Supplementary Materials to "MovingColor: Seamless Fusion of Fine-grained Video Color Enhancement"

This supplementary material provides additional details, results, and analyses to support the main paper "MovingColor: Seamless Fusion of Fine-grained Video Color Enhancement". We present the detailed architecture of MovingColor and discuss the implementation details, including data augmentation and training procedures. The proposed Texture Difference (TD) metric for quantifying textural differences between input and output frames is elaborated upon. We report additional quantitative results on the DAVIS and YouTube-VOS datasets, confirming MovingColor's effectiveness across multiple benchmarks. Comprehensive ablation studies are conducted to investigate the impact of structural variants and loss functions on the model's performance. Robustness tests with varying input resolutions and color adjustments demonstrate MovingColor's resilience to different conditions. The construction of the D5 dataset, a new benchmark for video color fusion, is described in detail. Moreover, we provide insights from a user study and an interview with a professional colorist, highlighting the practical applicability and usability of MovingColor in real-world scenarios. Finally, we discuss the limitations and future directions. The supplementary materials aim to offer a deeper understanding of MovingColor's architecture, performance, and potential for seamless video color enhancement.

A More Visual Results in Accompanying Video Website

We would be immensely appreciative if you would kindly refer to our supplementary video website, where you will find a compelling showcase of MovingColor's superior performance in achieving spatiotemporal consistency for color fusion across a diverse range of video clips. The website includes side-by-side comparisons with related baselines, clearly showcasing MovingColor's effectiveness.

B Network Structure

We present the detailed architecture of our MovingColor in Table 1, where a batch of 15 frames (10 local + 5 global) is assumed. For the downsample stages (Stage 1 to Stage 3), FFC_BN_ACT indicates a Fast Fourier Convolution (FFC) block followed by batch normalization and activation. The spatial dimensions are halved in each stage while the channel dimensions are doubled. In Stage 3, the features are split into local (convl2l) and global (convl2g) branches. The FFC resnet blocks consist of a series of convolutional layers with a SpectralTransform module. The temporal transformer block incorporates a SlideWindowAttention module with a window size of 5×9 and 4 attention heads, followed by layer normalization and a FusionFeedForward module. For the upsample stages (Stage 1 to Stage 3), we use transposed convolutions to increase the spatial dimensions while reducing the channel dimensions. Skip connections are used to concatenate features from the corresponding downsample stages. The final output is obtained through a reflection padding and a convolutional layer.

C Implementation Details

C.1 Training Data Preparation

To enhance the robustness of the pretext task learning, we introduce randomness in color adjustment and mask generation, leveraging the diverse content and color distributions in the large-scale dataset. A random parametric Look-Up Table (LUT) generator, parameterized by $\boldsymbol{p}$, modifies video frame batches to create color-enhanced versions $\tilde{\boldsymbol{I}}_t$. Randomly generated masks $\boldsymbol{M}_t$, incorporating both static and dynamic elements, are obtained using approaches similar to [5]. The training input $\boldsymbol{X}_t = \text{Concate}[\boldsymbol{I}_t, \tilde{\boldsymbol{I}}_t \odot \boldsymbol{M}_t, \boldsymbol{M}_t]$ is formed in $\mathbb{R}^{H \times W \times 7}$, with $\tilde{\boldsymbol{I}}_t$ serving as the ground truth.

C.2 Data Augmentation

To enhance the robustness of our model to various real-world variations in video data, we employ a comprehensive data augmentation strategy during training. The augmentation pipeline consists of two main components:

- **Edge Detection:** Applying Canny edge detection [1] to video frame masks, with randomized kernel sizes (3-5) and dilation rates (1-16), enhances the model's ability to generalize for color fusion tasks by exposing it to varied mask edge conditions during training.
- **Color Augmentation:** We introduce random adjustments to contrast, brightness, gamma, hue, saturation, vibrance, and warmth, each applied with a probability of 0.2. These color filters contribute to the model's invariance to color variations.

C.3 Training Procedure

The model is trained for 500,000 iterations on four NVIDIA Tesla V100 GPUs, which typically takes approximately 5 days. The training procedure is configured as follows:

- **Optimizer:** We employ the Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.99$. The initial learning rate is set to 0.0001.
- **Scheduler:** A MultiStepLR scheduler is used with $\gamma = 0.01$ and milestones set at 4000 iterations.
- **Batch Size:** Despite the complexity of the model and the extensive data augmentation, we maintain a batch size of 16 across the GPUs, striking a balance between memory constraints and training dynamics.

C.3.1 Inference Phase. During the inference stage for video color enhancement, the system processes input frames, masks, and color-adjusted frames, addressing spatio-temporal inconsistency at mask edges. Inference is performed in a sliding window fashion over batches of frames. Local neighboring frames defined by the window size and global frames spaced by a stride are selected. The model performs feature propagation and transformer-based video color fusion, outputting the predicted frames. Results from overlapping windows are averaged for the final output.

Table 1: Detailed architecture of MovingColor for a batch of 15 frames (10 local + 5 global).

Stage	Output Size	Block	Details
Downsample stage 1	[15, 128, 128, 128]	FFC_BN_ACT	Conv2d(64, 128, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1), bias=False, padding_mode=reflect)
Downsample stage 2	[15, 256, 64, 64]	FFC_BN_ACT	Conv2d(128, 256, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1), bias=False, padding_mode=reflect)
Downsample stage 3	[15, 512, 32, 32]	FFC_BN_ACT	(convl2l): Conv2d(256, 128, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1), bias=False, padding_mode=reflect) (convl2g): Conv2d(256, 384, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1), bias=False, padding_mode=reflect)
FFC resnet blocks	[15, 512, 32, 32]	FFCResnetBlock	Conv2d(128, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1)), Conv2d(128, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1)), Conv2d(384, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1)), SpectralTransform
Temporal transformer	[15, 512, 32, 32]	TemporalTransformer	SlideWindowAttention (dim=512, head=4, window_size=(5, 9)), LayerNorm, FusionFeedForward
Upsample stage 1	[15, 256, 64, 64]	ConvTranspose2d	ConvTranspose2d(512, 256, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1), output_padding=(1, 1))
Upsample stage 2	[15, 128, 128, 128]	ConvTranspose2d	ConvTranspose2d(256, 128, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1), output_padding=(1, 1))
Upsample stage 3	[15, 64, 240, 432]	ConvTranspose2d	ConvTranspose2d(128, 64, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1), output_padding=(1, 1))
Output	[15, 3, 240, 432]	Final_Conv	nn.ReflectionPad2d(3): Conv2d(64, 3, kernel_size=(7, 7), stride=(1, 1))

The presented configuration and training strategies are carefully selected to optimize the efficiency and effectiveness of our neural network in learning robust color fusion tasks, as evidenced by its good performance.

D Texture Difference Metric

Texture Difference (TD) metric is introduced to quantify the perceptual difference in texture between input and output frames. Inspired by [5], we decompose an image into base and detail layers using a guided filter [3], an edge-preserving low-pass filter. The guided filter parameters are set based on input from a professional colorist to align with human color perception. TD is defined as the mean absolute difference between the detail layers of the input and output frames, effectively quantifying the model's ability to preserve textural details during color fusion.

E Additional Quantitative Results

E.1 Quantitative Results on DAVIS and YouTube-VOS

We report MovingColor's quantitative results on the D5 dataset in the main paper. Here in this supplement, we further validate performance on two additional datasets: DAVIS and YouTube-VOS. Due to the coarse object masks in these datasets, we report non-edge area differences rather than full-frame results. Table 2 and Table 3 present the quantitative results, confirming our method's effectiveness across multiple benchmarks.

Table 2 presents the color fusion performance comparisons on the DAVIS dataset. Our MovingColor method achieves the best results in terms of non-edge difference metrics (PSNR and ΔE), texture preservation (TD), and temporal consistency (PVCS). In the non-edge difference category, MovingColor obtains a PSNR of 29.80 and a ΔE of 2.93, outperforming the second-best method, Color Matcher, by a significant margin. MovingColor also achieves the lowest TD score of 2.02, indicating better texture preservation compared to other methods. Regarding temporal consistency, our method achieves the best PVCS score of 0.44, demonstrating its superior ability to maintain visual steadiness across frames.

Table 3 shows the color fusion performance comparisons on the YouTube-VOS dataset. Similar to the results on DAVIS, MovingColor achieves the best performance in terms of non-edge difference metrics (PSNR and ΔE) and temporal consistency (PVCS). Our method obtains a PSNR of 29.65 and a ΔE of 3.14, surpassing the second-best method, Color Matcher. In terms of texture preservation, MovingColor and Color Matcher both achieve the lowest TD score of 1.10. For temporal consistency, MovingColor achieves the second-best PVCS score of 0.42, closely following Color Matcher's score of 0.40.

It is worth noting that while some methods, such as Deflicker and FSPBT, achieve better results in certain temporal consistency metrics (PC_{PSNR}, PC_{SSIM}, and E_{Warp}), they perform poorly in non-edge difference and texture preservation categories. In contrast, MovingColor demonstrates a strong balance across all evaluation categories, achieving state-of-the-art performance in both spatial and temporal aspects of color fusion.

Table 2: Color fusion performance comparisons between Color Matcher, Harmonizer, S2CRNet, StyA2K, PCTNet, Deflicker, StyA2K+Deflicker, and our method on the DAVIS dataset.

Category	Method	Non-Edge Difference		Texture	Temporal Consistency			
		PSNR↑	ΔE↓	TD↓	PC_{PSNR} ↑	PC_{SSIM} ↑	PVCS ↓	E_{Warp} ↓
Space	**Color Matcher**	26.84	4.08	2.06	33.24	0.88	0.48	1.04
	Harmonizer	19.10	10.71	2.14	28.58	0.86	0.54	1.75
	S2CRNet	19.08	11.17	2.19	28.13	0.85	0.63	1.99
	PCTNet	21.47	8.15	2.16	28.97	0.86	0.47	1.53
	StyA2K	22.54	6.05	2.15	30.30	0.86	0.73	1.88
Time	**Deflicker**	18.71	8.31	2.70	**35.05**	**0.90**	1.50	1.25
Space + Time	**FSPBT**	26.15	4.13	2.09	34.13	0.88	0.56	**0.92**
	StyA2K+Deflicker	17.38	10.07	2.70	33.72	**0.90**	1.57	1.81
Ours	**MovingColor**	**29.80**	**2.93**	**2.02**	32.86	0.88	**0.44**	1.06

Table 3: Color fusion performance comparisons between Color Matcher, Harmonizer, S2CRNet, StyA2K, PCTNet, Deflicker, StyA2K+Deflicker, and our method on the YouTube-VOS dataset.

Category	Method	Non-Edge Difference		Texture	Temporal Consistency			
		PSNR↑	ΔE↓	TD↓	PC_{PSNR} ↑	PC_{SSIM} ↑	PVCS ↓	E_{Warp} ↓
Space	**Color Matcher**	28.48	3.47	**1.10**	37.10	0.93	**0.40**	0.53
	Harmonizer	19.98	10.44	1.42	34.72	0.92	0.50	0.78
	S2CRNet	19.55	9.57	1.78	35.09	0.92	0.57	0.70
	PCTNet	20.01	9.38	1.44	35.37	0.92	0.44	0.77
	StyA2K	22.05	6.64	1.36	34.61	0.92	0.75	0.96
Time	**Deflicker**	16.29	10.84	2.42	38.10	**0.94**	1.60	0.46
Space + Time	**FSPBT**	27.81	3.99	1.17	**38.25**	**0.94**	0.49	**0.44**
	StyA2K+Deflicker	15.03	13.10	2.28	36.53	**0.94**	1.75	0.74
Ours	**MovingColor**	**29.65**	**3.14**	**1.10**	37.37	0.93	0.42	0.51

E.2 Comparisons with CFNet and DDColor

ChromaFusionNet (CFNet) [2] treats color fusion as a color inpainting problem, focusing only on chroma channels while leaving the lightness (L) channel unchanged. This limitation affects its applicability since lightness is crucial for color adjustments. MovingColor, with our novel self-supervised training scheme, supports color fusion across all LAB channels, enabling adjustments like making the sky appear bluer by decreasing the lightness channel. Table 4 compares MovingColor, CFNet, the colorization method DDColor (ICCV23) [4], and DDColor adapted for CFNet's chroma fusion setting (DDColor_adp). The higher ΔE values for CFNet and DDColor variants highlight the limitations of not adjusting the lightness (L) channel.

Table 4: Quantitative comparison of MovingColor with CFNet, DDColor, and DDColor adapted for CFNet's chroma fusion setting (DDColor_adp).

Method	PSNR ↑	ΔE ↓	TD ↓	PC_{PSNR} ↑	PVCS ↓	E^*_{Warp} ↓
DDColor	16.68	20.21	1.23	34.67	1.10	0.64
DDColor_adp	18.56	19.57	1.14	36.38	0.91	0.51
CFNet	19.20	19.38	1.08	36.42	0.78	0.51
MovingColor	**26.88**	**4.35**	**1.03**	**38.61**	**0.54**	**0.29**

E.3 Additional Comparison with Video Harmonization Methods

Besides Harmonizer (ECCV22), which supports both image and video harmonization, we have added two video-specific methods: TSA^2 [7] and CO2Net [6]. As shown in the Table 5, our method consistently outperforms the alternatives.

Table 5: Quantitative comparison of MovingColor with TSA^2 and CO2Net.

Method	PSNR ↑	ΔE ↓	TD ↓	PC_{PSNR} ↑	PVCS ↓	E^*_{Warp} ↓
CO2Net	18.56	19.71	1.25	34.52	0.93	0.58
TSA^2	23.15	6.83	1.86	38.33	0.96	0.35
MovingColor	**26.88**	**4.35**	**1.03**	**38.61**	**0.54**	**0.29**

F More Ablation Studies

F.1 Details of the Structural Variants Study

We have included a brief table with key metrics in the main paper. Here, in this supplementary material, we report the full table with all metrics.

Table 6 presents a comprehensive ablation study of MovingColor on the DAVIS, YouTube-VOS, and D5 datasets, considering the inclusion or exclusion of the Fast Fourier Convolution (FFC) encoder, local frame feature propagation, and global frame feature propagation.

Table 6: Color fusion performance of various variants of MovingColor on DAVIS, YouTube-VOS, and D5 dataset.

Dataset	Variants			Non-Edge Diff		Texture	Temporal Consistency			
	FFC	Local	Global	PSNR↑	ΔE ↓	TD↓	PC_{PSNR} ↑	PC_{SSIM} ↑	PVCS ↓	E_{Warp} ↓
DAVIS		✓	✓	24.72	6.88	2.11	32.35	0.87	0.74	1.33
	✓		✓	29.77	**2.93**	2.04	32.79	**0.88**	**0.44**	1.08
	✓	✓		29.74	2.94	2.09	32.42	**0.88**	0.46	1.12
	✓	✓	✓	**29.80**	**2.93**	**2.02**	**32.86**	**0.88**	**0.44**	**1.06**
YouTube-VOS		✓	✓	23.06	8.28	1.25	36.32	0.92	0.76	0.76
	✓		✓	29.64	**3.14**	1.13	37.17	**0.93**	0.43	0.52
	✓	✓		29.61	**3.14**	1.17	36.83	**0.93**	0.43	0.55
	✓	✓	✓	**29.65**	**3.14**	**1.10**	**37.37**	**0.93**	**0.42**	**0.51**
D5		✓	✓	22.50	7.65	1.47	38.46	0.96	1.18	0.50
	✓		✓	28.22	**3.39**	1.04	38.60	**0.97**	**0.54**	0.30
	✓	✓		28.21	**3.39**	1.06	38.57	**0.97**	0.55	0.33
	✓	✓	✓	**28.24**	**3.39**	**1.03**	**38.61**	**0.97**	**0.54**	**0.29**

Across all datasets, the full MovingColor model (FFC + Local + Global) achieves the best performance in most metrics. Removing any of the components leads to a decrease in performance, with the variant without the FFC encoder (Local + Global) exhibiting the most significant drop, particularly in non-edge difference and temporal consistency metrics.

The variants without local or global frame feature propagation (FFC + Global and FFC + Local, respectively) show slight performance decreases compared to the full model, highlighting the importance of both local and global context for optimal color fusion results.

In summary, the ablation study validates the effectiveness of the proposed components in MovingColor. The Fast Fourier Convolution encoder captures rich spectral-spatial features, while the local and global frame feature propagation modules improve spatial and temporal consistency. The combination of these components enables MovingColor to achieve state-of-the-art performance in color fusion tasks across multiple datasets.

F.2 Loss Function Ablation Study

Table 7 presents an ablation study of MovingColor's loss functions on the DAVIS, YouTube-VOS, and D5 datasets, considering the inclusion or exclusion of the reconstruction loss ($\mathcal{L}_1$), adversarial loss ($\mathcal{L}_G$), and perceptual loss ($\mathcal{L}_{perc}$).

Across all datasets, the full MovingColor model (with all three loss components) achieves the best overall performance. The variant without the reconstruction loss ($\mathcal{L}_G + \mathcal{L}_{perc}$) excels in texture preservation and temporal consistency, while the variant without the adversarial loss ($\mathcal{L}_1 + \mathcal{L}_{perc}$) shows competitive results in non-edge difference and temporal consistency. The variant without the perceptual loss ($\mathcal{L}_1 + \mathcal{L}_G$) achieves the best PSNR on some datasets but lacks in other metrics.

These observations suggest that each loss component contributes to different aspects of MovingColor's performance. The reconstruction loss captures low-level details, the adversarial loss enhances realism and coherence, and the perceptual loss promotes semantic similarity.

In summary, the ablation study demonstrates the effectiveness of the proposed loss functions in MovingColor, enabling the model to achieve state-of-the-art performance in color fusion tasks across multiple datasets.

G Detailed Robustness Test Results

G.1 Resolution

We have included a brief chart with key metrics (ΔE & TD) in the main paper. Here, in this supplementary material, we report the full table with all metrics.

Table 8 presents a robustness test of MovingColor's performance on the DAVIS, YouTube-VOS, and D5 datasets, considering various input resolutions: 240P, 360P, 480P, and 720P.

Across all datasets, MovingColor demonstrates consistent performance improvements as the input resolution increases. The 720P variant achieves the best results in non-edge difference (PSNR and ΔE), texture preservation (TD), and temporal consistency metrics (PC_{PSNR}, PC_{SSIM}, PVCS, and E_{Warp}), followed by the 480P variant.

On the DAVIS dataset, the 720P variant obtains a PSNR of 31.28 and a ΔE of 2.62, outperforming lower resolution variants. Similar observations can be made on the YouTube-VOS and D5 datasets, where the 720P variant consistently achieves the best results across most metrics.

These results highlight the robustness of MovingColor across different input resolutions. The model's performance consistently improves as the resolution increases, demonstrating its ability to effectively exploit the additional information provided by higher-resolution inputs.

In summary, the robustness test analysis demonstrates that MovingColor maintains its state-of-the-art performance across various input resolutions on multiple datasets, further validating its effectiveness and practicality in real-world color fusion applications.

Table 7: Color fusion performance of various loss functions of MovingColor on DAVIS, YouTube-VOS, and D5 dataset.

Dataset	Loss Function			Non-Edge Diff		Texture	Temporal Consistency			
	$\mathcal{L}_1$	$\mathcal{L}_G$	$\mathcal{L}_{perc}$	PSNR↑	ΔE ↓	TD↓	PC_{PSNR} ↑	PC_{SSIM} ↑	PVCS ↓	E_{Warp} ↓
DAVIS		✓	✓	29.59	3.14	**2.01**	32.80	**0.88**	**0.41**	1.09
	✓		✓	29.40	3.22	2.02	32.69	**0.88**	0.45	1.07
	✓	✓		**29.91**	3.02	2.06	32.70	0.87	0.43	**1.04**
	✓	✓	✓	29.80	**2.93**	2.02	**32.86**	**0.88**	0.44	1.06
YouTube-VOS		✓	✓	29.04	3.52	**1.09**	37.32	**0.93**	**0.39**	0.54
	✓		✓	29.63	3.27	1.10	37.05	**0.93**	0.43	0.52
	✓	✓		28.87	3.50	1.17	37.24	**0.93**	0.40	0.54
	✓	✓	✓	**29.65**	**3.14**	1.10	**37.37**	**0.93**	0.42	**0.51**
D5		✓	✓	27.71	3.74	**1.00**	38.32	**0.97**	**0.50**	0.30
	✓		✓	27.90	3.69	1.04	38.46	**0.97**	0.56	**0.29**
	✓	✓		**28.86**	**3.33**	1.11	38.15	**0.97**	0.56	0.31
	✓	✓	✓	28.24	3.39	1.03	**38.61**	**0.97**	0.54	**0.29**

Table 8: Color fusion performance of various resolutions of MovingColor on DAVIS, YouTube-VOS, and D5 dataset.

Dataset	Size	Non-Edge Diff		Texture	Temporal Consistency			
		PSNR↑	ΔE ↓	TD↓	PC_{PSNR} ↑	PC_{SSIM} ↑	PVCS ↓	E_{Warp} ↓
DAVIS	240P	29.80	2.93	2.02	32.86	0.88	0.44	1.06
	360P	29.87	2.85	2.01	34.61	0.89	0.40	1.06
	480P	30.55	2.75	1.95	35.75	**0.90**	**0.37**	**1.00**
	720P	**31.28**	**2.62**	**1.78**	**36.40**	0.89	0.38	1.03
YouTube	240P	29.65	3.14	1.10	37.37	0.93	0.42	0.51
	360P	29.97	3.08	1.14	38.35	0.93	0.39	0.50
	480P	30.47	3.00	1.06	38.67	**0.94**	0.37	0.50
	720P	**31.26**	**2.84**	**0.96**	**39.11**	**0.94**	**0.36**	**0.46**
D5	240P	28.24	3.39	1.03	38.61	**0.97**	0.54	0.29
	360P	28.54	3.22	0.98	38.71	**0.97**	0.52	0.29
	480P	28.87	3.14	**0.95**	39.27	**0.97**	0.50	0.27
	720P	**29.25**	**3.00**	0.93	**40.29**	**0.97**	**0.48**	**0.28**

G.2 Detailed Color Adjustment Analysis

We have included a brief chart with key metrics in the main paper. Here, in this supplementary material, we report the full table with all metrics.

Table 9 presents the performance of MovingColor on the DAVIS, YouTube-VOS, and D5 datasets under various color adjustments, including brightness, contrast, exposure, gamma, hue, saturation, vibrance, and warmth.

Across all datasets, MovingColor demonstrates robustness to color adjustments, maintaining strong performance in non-edge difference (PSNR and ΔE), texture preservation (TD), and temporal consistency metrics (PC_{PSNR}, PC_{SSIM}, PVCS, and E_{Warp}).

On the DAVIS dataset, the $Saturation^-$ adjustment achieves the best PSNR (30.79) and ΔE (2.74), while the $Contrast^-$ adjustment yields the highest PC_{PSNR} (33.23) and lowest E_{Warp} (0.97). Similar trends are observed on the YouTube-VOS and D5 datasets, with the $Saturation^-$ adjustment consistently achieving top results in non-edge difference metrics.

These results demonstrate MovingColor's robustness to various color adjustments, highlighting its ability to maintain state-of-the-art performance under different color conditions. This robustness is essential for practical applications, where input videos may exhibit varying color characteristics.

In summary, the color adjustment analysis confirms MovingColor's resilience to color variations, further validating its effectiveness and reliability in real-world color fusion tasks.

H D5 Dataset Construction

We introduce the D5 dataset, a new benchmark for comprehensive evaluation of video color fusion methods. The dataset comprises 121 high-quality video sequences sourced from 12 community 3D scenes contributed by professional users in the D5 render community. These scenes encompass a diverse range of content, including nature scenes, urban environments, and human activities, ensuring a broad spectrum of color variations, textures, and lighting conditions. We will release the url to associated scenes and the whole

Table 9: Color fusion performance of various color adjustments of MovingColor on DAVIS, YouTube-VOS, and D5 dataset.

Dataset	Adjustments	Non-Edge Diff		Texture	Temporal Consistency			
		PSNR↑	ΔE ↓	TD↓	PC_{PSNR} ↑	PC_{SSIM} ↑	PVCS ↓	E_{Warp} ↓
DAVIS	$Brightness^{+}$	28.73	3.04	**2.01**	32.80	**0.88**	0.47	1.06
	$Contrast^{+}$	30.00	2.92	**2.01**	32.34	0.87	0.42	1.14
	$Exposure^{+}$	29.39	3.01	**2.01**	32.78	**0.88**	0.44	1.07
	$Gamma^{+}$	30.01	2.89	2.02	32.78	**0.88**	0.47	1.07
	Hue^{+}	29.98	3.08	2.05	32.80	**0.88**	0.45	1.04
	$Saturation^{+}$	29.77	2.96	2.03	32.80	**0.88**	**0.41**	1.06
	$Vibrance^{+}$	30.09	2.89	2.02	32.99	**0.88**	**0.41**	1.05
	$Warmth^{+}$	30.28	2.77	2.02	32.89	**0.88**	0.42	1.04
	$Brightness^{-}$	30.15	2.89	2.02	32.96	**0.88**	0.47	1.04
	$Contrast^{-}$	30.15	2.85	2.02	**33.23**	**0.88**	0.42	**0.97**
	$Exposure^{-}$	30.39	2.83	2.02	33.08	**0.88**	0.44	1.03
	$Gamma^{-}$	28.66	3.11	**2.01**	33.03	**0.88**	0.48	1.03
	Hue^{-}	30.20	2.89	2.02	32.86	**0.88**	**0.41**	1.04
	$Saturation^{-}$	**30.79**	**2.74**	**2.01**	32.95	**0.88**	**0.41**	1.04
	$Vibrance^{-}$	30.30	2.86	**2.01**	32.86	**0.88**	**0.41**	1.04
	$Warmth^{-}$	29.23	3.01	**2.01**	32.89	**0.88**	0.44	1.06
YouTube	$Brightness^{+}$	28.85	3.21	1.10	37.36	**0.93**	0.47	0.51
	$Contrast^{+}$	29.81	3.06	1.11	36.92	**0.93**	0.43	0.59
	$Exposure^{+}$	29.43	3.17	**1.09**	37.17	**0.93**	0.42	0.52
	$Gamma^{+}$	29.93	3.05	1.11	37.46	**0.93**	0.42	0.54
	Hue^{+}	29.75	3.18	1.14	37.34	**0.93**	0.43	0.52
	$Saturation^{+}$	29.73	3.10	1.11	37.34	**0.93**	0.40	0.52
	$Vibrance^{+}$	30.05	3.06	**1.09**	37.52	**0.93**	0.40	0.51
	$Warmth^{+}$	29.44	3.11	1.10	37.35	**0.93**	0.41	0.51
	$Brightness^{-}$	29.97	3.05	**1.09**	37.44	**0.93**	0.47	0.52
	$Contrast^{-}$	30.04	3.09	**1.09**	**38.10**	**0.93**	0.43	**0.45**
	$Exposure^{-}$	30.12	3.03	1.10	37.69	**0.93**	0.43	0.50
	$Gamma^{-}$	28.91	3.28	**1.09**	37.67	**0.93**	0.44	0.50
	Hue^{-}	29.90	3.14	**1.09**	37.27	**0.93**	0.40	0.51
	$Saturation^{-}$	**30.34**	**2.97**	**1.09**	37.33	**0.93**	**0.39**	0.50
	$Vibrance^{-}$	30.11	3.04	**1.09**	37.56	**0.93**	0.40	0.51
	$Warmth^{-}$	29.60	3.20	**1.09**	37.50	**0.93**	0.42	0.51
D5	$Brightness^{+}$	27.58	3.46	**1.02**	38.58	**0.97**	0.63	0.30
	$Contrast^{+}$	28.55	3.29	1.04	38.26	0.96	0.52	0.33
	$Exposure^{+}$	28.25	3.37	**1.02**	38.43	**0.97**	0.55	0.30
	$Gamma^{+}$	27.94	3.40	1.04	38.57	0.96	0.59	0.29
	Hue^{+}	28.68	3.71	1.05	38.66	0.96	0.52	0.30
	$Saturation^{+}$	28.65	3.35	1.03	38.44	0.96	0.50	0.29
	$Vibrance^{+}$	28.81	3.29	**1.02**	38.52	**0.97**	0.48	0.29
	$Warmth^{+}$	28.23	3.38	1.03	38.58	**0.97**	0.52	0.30
	$Brightness^{-}$	27.96	3.38	1.03	38.41	0.96	0.62	0.29
	$Contrast^{-}$	28.64	3.33	**1.02**	**39.23**	**0.97**	0.53	**0.26**
	$Exposure^{-}$	28.43	3.33	1.03	38.96	**0.97**	0.56	0.28
	$Gamma^{-}$	27.53	3.53	**1.01**	38.70	**0.97**	0.62	0.30
	Hue^{-}	28.84	3.33	**1.02**	38.51	**0.97**	0.48	0.29
	$Saturation^{-}$	**29.05**	**3.21**	**1.02**	38.80	**0.97**	**0.47**	0.29
	$Vibrance^{-}$	28.94	3.25	**1.02**	38.61	**0.97**	0.48	0.29
	$Warmth^{-}$	28.61	3.41	**1.02**	38.73	**0.97**	0.52	0.29

dataset at full resolution if the D5 community agrees to share it publicly.

Each video sequence in the D5 dataset has a duration of 30 seconds and a resolution of 3960 × 2160 pixels at 25 frames per second. The high spatial and temporal resolution allows for a detailed assessment of color fusion performance, particularly in terms of spatial consistency and temporal stability.

Moreover, the majority of the video sequences in D5 feature camera movement, introducing additional challenges that are representative of real-world video color fusion scenarios. This inclusion of dynamic camera motion enables a more rigorous evaluation of a method's ability to maintain spatio-temporal consistency and handle complex scene geometry.

Table 10: D5 dataset summary. Example clips are available at the supplementary website.

Name	# clips	Camera	Objects	Moving Object	In/Out	Lighting
Big Office	9	Pan, Zoom	sky, chair, tree, table, electronic device	person	Indoor	Indoor light, Sun
Condo	12	Pan, Zoom, Still	sky, tree, flower, cat, house, grass	dog, cat, person	Outdoor	Outdoor light, Sun
Interior Design	8	Pan, Zoom, Still	dog, chair, table	dog	Indoor	Indoor light
Modern Home	18	Pan, Zoom, Still	sky, person, chair, table, book	person	Indoor	Indoor light, Sun
Mooncakes	7	Pan, Zoom	mooncake, table, tree, book		Indoor	Indoor light
Exhibition Hall	9	Pan, Zoom	person, ball	person, ball	Indoor	Indoor
Shopping Center	8	Pan, Zoom	sky, building, tree, person, car	car, person	Outdoor	Outdoor light, Sun
The Last of Us	6	Pan, Zoom	sky, tree, grass, snow, person, house	person	Outdoor	Outdoor light, Sun
Wooden Architecture	11	Pan, Zoom	sky, tree, grass, building, chair, table, person, bird	person, bird	Indoor	Indoor light, Sun
Indoor Pool	10	Pan, Zoom	sky, swimming pool, chair		Indoor	Indoor light
French Manor	13	Pan, Zoom	sky, tree, grass, building, bird	bird	Outdoor	Outdoor light
Tower	10	Pan	sky, building, bird, moon, tree		Outdoor	Outdoor light, Sun

In summary, as shown in Table 10, the D5 dataset provides a comprehensive and effective benchmark for video color fusion methods, with its diverse content, high-quality video sequences, and inclusion of camera motion. This dataset facilitates a thorough evaluation of MovingColor and other state-of-the-art techniques, promoting the development of advanced video color fusion algorithms.

I User Study

To further evaluate the performance of MovingColor, we have conducted a set of user study and a professional colorist interview. The user study aims to assess the visual quality of MovingColor's color fusion results compared to the baselines. Its results have been reported in the main paper. The professional colorist interview provides insights into the practicality and usability of MovingColor in professional video post-production workflows.

I.1 User Study Settings and Demographics

We conducted a user study with 68 participants to evaluate the performance and user experience of MovingColor. Its results have been reported in the main paper. The demographic information of the participants is as follows:

- **Gender Distribution:** The study included 47 male participants (69.12%) and 21 female participants (30.88%).
- **Age Distribution:** The majority of the participants were between 30 and 40 years old (28 participants, 41.18%), followed by those between 20 and 30 years old (26 participants, 38.24%). The study also included 7 participants (10.29%) who were 20 years old or younger and 7 participants (10.29%) between 40 and 50 years old.
- **Education Distribution:** Most of the participants held a Bachelor's degree (43 participants, 63.24%), while 18 participants (26.47%) had a Master's degree or above. The remaining 7 participants (10.29%) had a high school diploma.
- **Experience Distribution:** The participants' experience levels varied, with 35 participants (51.47%) identifying as amateurs, 23 participants (33.82%) having no prior experience, and 10 participants (14.71%) being professionals in the field of video color enhancement or related areas.

The diverse demographics of the participants in terms of gender, age, education, and experience levels provide a representative sample for evaluating MovingColor's performance and user experience across a wide range of users.

I.2 User Interview with a Professional Colorist

To assess the real-world applicability and effectiveness of MovingColor, we conducted a qualitative interview with a professional colorist. The colorist interacted directly with MovingColor's demo app to perform color enhancements and fusion on various video sequences, comparing the results with those obtained using traditional methods.

- **Spatial and Temporal Consistency**: The colorist praised MovingColor's exceptional ability to blend colors seamlessly within the video frames while maintaining spatial and temporal consistency, even in challenging scenarios involving complex color gradients and transitions. The tool's capability to execute precise color adjustments without disrupting the spatio-temporal harmony of the video was seen as a significant advancement over traditional color grading methods.
- **Texture Preservation**: MovingColor's architecture garnered special praise for its innovative approach to texture preservation. The colorist highlighted that while most color enhancement tools tend to compromise the video's original texture during color adjustment, MovingColor excelled in retaining textural details, which is critical for high-quality visual outputs.
- **Temporal Consistency**: The colorist commended MovingColor's ability to maintain temporal consistency across video frames. The tool's global-local feature propagation module was noted for its effectiveness in ensuring a smooth and coherent color enhancement throughout the video sequence, minimizing any flickering or inconsistencies that often arise with traditional methods.
- **User Experience and Integration**: The user interface of MovingColor was lauded for its intuitiveness and ease of use, with the colorist appreciating the streamlined workflow

and the tool's processing efficiency. The remarkable integration ease of MovingColor with existing color enhancement pipelines was also highlighted, as it aligns seamlessly with conventional workflows and requires no extensive reconfiguration or steep learning curve.

Overall, the detailed feedback from the professional colorist affirmed the effectiveness of MovingColor in addressing the nuances of video color enhancement, underscoring its potential to revolutionize the field with its advanced features, user-centric design, and seamless integration into traditional workflows.

J Limitations and Future Work

MovingColor struggles to accurately process long, thin objects, such as sticks, and transparent objects. This limitation stems from the difficulty in segmenting long, thin objects and transparent objects, which remains a challenging problem in computer vision. Addressing this issue and improving the method's ability to handle these complex objects will be a focus of our future work. Despite their limitations in generating high-resolution, long videos, Diffusion-based models show promise for enhanced color fusion, which we aim to explore in the future.

References

[1] John Canny. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8, 6 (1986), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

[2] Yi Dong, Yuxi Wang, Ruoxi Fan, Wenqi Ouyang, Zhiqi Shen, Peiran Ren, and Xuansong Xie. 2024. ChromaFusionNet (CFNet): Natural Fusion of Fine-Grained Color Editing. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 2 (Mar. 2024), 1591–1599. https://doi.org/10.1609/aaai.v38i2.27925

[3] Kaiming He, Jian Sun, and Xiaoou Tang. 2010. Guided Image Filtering. In *Computer Vision – ECCV 2010*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–14.

[4] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. 2023. DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 328–338.

[5] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. 2019. Deep Video Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5792–5801.

[6] Xinyuan Lu, Shengyuan Huang, Li Niu, Wenyan Cong, and Liqing Zhang. 2022. Deep Video Harmonization With Color Mapping Consistency. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1232–1238. https://doi.org/10.24963/ijcai.2022/172 Main Track.

[7] Zeyu Xiao, Yurui Zhu, Xueyang Fu, and Zhiwei Xiong. 2024. TSA2: Temporal Segment Adaptation and Aggregation for Video Harmonization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 4136–4145.