in the 2304-dimensional feature. For

chrominance stretching, we set $\gamma = 0.5$ and a scale factor of $S = 128$. The Transformer decoder features 6 layers with a hidden dimension of $d_{model} = 512$ and 8 attention heads. The network is optimized using AdamW [35] for 200 epochs. The learning rate is initialized at 3.0×10^{-4} with a weight decay of 0.05, incorporating a 5-epoch linear warmup followed by a cosine annealing schedule decaying to 1.0×10^{-7}. The empirical weights for the objective functions are set to $\lambda_{lut} = 1.0$, $\lambda_{img} = 0.5$, $\lambda_{mono} = 5.0$, and $\lambda_{tv} = 0.0001$. The model is trained on 8 NVIDIA A800 GPUs for approximately 15 hours. During inference on a single A800 GPU, the entire image-driven pipeline (including feature extraction, LUT prediction, and mapping) takes less than 50 ms per image. Notably, the final 3D LUT mapping step is extremely efficient, requiring less than 0.1 ms.

Text-Driven Pipeline Details. To construct the training pairs, we employ Qwen3.5[36] to generate stylistic captions for all images in the COCO dataset. The generation process is strictly constrained to describe only luminance, chrominance, and their coupling effects, explicitly excluding any object-level semantics. These captions are then paired with the corresponding 2304-dimensional statistical features extracted by our Lab-Extractor to form the text-feature dataset. For the H-Diffuser, the 2304-dimensional features are tokenized into 36 patches, processed by an 8-block DiT architecture (hidden size of 768 and 12 heads). The diffusion model is trained for 500K iterations on 8 NVIDIA A800 GPUs with a total batch size of 4096. We apply a learning rate of 1.0×10^{-4} with a 3000-step linear warmup and a cosine decay schedule. The balancing weights for the diffusion objective are empirically set to $\lambda_L = 1.0$, $\lambda_{ab} = 1.2$, and $\lambda_M = 1.5$. The forward diffusion process spans 1000 timesteps. The entire training process takes approximately 50 hours. During inference on a single A800 GPU, the text-driven pipeline takes about 2 seconds per image, with the final 3D LUT mapping step remaining highly efficient at less than 0.1 ms. The substantial latency of the text-driven branch originates from the iterative generation process from text to features, which can be formulated as $\text{MACs}_{total} = \text{MACs}_{CLIP} + 200 \times \text{MACs}_{DiT_{step}}$ due to the 200-step DDIM sampling. However, this process is only triggered when the text description is updated. In most practical applications, the inference time is close to that of the image-driven branch (less than 50 ms).

C Application to Video Color Grading

To further demonstrate the practical applicability and robustness of our proposed StatLUT, we apply it to video color grading. In this experiment, we extract the color transformation from a single reference style image and apply it frame-by-frame to a source video.

As shown in Figure 8, our method successfully transfers the desired color style to the target video sequence. More importantly, because our method relies on stable statistical color mapping, the stylized video maintains excellent temporal consistency without introducing noticeable flickering or artifacts across consecutive frames. This indicates that StatLUT can be naturally extended to video processing tasks.

D Generalization to Cross-Semantic Scenarios

Experimental Setup. To further demonstrate the generalization capability of StatLUT for PST tasks across diverse semantic scenarios, we selected style images from the WikiArt[37] dataset, covering various artistic movements such as Abstract Expressionism, Impressionism, and Rococo. These styles present a significant domain gap compared to the PST50 dataset. To maximize the semantic challenge, we deliberately paired content and style images from entirely different categories, including portraits, architecture, landscapes, and objects.

Results and Analysis. As illustrated in Figure 9, we present PST results across 18 diverse cross-semantic pairs. The consistently high similarity scores suggest that the performance of StatLUT is not strictly dependent on semantic alignment between the content and style images. A representative case is the transfer of a classical portrait style to ancient temple architecture (Row 6, Col 3), which achieves a Style Similarity (SS) of 0.96 and a Content Similarity (CS) of 0.85. Such stability, even amidst significant semantic gaps, validates our core design: by leveraging spatially-agnostic features to decouple structural semantics from luminance and color information, StatLUT ensures robust and natural color mapping even across disparate semantic domains.

E Semantic Decoupling Analysis: Patch-Shuffling Experiment

In Section 4.1.4 of the main manuscript, we briefly introduced the style patch-shuffling experiment. Here, we provide the detailed experimental setups, visual comparisons, and quantitative analyses to further demonstrate the robust semantic decoupling capability of StatLUT.

Experimental Setup. To explicitly isolate the influence of spatial structural semantics from color and luminance information, we divide a given reference style image into $P \times P$ patches (where $P \in \{2, 3, 4, 5\}$) and randomly shuffle their spatial positions. This operation completely disrupts the structural semantics (e.g., the

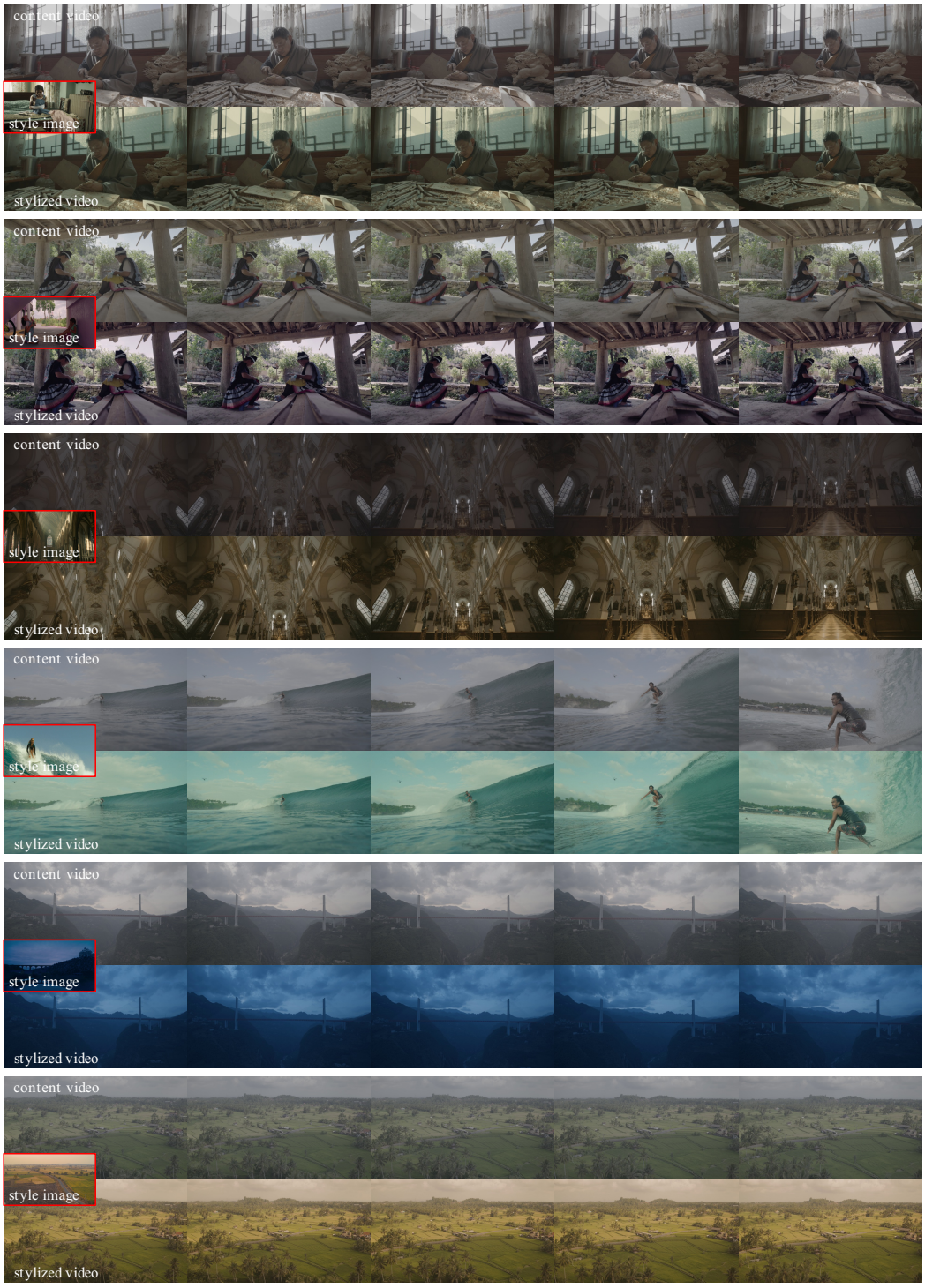

Figure 8: Qualitative results of video color grading. For each example, the top row shows the original content video frames, the left inset is the reference style image, and the bottom row presents our stylized video frames. Our method effectively transfers the desired color style while maintaining temporal consistency across consecutive frames.



Figure 9: Cross-semantic results (PST50 content, WikiArt style). For each pair: content (left), stylized result (right), and style reference (inset). CS/SS denote content/style similarity.

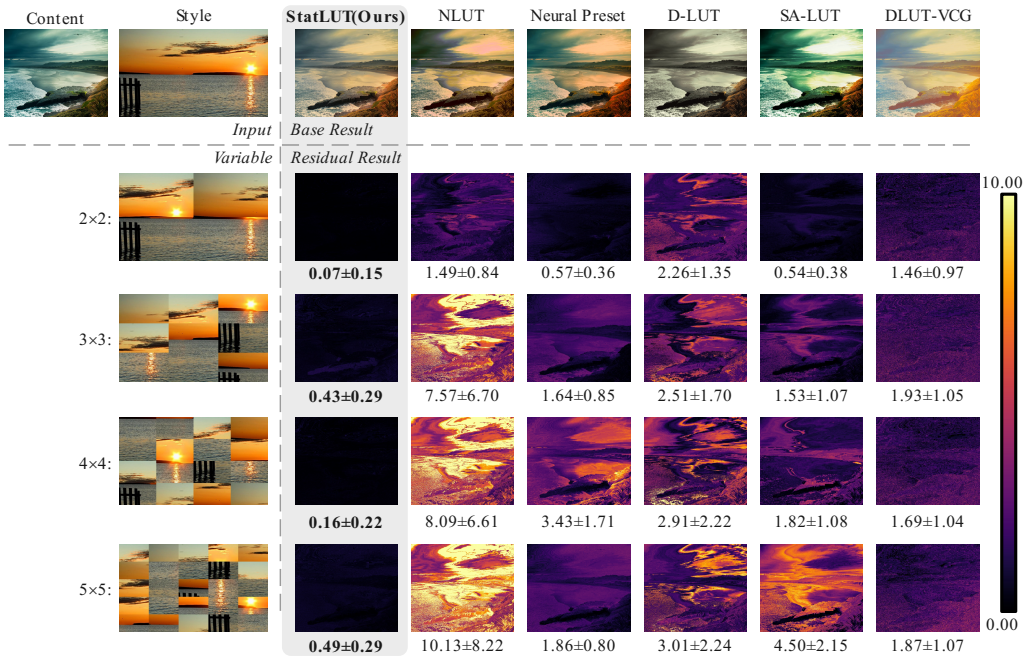

Figure 10: Visual and quantitative results of the style patch-shuffling experiment. The top row displays the content image, the original intact style image, and the base stylized results generated by different methods. The subsequent rows present the randomly shuffled style images ($P \times P$ patches, where $P \in \{2, 3, 4, 5\}$) alongside the corresponding residual maps. These maps represent the absolute difference between the results guided by the shuffled styles and the base results. The mean and standard deviation of the residuals are reported below each map. StatLUT (Ours) exhibits near-zero residuals (darkest maps), demonstrating superior robustness against spatial semantic interference.

shape of the sky, ocean, or sun) while strictly preserving the complete color and luminance distribution of the original style image. Subsequently, we perform style transfer on the same content image using these shuffled style variants. To quantify the stability of the color mapping, we calculate the residual maps between the stylized results guided by the shuffled styles and the result guided by the original, intact style image. We also record the mean and standard deviation of these residuals.

Results and Analysis. As illustrated in Figure 10, baseline methods (such as NLUT, D-LUT, and SA-LUT) exhibit significant variations and high residuals when the style image's spatial structure is altered. This clearly indicates that their color mapping processes are heavily entangled with the spatial semantics of the reference image.

In stark contrast, StatLUT consistently achieves the lowest residuals across all $P \times P$ configurations. The residual maps for our method are nearly entirely dark, proving that our Lab-Extractor effectively extracts pure statistical color features, rendering the color mapping process completely invariant to spatial structural changes.

It is worth noting that the residuals for StatLUT are not strictly zero (e.g., 0.07 ± 0.15 for the 2×2 configuration). This minor deviation is mathematically expected and reasonable. Before extracting the 1D and 2D histogram features, the input images undergo a resizing operation. For the shuffled style images, this resizing triggers resampling and interpolation at the newly formed artificial patch boundaries, causing slight pixel-level deviations from the original intact image. Consequently, these interpolation differences introduce negligible variations in the extracted statistical features, leading to the minimal non-zero residuals observed.

F User Study Details

To evaluate the perceptual quality and aesthetic appeal of our method in real-world scenarios, we conducted a comprehensive user study. We recruited 50 participants for this study. From the combined NAS and PST50 datasets (90 pairs in total), we randomly sampled 20 content-style pairs for each participant to mitigate visual fatigue.

For each trial, the participant was presented with the reference content and style images, followed by six stylized results generated by NLUT, Neural Preset, D-LUT, SA-LUT, DLUT-VCG, and our StatLUT. To ensure a fair comparison and eliminate positional bias, the six generated images were displayed in a completely randomized order for each trial. Participants were asked to rank the six results from 1 (best) to 6 (worst) based on three primary criteria: style similarity to the reference, content preservation (e.g., absence of artifacts), and overall visual aesthetics.

G Ablation Study Details

To validate the effectiveness of the key components proposed in our method, we conduct extensive ablation studies under identical datasets and training settings. This section primarily investigates the impact of the statistical feature combinations, the design of the condition mapping module, and the LUT size on the final color transfer performance and model efficiency.

G.1 Effectiveness of Statistical Feature Combinations

To evaluate the superiority of the statistical features designed in the Lab color space, we design three feature combinations for comparison. Since the feature dimensions vary across these variants, they cannot be directly processed by our proposed MR-Mapper. For a fair comparison, all three variants utilize a unified Bottleneck MLP combined with base token mapping to generate the condition vectors.

- **Features A (RGB 1D-hist)**: Extracts independent 1D histograms of the three channels in the RGB space (yielding a 3×256-dimensional feature).
- **Features B (Lab L+ab)**: Extracts independent lightness histograms (L-hist, 256 bins) and 2D color histograms (ab-hist, 32×32 bins) in the Lab space.
- **Features C (Lab L+ab+abL)**: Employs the complete feature input proposed in this paper, adding the ab-conditioned mean lightness (32×32 bins) to Features B.

The quantitative comparison results are presented in Table 5. By independently calculating the histograms in the RGB space (Features A), where the lightness and color information are highly coupled across the three channels, the model learns certain mapping relationships but achieves only suboptimal performance. When shifting to the Lab space (Features B) and completely disentangling lightness from color, providing only independent lightness and color histograms leads to the loss of correspondence between specific colors and their specific lightness levels (e.g., the blue of the sky is bright, while the green of the trees is dark). Consequently, the model becomes confused during color mapping, resulting in the lowest overall performance. Features C, by

introducing the ab-conditioned mean lightness (abL), explicitly re-establishes the binding relationship between specific colors and illumination while maintaining the advantages of lightness-color decoupling. This design avoids the dimensional explosion associated with 3D joint histograms and provides the most precise color-lightness mapping prior, thereby achieving the optimal performance across all metrics (with the Dist. to Ideal dropping to 0.281).

Table 5: Ablation study on the input statistical feature combinations. All variants use a unified Bottleneck MLP for fair comparison.

Feature Components	Content Sim. ↑	Style Sim. ↑	Dist. to Ideal ↓
Features A: RGB 1D-hist	0.714	**0.874**	0.286
Features B: Lab L-hist + ab-hist	0.716	0.848	0.304
Features C: Lab L-hist + ab-hist + abL-mean	**0.726**	0.871	**0.281**

G.2 Design of the Condition Mapping Module

We further investigate the advantages of the proposed MR-Mapper in projecting statistical features into Transformer condition vectors by comparing it against two baseline mapping strategies using our complete Lab feature set:

- **Simple MLP**: Flattens all features and maps them via a standard 3-layer MLP.
- **Bottleneck MLP**: Utilizes a bottleneck MLP architecture combined with base token mapping (identical to Features C in the previous section).
- **MR-Mapper (Ours)**: Employs our proposed architecture, which processes features through base, residual (delta), and global branches using 1D and 2D convolutions to preserve spatial topology.

As shown in Table 6, the Simple MLP employs a naive flattening strategy, which destroys the structural information of the features and leads to the worst performance (a Dist. to Ideal of 0.286). The Bottleneck MLP filters out redundant information to some extent by reducing dimensionality before mapping, yielding a performance improvement (Dist. to Ideal drops to 0.281). In contrast, our MR-Mapper achieves the best generation quality (the Dist. to Ideal reaches an optimal 0.259, and the Content Similarity significantly improves to 0.760). More importantly, the Params across these three mapping strategies exhibits a progressively decreasing trend. Traditional MLPs introduce a large number of redundant parameters, whereas the MR-Mapper explicitly models residual differences and global statistics while preserving spatial topology via weight-sharing convolutions. Although the MACs are slightly higher than the Bottleneck MLP due to the sliding window mechanism of convolutions, MR-Mapper achieves the most efficient multi-dimensional feature aggregation with significantly fewer parameters (only 0.38M), demonstrating its dual superiority in both performance and architectural efficiency.

Table 6: Quantitative comparison of different condition mapping modules. The proposed MR-Mapper achieves better style transfer performance while requiring significantly fewer parameters.

Mapping Module	Content Sim. ↑	Style Sim. ↑	Dist. to Ideal ↓	Params (M) ↓	MACs (M) ↓
Simple MLP	0.723	0.866	0.286	428.97	428.87
Bottleneck MLP	0.726	**0.871**	0.281	5.24	**5.24**
MR-Mapper (Ours)	**0.760**	0.870	**0.259**	**0.38**	13.82

G.3 Selection of LUT Size

The size of the 3D LUT, denoted as D, directly determines the precision of color mapping and the computational overhead of the model. To investigate the impact of the LUT size on the color transfer performance, we conduct an ablation study by setting $D \in \{8, 12, 16, 20, 24\}$.

As illustrated in Table 7, when the LUT size increases from 8 to 16, the model exhibits noticeable improvements in both content preservation and stylization capabilities. However, as D continues to increase to 20 or 24, the changes in performance metrics become marginal, with no significant further gains. It is worth noting that the number of parameters and memory footprint of the 3D LUT grow cubically, with a spatial complexity of $\mathcal{O}(D^3 \times 3)$. Therefore, larger LUT sizes introduce drastically higher memory consumption and computational overhead without providing proportional performance benefits. Taking both visual quality and real-time inference efficiency into comprehensive consideration, we ultimately select $D = 16$ as the hyperparameter setting for all experiments in this paper.

Table 7: Quantitative comparison of different LUT sizes.

LUT Size (D)	Content Sim. ↑	Style Sim. ↑	Dist. to Ideal ↓
8	0.705	0.834	0.322
12	0.706	0.858	0.303
16	0.760	0.870	0.259
20	0.752	0.865	0.268
24	0.754	0.874	0.260

H Broader Impacts

Positive Impacts. By leveraging the highly efficient 3D Look-Up Table (LUT) architecture, our stylization framework ensures low energy consumption and a reduced carbon footprint. This lightweight design aligns with Green AI goals and democratizes high-quality editing by enabling real-time processing on resource-constrained edge devices (e.g., mobile phones) without relying on heavy cloud infrastructure.

Negative Impacts and Mitigation. While our method only adjusts low-level colors and luminance, it could inadvertently alter human skin tones, raising fairness and representation concerns regarding racial identity. Additionally, it could be misused to visually enhance fabricated content. To mitigate these risks, we recommend incorporating skin-tone preservation masks or fairness-aware constraints when processing human subjects. Furthermore, real-world deployments should adopt provenance tracking (e.g., C2PA standards) or digital watermarking to ensure media authenticity.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our main contributions, including the Stat-LUT pipeline, the statistical feature representations, and the MR-Mapper. These claims are directly supported by the theoretical formulation in Section 3 and the experimental results in Section 4. Furthermore, we explicitly discuss the scope and limitations of our method in Section 5.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We explicitly discuss the limitations of our work in Section 5. Specifically, we address the reliance on absolute style descriptions, the inability of global 3D LUTs to perform semantic-aware local fine-tuning, and the potential for suboptimal results when there are extreme content-style gaps.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: Our paper focuses on an empirical methodology and architectural design for image and video color grading, and thus does not include formal theoretical results or mathematical proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide comprehensive implementation details, including network architectures, training datasets, and hyperparameters, in Section(Experiments) and the Appendix(Detailed Experimental Setups). Furthermore, we plan to release the source code and pre-trained models upon publication to facilitate full reproducibility.

   Guidelines:

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

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

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

   Answer: [No]

   Justification: The code and data are not included in the current submission due to ongoing internal review and cleanup. However, we are committed to open-sourcing the codebase, pre-trained models, and data processing scripts upon acceptance of the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We specify all necessary training and testing details, including the datasets used, hyperparameters, network configurations, and optimization strategies, in Section(Experiments) and the Appendix(Detailed Experimental Setups).

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Following standard practices in the field of image/video color grading and editing, we report the average performance metrics (e.g., Content Sim. and Style Sim.) over test dataset. Computing error bars across multiple training runs is computationally prohibitive due to the high cost of training diffusion models, and the inference process of our 3D LUTs is deterministic.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide comprehensive details on the computing resources in the Appendix. Specifically, we detail the hardware used (8 NVIDIA A800 GPUs) and the training time for both pipelines ($\sim$15 hours for image-driven and $\sim$50 hours for text-driven). Furthermore, we report the precise inference latency on a single A800 GPU (total inference takes < 50 ms and $\sim$2 seconds, respectively), highlighting that our core 3D LUT mapping consistently requires < 0.1 ms.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our research focuses on image stylization using 3D LUTs. While we conducted a user study to evaluate subjective visual preferences for the generated results, this study poses minimal risk to participants and does not collect any personally identifiable information (PII). Furthermore, all participants were fairly compensated at a rate that exceeds the local minimum hourly wage, strictly adhering to ethical guidelines for human evaluation. Therefore, our work conforms to the NeurIPS Code of Ethics in every respect.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss both positive and negative societal impacts, as well as potential mitigation strategies, in the Appendix(Broader Impacts).

    Guidelines:

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

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our work focuses on low-level image stylization (color and luminance adjustments via 3D LUTs) and does not generate new semantic content. Therefore, it does not pose the high risks associated with generative models or large-scale scraped datasets that would require strict release safeguards.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We properly cite the existing datasets used in our experiments. Specifically, the NAS dataset is distributed under the MIT license, and the PST50 dataset is available under the CC-BY 4.0 license. Detailed citations are provided in the paper.

    Guidelines:

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

13. **New assets**

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

    Answer: [No]

    Justification: The code is not included in the current submission, but we plan to publicly release the well-documented source code and pre-trained models upon the acceptance and publication of this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [Yes]

    Justification: We conducted a user study to evaluate the visual quality of our results. The details of the user study are provided in the supplementary Appendix(User Study Details).

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: While we conducted a user study to evaluate visual quality, it only involved standard visual preference questions. It collected no personally identifiable information (PII) and posed no physical or psychological risks to participants. Therefore, formal IRB approval was not required by our institution.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: We utilized the Qwen 3.5 LLM to generate text descriptions of images for constructing our training dataset. The details of this data generation process are described in the Appendix(Detailed Experimental Setups).

    Guidelines:

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