COLOR3D: CONTROLLABLE AND CONSISTENT 3D COLORIZATION WITH PERSONALIZED COLORIZER

Yecong Wan[1], **Mingwen Shao**[2], **Renlong Wu**[1], **Wangmeng Zuo**[1]
[1]Harbin Institute of Technology, [2]Shenzhen University of Advanced Technology

Figure 1: **Exemplary Visual Results of Color3D**. Color3D is a unified controllable 3D colorization framework for both static and dynamic scenes, producing vivid and chromatically rich renderings with strong cross-view and cross-time consistency. Our method supports diverse colorization controls, including language-guided (left), automatic inference (middle), and reference-based (right), showcasing its versatility and practical value.

ABSTRACT

In this work, we present **Color3D**, a highly adaptable framework for colorizing both static and dynamic 3D scenes from monochromatic inputs, delivering visually diverse and chromatically vibrant reconstructions with flexible user-guided control. In contrast to existing methods that focus solely on static scenarios and enforce multi-view consistency by averaging color variations which inevitably sacrifice both chromatic richness and controllability, our approach is able to preserve color diversity and steerability while ensuring cross-view and cross-time consistency. In particular, the core insight of our method is to colorize only a single key view and then fine-tune a personalized colorizer to propagate its color to novel views and time steps. Through personalization, the colorizer learns a scene-specific deterministic color mapping underlying the reference view, enabling it to consistently project corresponding colors to the content in novel views and video frames via its inherent inductive bias. Once trained, the personalized colorizer can be applied to infer consistent chrominance for all other images, enabling direct reconstruction of colorful 3D scenes with a dedicated Lab color space Gaussian splatting representation. The proposed framework ingeniously recasts complicated 3D colorization as a more tractable single image paradigm, allowing seamless integration of arbitrary image colorization models with enhanced flexibility and controllability. Extensive experiments across diverse static and dynamic 3D colorization benchmarks substantiate that our method can deliver more consistent and chromatically rich renderings with precise user control. Project Page `https://yecongwan.github.io/Color3D/`

1 INTRODUCTION

The emerging 3D reconstruction techniques, such as Neural Radiance Fields (NeRF) Mildenhall et al. (2021) and 3D Gaussian Splatting (3DGS) Kerbl et al. (2023), have catalyzed high-fidelity novel-view synthesis from the observed color images. These methods evolves rapidly Chen et al. (2022); Fridovich-Keil et al. (2022); Garbin et al. (2021); Müller et al. (2022); Yu et al. (2024); Cheng et al. (2024a); Wu et al. (2024); Lu et al. (2024), and are extended to dynamic scenes Liu et al. (2023b); Guo et al. (2023); Shao et al. (2023); Duan et al. (2024); Li et al. (2024b); Yang et al. (2023); Wu et al. (2024); Yang et al. (2024c) with object motion modeling. Despite these advances, there is still a compelling challenge, i.e., reconstructing colorful 3D scenes from monochromatic inputs. This can significantly enhance the visual realism and expression of 3D reconstructions, unlocking new applications across digital art, artistic creation, and cultural heritage preservation.

In light of this, a straightforward way is to train a 3D colorization model on massive collections of 3D scene data. However, the formidable complexity of 3D geometry Lu et al. (2024), and the temporal intricacies in dynamic scenes Yan et al. (2024), making it prohibitively infeasible in practice. An alternative solution is to first colorize multi-view monochromatic images with 2D colorization models and then reconstruct the 3D content. Nevertheless, its naive implementation leads to severe cross-view color shift, due to latest 2D colorization models Kang et al. (2023); Yang et al. (2024a) struggling to colorize multi-view images consistently (See Fig. 9).

To mitigate the color inconsistency, a few attempts Dhiman et al. (2023); Cheng et al. (2024b) aims to average multi-view color variations via distillation Dhiman et al. (2023) or dynamic color injection Cheng et al. (2024b). While partially effective, they inevitably dilute the palette richness, producing desaturated and tone-flattened results. Furthermore, smoothing color variations makes the colorized results unpredictable, sacrificing user-controlled colorization ability. Moreover, existing studies only focus on static scenes, while controllable colorization of dynamic ones remains an open and unexplored problem, where color consistency in spatial and temporal dimensions should be maintained.

In this work, we suggest a novel paradigm for 3D colorization, i.e., first colorizing a key image and then propagating the color information to novel views and time steps by fine-tuning a personalized colorizer. The benefits are three-fold. First, it transforms the complex 3D colorization into a simpler color information propagation task, enabling more consistent results. Second, since only a key image needs to be controlled, it is easier to achieve controllable 3D colorization. Third, it provides a chance for unifying 3D colorization of both static and dynamic scenes.

Driven by the motivations, we propose **Color3D**, a unified controllable 3D colorization framework for both static and dynamic scenes (Fig. 1), where a personalized colorizer is optimized for each scene to enable consistent color propagation. Specifically, Color3D first selects the most informative view from the monochromatic images, and then colorizes it with a desirable off-the-shelf image colorization model. Next, to colorize other views or timesteps consistently, Color3D personalizes a scene-specific colorizer to learn the deterministic color mapping underlying this colorized view. In which a single view augmentation strategy is devised to expand the sample space, thereby enhancing the colorizer's generalization and its capacity to handle unseen visual content. Finally, the personalized colorizer is employed to infer consistent chromatic content of remaining views or frames, followed by direct reconstruction of colorful static or dynamic 3D scenes via a dedicated Lab color space Gaussian splatting representation, where luminance and chrominance components are optimized separately. Extensive experiments on public static and dynamic 3D datasets substantiate that our Color3D can produce colorful 3D content that aligns with user intentions, outperforming alternatives quantitatively and qualitatively. Additionally, Color3D also achieves visually promising results on real-world in-the-wild scenes, further demonstrating the practicality of our method in legacy restoration.

In conclusion, the main contributions are summarized as follows:

- We propose a unified and versatile 3D colorization framework, termed Color3D, which enables user-guided colorization of both static and dynamic scenes by tuning a personalized colorizer for each scene, advancing the controllability and interactivity of 3D colorization.
- To facilitate personalized colorizer tuning, we propose a customized key view selection strategy along with a single view augmentation scheme to enhance coloring richness and generalization. Furthermore, we introduce a Lab Gaussian representation to improve color reconstruction fidelity and better preserve scene structures.

- Our Color3D delivers vibrant, controllable, consistent results that faithfully align with user intent, significantly outperforming alternatives on a variety of benchmarks.

2 Related Work

Radiance Fields for Static and Dynamic Scenes. Radiance field representations Mildenhall et al. (2021); Kerbl et al. (2023) have driven a paradigm shift in novel view synthesis. Neural Radiance Fields (NeRF) Mildenhall et al. (2021) pioneered implicit volumetric modeling with coordinate-based neural networks, inspiring numerous extensions Barron et al. (2021; 2022; 2023); Verbin et al. (2022); Chen et al. (2022); Fridovich-Keil et al. (2022); Garbin et al. (2021); Müller et al. (2022) and subsequent work on dynamic scenes Park et al. (2021a;b); Wang et al. (2023); Fang et al. (2022); Liu et al. (2023b); Guo et al. (2023); Shao et al. (2023). However, NeRF-based methods remain hindered by costly training and slow rendering. To overcome this, 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) introduced an explicit Gaussian representation that enables real-time, photorealistic rendering. Recent advances further extend Gaussian splatting to dynamic reconstruction Duan et al. (2024); Li et al. (2024b); Yang et al. (2023); Wu et al. (2024); Yang et al. (2024c), combining efficiency and fidelity. Motivated by these developments, we employ 3DGS and 4DGS as canonical static and dynamic representations, building a unified framework for controllable 3D colorization with strong spatial-temporal consistency.

From 2D Colorization to 3D. Image colorization aims to recover plausible chromatic content from grayscale inputs, with decades of progress improving realism and controllability Zhang et al. (2017); Wu et al. (2021); Kim et al. (2022); Kang et al. (2023); Ji et al. (2022). Beyond automatic approaches, user-guided methods leverage reference images He et al. (2018); Huang et al. (2022); Zhao et al. (2021) or language prompts Weng et al. (2023); Zabari et al. (2023); Weng et al. (2022); Chang et al. (2023); Liang et al. (2024). Extending to videos introduces temporal consistency challenges, tackled by matching Yang et al. (2024b), palette transfer Wang et al. (2025), attention Li et al. (2024a), and memory propagation Yang et al. (2024a). Yet these remain limited to 2D domains, lacking cross-view consistency. Efforts on 3D scene colorization are still nascent. GBC Liao et al. (2024) exploits video models for continuous inputs, while ChromaDistill Dhiman et al. (2023) and ColorNeRF Cheng et al. (2024b) transfer knowledge from pretrained colorizers by averaging inconsistent predictions, often sacrificing controllability and vividness. Moreover, colorizing dynamic 3D scenes while ensuring spatial-temporal consistency remains unaddressed. Our work bridges this gap with a unified framework for both static and dynamic settings, achieving visually rich, consistent, and user-controllable 3D colorization.

3 Methodology

Our core idea is to fine-tune a personalized colorizer for each scene using a single colorized key view and then employ it to propagate the color information to novel views and time steps. By specializing in the scene-specific deterministic color mapping derived from this reference view, the colorizer is able to project the same content in novel views to the corresponding colors in the reference view through the inherent inductive bias capability, thereby elegantly addressing the challenge of color consistency across both static and dynamic 3D scenes.

3.1 Overview of Color3D

The schematic illustration of the proposed Color3D is depicted in Fig. 2. Color3D is primarily composed of two phases, i.e., the personalized colorizer training stage and the 3D scene colorization stage. In the first stage, we begin by selecting a single key view from the inputs and apply an off-the-shelf image colorization model to generate a colorized reference view. With the augmented data from a customized single view augmentation scheme, we then fine-tune a personalized colorizer to learn the definite color mapping underlying this reference view. In the second stage, the personalized colorizer is employed to infer consistent chromatic content for the remaining views or frames, which are directly leveraged to optimize the vibrant 3D scene with a tailored Lab Gaussian representation.

3.2 Stage 1: Personalized Colorizer Training

In the initial stage, we aim to fine-tune a personalized colorizer using a single colorized reference view, which enables consistent color propagation to novel views and frames. In contrast to training a standard image colorization model on diverse samples with inherent one-to-many color mappings, which results in inconsistent color predictions under minor contextual changes, our personalized

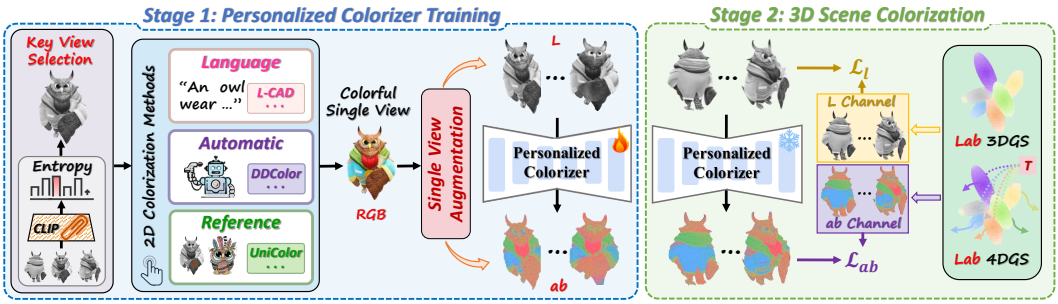


Figure 2: The overall pipeline of Color3D. Our framework comprises two primary stages. In the first stage, we initially identify the most informative key view from the given monochromatic images and video frames, and employ an off-the-shelf image colorization model to generate a colorized single view. Then, a single view augmentation scheme is elaborated to amplify the data, and the augmented samples are subsequently used to fine-tune a per-scene personalized colorizer. In the second stage, this personalized colorizer is utilized to infer consistent chromatic content of the remaining views or frames, and directly reconstruct the colorful 3D scene with Lab color space 3DGS or 4DGS.

colorizer is explicitly trained to learn the scene-specific one-to-one color mapping derived from a single image. This allows for consistent chromatic outputs across both spatial and temporal variations.

Key View Selection. Since our goal is to fine-tune a personalized colorizer using a single view, this view should capture the widest range of visual information while minimizing redundancy, which is crucial to ensure that the colorizer can be more robust to handle other views of the scene. To achieve this, we propose a heuristic strategy that selects the most informative and representative view by comparing feature similarity entropies across all candidates.

Specifically, given a set of N candidate monochromatic inputs, we first extract their single dimension feature representations using CLIP Radford et al. (2021) and the concated feature matrix can be denoted as $F \in \mathbb{R}^{N \times D}$, where D is the dimension of the feature space. To ensure that similarity comparisons reflect angular distances, we normalize each feature vector to unit length $\hat{F}_i = \frac{F_i}{\|F_i\|_2}$, for $i = 1, 2, \ldots, N$. Then, we compute the pairwise cosine similarity matrix to measure the relationships between different views $S = \hat{F}\hat{F}^T$, where S_{ij} represents the similarity between view i and view j. For each view i, we define a probability distribution P_{ij} over its similarity with all other views, and use this distribution to compute the information entropy $H(I_i)$ associated with view i:

$$H(I_i) = -\sum_j P_{ij} \log P_{ij}, \qquad P_{ij} = \frac{\exp(S_{ij})}{\sum_k \exp(S_{ik})} \tag{1}$$

A higher entropy value indicates that the view is more evenly related to all other views, meaning it encapsulates a broader and more diverse set of visual information. We define the optimal key view as the one that maximizes the entropy:

$$I^* = \arg\max_{I_i} H(I_i). \tag{2}$$

Key View Colorization. After obtaining the key view, users can apply any off-the-shelf image colorization model to perform colorization on the single view. Such a single-view-based strategy allows for the seamless integration of various mature techniques, including language-guided Weng et al. (2023); Chang et al. (2023); Liang et al. (2024), reference-based He et al. (2018); Huang et al. (2022), and automatic Kang et al. (2023); Ji et al. (2022) image colorization methods without any additional tuning or modification, thereby offering more flexible and convince controllability over both static and dynamic 3D scene colorization.

Single View Augmentation. In order to avoid overfitting and more robustly generalize to novel views and video frames, we propose a single view augmentation scheme that combines generative augmentations and traditional augmentations to generate more samples in which the same content consistently retains its color across all instances. As illustrated in Fig. 3(a), the proposed scheme first leverages generative models to produce plausible scene content beyond the provided single view, while ensuring chromatic coherence with the input to enrich the diversity of colorization supervision. Specifically, given a single colored view, we incorporate three complementary generative strategies: (1) Outpainting: we divide the image into a 2×2 grid and apply Stable Diffusion Rombach et al. (2022) to each of the four regions individually to imagine plausible scene content beyond their outer boundaries. Meanwhile, we employ LLaVA Liu et al. (2023a) to generate a descriptive caption based

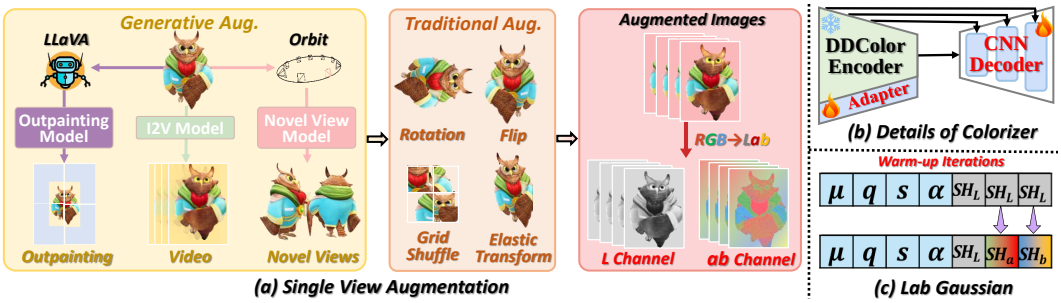


Figure 3: (a): Illustration of the proposed single view augmentation scheme that combines generative augmentations and traditional augmentations to enrich the single colored view with consistent color distribution. (b): Architecture of the colorizer consists of a frozen DDColor encoder alongside a trainable adapter and CNN decoder. (c): Lab Gaussian that first warms up with three L channels and then switches to full Lab channels for color optimization.

on the image, which serves as the text prompt to guide the diffusion process. (2) Image-to-Video: we generate continuous video frames through Stable Video Diffusion Blattmann et al. (2023) conditioned on the single image to provide coherent object motion and gradual changes in viewpoint. (3) Novel View: to simulate consistent novel viewpoints, we employ Stable Virtual Camera Zhou et al. (2025) to generate views along a predefined orbital trajectory with consistent content across perspectives. Notably, we do not require the generated content to be completely the same as the scene to be colored; instead, we emphasize preserving consistent chromatic styles with the input view, which is sufficient to improve coloring robustness and enhance coloring richness for scene content not included in the single view.

Afterward, we apply a series of traditional augmentations to further enhance the samples, including rotation, flip, grid shuffle that divides the image into grids and randomly disrupts them, and elastic transform Simard et al. (2003) that introduces smooth deformations to simulate realistic variations in object shape and structure. This can facilitate the learning of more robust and variation-agnostic color mapping.

Finally, these augmented images are converted from RGB to the CIE Lab color space Iizuka et al. (2016), where the L channel is used as the input to the colorizer and the ab channels serve as the target for color prediction. The Lab color space not only decouples luminance from chromatic components but also offers a perceptually uniform representation that better aligns with human visual perception.

Colorizer Tuning. With the augmented samples, we can fine-tune a personalized colorizer to learn the scene-specific, variation-agnostic, deterministic color mapping underlying the single view. The structure of the colorizer is illustrated in Fig. 3(b). We adopt an encoder from the pre-trained image colorization model DDColor Kang et al. (2023) and freeze its parameters to preserve its strong capability in extracting high-level color-relevant features. This enables our model to learn semantic-level color mappings from limited samples, which are inherently more robust to viewpoint and motion variations than low-level luminance-to-color correspondences. On the other hand, we deliberately avoid using a pre-trained decoder, whose built-in color priors are a major source of multi-view color inconsistency. Instead, we employ a lightweight, learnable CNN-based decoder initialized from scratch to serve as a clean and unbiased slate for learning personalized color mappings. Furthermore, we equip the encoder with additional learnable adapters Chen et al. (2024a) to better adapt to the current scenario while preserving its original pre-trained priors. The personalized colorizer is fine-tuned using a simple L1 loss:

$$\mathcal{L}_{colorizer} = \| P^{ab} - G^{ab} \|_1 \tag{3}$$

where P^{ab} and G^{ab} denote the predicted ab channels by the personalized colorizer and the ground-truth ab channels, respectively. Please refer to Appendix A.2 for additional network details and a more in-depth explanation of why the personalized colorizer can achieve consistent color propagation.

3.3 Stage 2: 3D Scene Colorization

With the personalized colorizer, in this stage, we aim to generate consistent colors for other views and video frames, which are then directly utilized to reconstruct colorful static and dynamic 3D scenes. Considering that the luminance information of the scene is already known and it can serve as reliably scene structure constraint, we advocate optimizing in CIE Lab color space Iizuka et al. (2016). By decoupling the known luminance (L channel) from the predicted chrominance (ab channels), our

approach can constrain them separately to more stably optimize the 3D scene, attenuating the effect of perturbations in the predicted chromatic information. To achieve this, we deploy 3DGSKerbl et al. (2023) and 4DGS Wu et al. (2024) to model static and dynamic 3D scenes and propose a Lab Gaussian representation for rendering images in Lab color space and facilitating separate optimization of luminance and chrominance.

Lab Gaussian. The vanilla Gaussian attributes including position (μ), rotation (q), scaling (s), opacity (α), and Spherical Harmonics (SH) coefficients (SH_R, SH_G, SH_B) for modeling RGB color channels. To better deouple luminance and chrominance, we retain the same architecture but reformulate the three sets of SH coefficients to the Lab color space as $\{SH_L, SH_a, SH_b\}$, which separately parameterize the luminance (L) and chrominance (a and b) channels. Following the original splatting pipeline Kerbl et al. (2023), we can render images with Lab color space representation directly from these SH coefficients. This strategy allows for fully leveraging the high-fidelity luminance information to provide more stable optimization signals for Gaussian primitives such as position and shape, thereby improving training stability and scene convergence.

Due to the differing numerical ranges between L and ab channels in the Lab color space, directly modeling chrominance and luminance jointly with the same Gaussian primitives can lead to unstable optimization and convergence. To mitigate this issue, we normalize rendered Lab channels into a unified range of $[0, 1]$. Specifically, the luminance channel L, originally in $[0, 100]$, is scaled by a factor of $1/100$. The chrominance channels a and b, originally spanning $[-128, 127]$, are first shifted by 128 and then scaled by $1/255$ accordingly. The normalization process can be formulated as:

$$L' = \frac{L}{100}, \quad a' = \frac{a + 128}{255}, \quad b' = \frac{b + 128}{255} \tag{4}$$

Optimization objectives. We deploy the same loss functions for both static and dynamic scenarios and otherwise remain identical with the original 3DGS and 4DGS. For the rendered image in the Lab color space, we optimize the L and ab channels separately. Since the L channel contains the core structural and high-frequency detail information of the scene, we deploy an edge loss $\mathcal{L}_{\text{edge}}$ to encourage the Gaussian primitives to better capture fine textures and structural details.

$$\mathcal{L}_{\text{edge}} = \sum \sqrt{\left(\nabla(R^L) - \nabla(G^L)\right)^2 + \epsilon^2} \tag{5}$$

where ∇ denotes the Laplacian operator, and R^L, G^L are the rendered and ground-truth luminance channels, respectively. In addition, we also deploy the $\mathcal{L}_1$ and $\mathcal{L}_{D\text{–}SSIM}$ from 3DGS and the overall objective for the L channel can be written as:

$$\mathcal{L}_l = (1-\beta)\mathcal{L}_1 + \beta\mathcal{L}_{D\text{–}SSIM} + \mathcal{L}_{edge}, \tag{6}$$

where β is set to 0.2 similar to the original 3DGS Kerbl et al. (2023). For the ab channels, which primarily contain low-frequency chromatic information, we exclude the edge loss term and utilize only the following objective:

$$\mathcal{L}_{ab} = (1-\beta)\mathcal{L}_1 + \beta\mathcal{L}_{D\text{–}SSIM}. \tag{7}$$

Warm-Up. To facilitate more accurate 3D structural and temporal motion modeling, we propose to first optimize the structure and deformation of the scene with only luminance supervision for warm-up and then co-optimize the scene with color constraints. More concretely, as illustrated in Fig. 3(c), during the first half of training iterations in 3DGS or 4DGS, we allocate all three sets of SH coefficients to represent the L channel and render a three-channel luminance image, supervised by the luminance loss $\mathcal{L}_l$. In the latter half of the training, we then reassign two of the SH coefficient sets to represent the a and b chrominance channels for color modeling, and optimize them following the above-presented objectives. This warm-up scheme stabilizes early-stage geometry learning while providing a strong initialization for color modeling, ultimately yielding more robust colorization.

4 Experiments and Analysis

4.1 Experimental Settings

Implementation details. All experiments are conducted using the PyTorch framework on NVIDIA RTX A6000 GPUs. We evaluate our method on two static datasets (LLFF Mildenhall et al. (2019) and Mip-NeRF 360 Barron et al. (2022)) and one dynamic dataset (DyNeRF Li et al. (2022)), all of which are converted to monochrome images for colorization. Structure-from-Motion (SfM) is employed to initialize Gaussian points from monochrome inputs in all experiments. Without loss of generality, both our method and competing approaches deploy the same set of image colorization models for fair comparison: ControlColor Liang et al. (2024) for language-guided colorization, DDColor Kang et al.

Table 1: Quantitative comparisons on static (LLFF and Mip-NeRF 360) and dynamic (DyNeRF) 3D scene colorization benchmarks. The top-performing results are highlighted with color.

Method	Language			Automatic			Reference		
	FID↓	CLIP Score↑	ME↓	FID↓	Colorful↑	ME↓	FID↓	Ref-LPIPS↓	ME↓
LLFF (Static 3D Scenes)									
GaussianEditor Chen et al. (2024b)	136.49	0.6346	0.093	-	-	-	-	-	-
GenN2N Liu et al. (2024)	-	-	-	52.63	25.24	0.078	-	-	-
Ref-NPR Zhang et al. (2023)	-	-	-	-	-	-	111.93	0.7958	0.112
ColorNeRF Cheng et al. (2024b)	138.62	0.6388	0.089	58.26	21.78	0.062	103.68	0.7891	0.086
3DGS+2DColorizer	87.45	0.6415	0.178	40.25	**34.07**	0.084	67.48	0.7366	0.131
Color3D (Ours)	**62.84**	**0.6544**	**0.051**	**35.10**	33.99	**0.056**	**41.36**	**0.6823**	**0.055**
Mip-NeRF 360 (Static 3D Scenes)									
GaussianEditor Chen et al. (2024b)	123.71	0.6045	0.088	-	-	-	-	-	-
GenN2N Liu et al. (2024)	-	-	-	83.69	23.44	0.137	-	-	-
Ref-NPR Zhang et al. (2023)	-	-	-	-	-	-	146.32	0.7928	0.143
ColorNeRF Cheng et al. (2024b)	163.12	0.6075	0.093	87.42	18.90	0.093	121.57	0.7852	0.106
3DGS+2DColorizer	112.33	0.6148	0.202	62.32	29.32	0.135	86.25	0.7521	0.178
Color3D (Ours)	**68.23**	**0.6246**	**0.058**	**39.03**	**33.36**	**0.082**	**48.62**	**0.7028**	**0.079**
DyNeRF (Dynamic 3D Scenes)									
Instruct 4D-to-4D Mou et al. (2024)	112.35	0.6082	0.062	-	-	-	-	-	-
4DGS+2DColorizer	89.39	0.6159	0.124	39.17	29.45	0.080	84.65	0.7446	0.128
Color3D (Ours)	**58.62**	**0.6271**	**0.041**	**37.28**	**32.75**	**0.062**	**52.77**	**0.6717**	**0.052**

(2023) for automatic colorization, and UniColor Huang et al. (2022) for reference-based colorization. Apart from our proposed Lab Gaussian representation and optimization objectives, all other settings are kept identical to the original 3DGS and 4DGS frameworks. Notably, our entir pipeline incurs only about eight additional minutes for personalized colorizer tuning, which is considered acceptable for chromatic 3D scene reconstruction.

Metrics. We adopt CLIP score Radford et al. (2021) for language-guided colorization to measure text-image alignment, Colorful Score Hasler & Suesstrunk (2003) for automatic colorization to assess color vividness, and Ref-LPIPS Zhang et al. (2023) for reference-based colorization to evaluate perceptual similarity. Fréchet Inception Distance (FID) Heusel et al. (2017) is employed as a standard metric for assessing the rendering quality of all colorization tasks. Additionally, to evaluate the consistency of colorization across views and time, we propose a metric termed Matching Error (ME). It utilizes a dense image matching model Shen et al. (2024) to identify pixel correspondences between ground-truth color images from different views or time steps, and computes the average color difference at these matched locations in the rendered colorized results.

4.2 Results on Static 3D Scene Colorization

The experimental results for quantitative comparisons in static 3D scenes are shown in the middle and upper portions of Tab. 1. It is observed that our Color3D delivers remarkable performance gains and outperforms all competitive methods significantly across diverse colorization tasks. A direct combination of 3DGS and 2D colorization models leads to significantly higher Matching Error (ME), indicating severe multi-view inconsistency. While ColorNeRF Cheng et al. (2024b) improves 3D color consistency by averaging inconsistent 2D colorizations, this strategy inevitably sacrifices color richness and controllability, resulting in noticeable degradation in FID scores and other task-specific evaluation metrics. We also demonstrate visual comparisons in Fig. 4. As suggested, our method effectively accommodates various forms of user control and accurately generates the desired colorized scenes while preserving strong multi-view consistency. In contrast, direct integration of 2D colorizers introduces obvious view inconsistency and visual artifacts, while ColorNeRF produces overly uniform color tones with limited color diversity and controllability.

4.3 Results on Dynamic 3D Scene Colorization

The quantitative and qualitative experimental results for the dynamic 3D scenarios are exhibited in the bottom portion of Tab. 1 and Fig. 5. It can be observed that our method significantly outperforms the combination of 4DGS and 2D colorizers, particularly under language-guided and reference-based settings, achieving better FID scores along with substantial reductions in Matching Error (ME) by 0.83 and 0.70, respectively. Moreover, the visual results further demonstrate that our method offers more precise controllability and superior consistency both spatially and temporally.

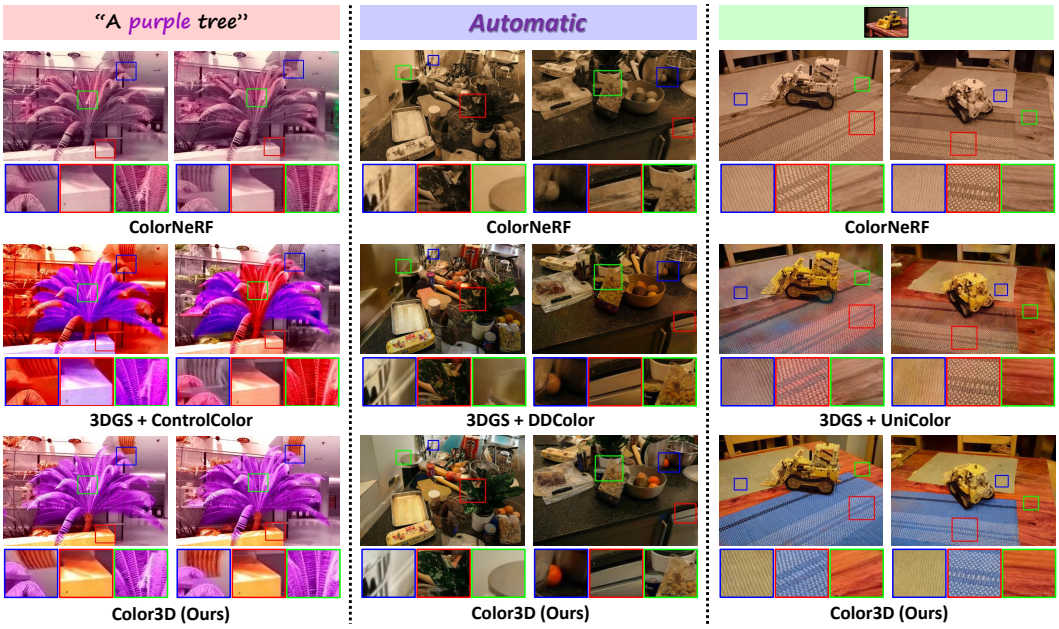


Figure 4: Qualitative comparisons on static 3D scene colorization benchmarks. Our method produces more color-accurate and color-rich results while maintaining multi-view consistency.

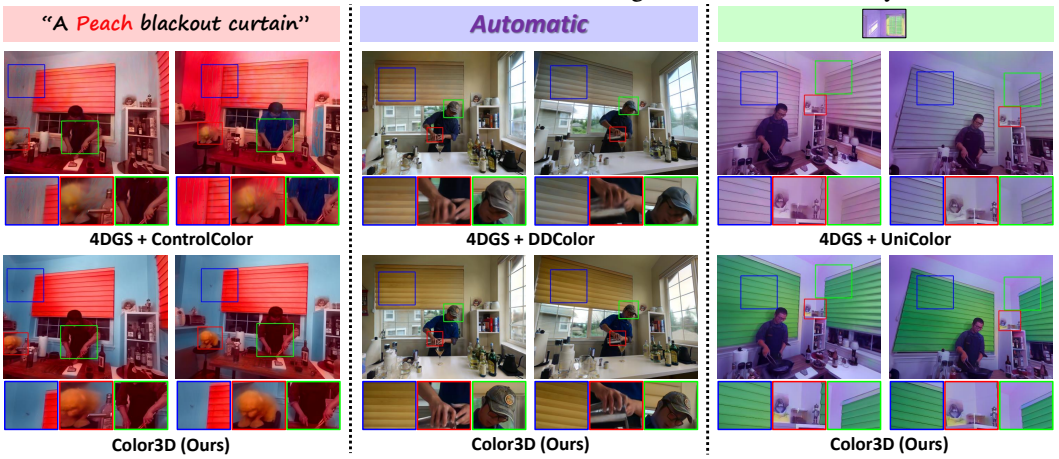


Figure 5: Qualitative comparisons on dynamic 3D scene colorization benchmarks. Our method consistently yields spatial-temporal coherent results with vivid and perceptually realistic color.

4.4 Results on Real-World Applications

To more rigorously validate the effectiveness of our proposed Color3D, we also conduct experiments on collected in-the-wild monochrome multi-view images and old movies. As illustrated in Fig. 6, our method is capable of producing realistic and vivid colorizations while maintaining impressive color consistency across viewpoints and time, demonstrating the effectiveness of our approach in revitalizing historical or legacy visual content.

4.5 Ablation Studies

In Tab. 2, we conduct ablation experiments on the dedicated components introduced in Color3D. The effectiveness of each proposed component in Color3D is evaluated by gradually integrating them into the model, revealing their individual contributions to the overall performance. More detailed analyses and visual demonstrations are as below.

Effort of key view selection. As shown in Fig. 7(a), randomly selected key view often fail to sufficiently capture the full content distribution of the scene, leading to fine-tuned colorizers exhibiting limited color richness when generalized to unseen regions of the scene. In contrast, our key view selection scheme identifies more informative and representative key view, thereby enabling the propagation of richer and more diverse colorizations across novel views.

Figure 6: *Left:* Novel views of static 3D scene colorization from in-the-wild monochrome multi-view images. *Right:* Novel views of dynamic 3D scene colorization from a historical monochrome video.

Table 2: Ablation studies on language-guided colorization on the Mip-NeRF 360 dataset.

Variant	FID↓	CLIP Score↑	ME↓
Baseline	115.66	0.6145	0.078
+ Key View Selection	100.28 (+15.38)	0.6175 (+0.0030)	0.074 (+0.004)
+ Single View Augmentation	85.25 (+15.03)	0.6196 (+0.0021)	0.069 (+0.005)
+ Fine-Tuning-Based Colorizer	75.48 (+9.77)	0.6225 (+0.0029)	0.064 (+0.005)
+ Lab Gaussian	68.23 (+7.25)	0.6246 (+0.0021)	0.058 (+0.006)

Effort of single view augmentation. Fig. 7(b) demonstrates that the proposed single view augmentation strategy can effectively extrapolate potential visual content beyond the selected view and expand the sample space, thereby facilitating the colorizer to generate more plausible and enriched colorizations for objects not originally visible in the key view and better preserve scene saturation.

Effort of personalized colorizer. As illustrated in Fig. 7(c), a randomly initialized colorizer, due to its suboptimal feature extraction capabilities, tends to cause color drifting and miscolorization when generalized to novel views. On the contrary, our personalized colorizer fine-tunes a pre-trained DDColor encoder to learn a high-level semantically aware color mapping, ensuring more accurate and consistent colorization across novel views in the scene.

Effort of Lab Gaussian. As depicted in Fig. 7(d), the vanilla RGB Gaussian representation, which entangles luminance and chrominance, tends to introduce blurring artifacts and ghosting since color perturbations will simultaneously affect all three channels. In contrast, our proposed Lab Gaussian representation decouples luminance from predicted chrominance for independent optimization, resulting in significantly sharper textures and more faithful structural details.

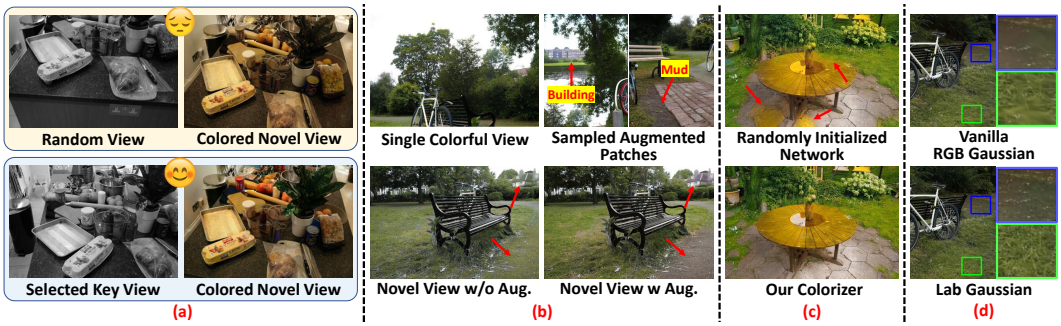


Figure 7: Visual ablation study illustrating the impact of (a) key view selection, (b) single view augmentation, (c) personalized colorizer design, and (d) Lab Gaussian representation.

5 Concluding Remarks

In this work, we propose Color3D, an innovative framework that can efficiently reconstruct colorful static and dynamic 3D scenes from monochromatic inputs with desirable controllability and consistency. In contrast to existing methods that rely on color averaging strategies to enforce multi-view consistency, which often leads to monotonous hues and uncontrollable results, our approach elegantly tackles cross-view and cross-time consistency by fine-tuning a personalized colorizer using a single colorized view. By specializing in the deterministic color mapping underlying this reference view, the colorizer can consistently propagate the intended colors to novel views and frames, preserving color consistency without sacrificing vividness and diversity. Such a design also empowers users to control the overall scene via controlling only one view, enabling flexible and concise controlled colorization. Extensive experiments on various static and dynamic 3D scene colorization benchmarks manifest the effectiveness, superiority, and controllability of our method.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5855–5864, 2021.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19697–19705, 2023.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-coins: Language-based colorization with instance awareness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19221–19230, 2023.

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pp. 333–350. Springer, 2022.

Hao Chen, Ran Tao, Han Zhang, Yidong Wang, Xiang Li, Wei Ye, Jindong Wang, Guosheng Hu, and Marios Savvides. Conv-adapter: Exploring parameter efficient transfer learning for convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1551–1561, 2024a.

Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21476–21485, 2024b.

Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. In *Forty-first International Conference on Machine Learning*, 2024a.

Yean Cheng, Renjie Wan, Shuchen Weng, Chengxuan Zhu, Yakun Chang, and Boxin Shi. Colorizing monochromatic radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1317–1325, 2024b.

Ankit Dhiman, R Srinath, Srinjay Sarkar, Lokesh R Boregowda, and R Venkatesh Babu. Corf: Colorizing radiance fields using knowledge distillation. *arXiv preprint arXiv:2309.07668*, 2023.

Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5501–5510, 2022.

Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14346–14355, 2021.

Xiang Guo, Jiadai Sun, Yuchao Dai, Guanying Chen, Xiaoqing Ye, Xiao Tan, Errui Ding, Yumeng Zhang, and Jingdong Wang. Forward flow for novel view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16022–16033, 2023.

David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pp. 87–95. SPIE, 2003.

Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016.

Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. Colorformer: Image colorization via color memory assisted hybrid-attention transformer. In *European Conference on Computer Vision*, pp. 20–36. Springer, 2022.

Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. Ddcolor: Towards photo-realistic image colorization via dual decoders. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 328–338, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. Bigcolor: Colorization using a generative color prior for natural images. In *European Conference on Computer Vision*, pp. 350–366. Springer, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Jiaxing Li, Hongbo Zhao, Yijun Wang, and Jianxin Lin. Towards photorealistic video colorization via gated color-guided image diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10891–10900, 2024a.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5521–5531, 2022.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8508–8520, 2024b.

Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Control color: Multimodal diffusion-based interactive image colorization. *arXiv preprint arXiv:2402.10855*, 2024.

Jiecheng Liao, Shi He, Yichen Yuan, and Hui Zhang. Gbc: Gaussian-based colorization and super-resolution for 3d reconstruction. In *The 19th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pp. 1–9, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Xiangyue Liu, Han Xue, Kunming Luo, Ping Tan, and Li Yi. Genn2n: Generative nerf2nerf translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5105–5114, 2024.

Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13–23, 2023b.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Linzhan Mou, Jun-Kun Chen, and Yu-Xiong Wang. Instruct 4d-to-4d: Editing 4d scenes as pseudo-3d scenes using 2d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20176–20185, 2024.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15, 2022.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5865–5874, 2021a.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16632–16642, 2023.

Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*, 2024.

Patrice Y Simard, David Steinkraus, and John C Platt. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*. IEEE, 2003.

Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023.

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5481–5490. IEEE, 2022.

Feng Wang, Zilong Chen, Guokang Wang, Yafei Song, and Huaping Liu. Masked space-time hash encoding for efficient dynamic scene reconstruction. *Advances in neural information processing systems*, 36:70497–70510, 2023.

Han Wang, Yuang Zhang, Yuhong Zhang, Lingxiao Lu, and Li Song. Consistent video colorization via palette guidance. *arXiv preprint arXiv:2501.19331*, 2025.

Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. L-code: Language-based colorization using color-object decoupled conditions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2677–2684, 2022.

Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, Boxin Shi, et al. L-cad: Language-based colorization with any-level descriptions using diffusion priors. *Advances in Neural Information Processing Systems*, 36:77174–77186, 2023.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20310–20320, 2024.

Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14377–14386, 2021.

Jinbo Yan, Rui Peng, Luyang Tang, and Ronggang Wang. 4d gaussian splatting with scale-aware residual field and adaptive optimization for real-time rendering of temporally complex dynamic scenes. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7871–7880, 2024.

Yixin Yang, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Colormnet: A memory-based deep spatial-temporal feature propagation network for video colorization. In *European Conference on Computer Vision*, pp. 336–352. Springer, 2024a.

Yixin Yang, Jinshan Pan, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, and Jinhui Tang. Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20331–20341, 2024c.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19447–19456, 2024.

Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image colorization with text guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017.

Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, and Jiaya Jia. Ref-npr: Reference-based non-photorealistic radiance fields for controllable scene stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4242–4251, 2023.

Hengyuan Zhao, Wenhao Wu, Yihao Liu, and Dongliang He. Color2embed: Fast exemplar-based image colorization using color embeddings. *arXiv preprint arXiv:2106.08017*, 2021.

Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.

A Appendix

A.1 3DGS & 4DGS

3D Gaussian Splatting (3DGS) is an emerging and highly effective representation for novel view synthesis, which models a 3D scene as a set of G anisotropic Gaussians. Each Gaussian g_i defines a spatial density function:

$$g_i(x) = \exp\left(-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)\right), \tag{8}$$

where $1 \le i \le G$, $\mu_i \in \mathbb{R}^3$ denotes the mean (i.e., the Gaussian center), and $\Sigma_i \in \mathbb{R}^{3\times3}$ is the covariance matrix encoding its anisotropic shape. In practice, Σ_i is parameterized by a positive scaling vector $s_i \in \mathbb{R}^3_+$ and a unit quaternion $q_i \in \mathbb{R}^4$ representing its orientation. Each Gaussian is further associated with an opacity value $\alpha_i \in \mathbb{R}_+$ and a set of spherical harmonics (SH) coefficients c_i for view-dependent RGB color modeling.

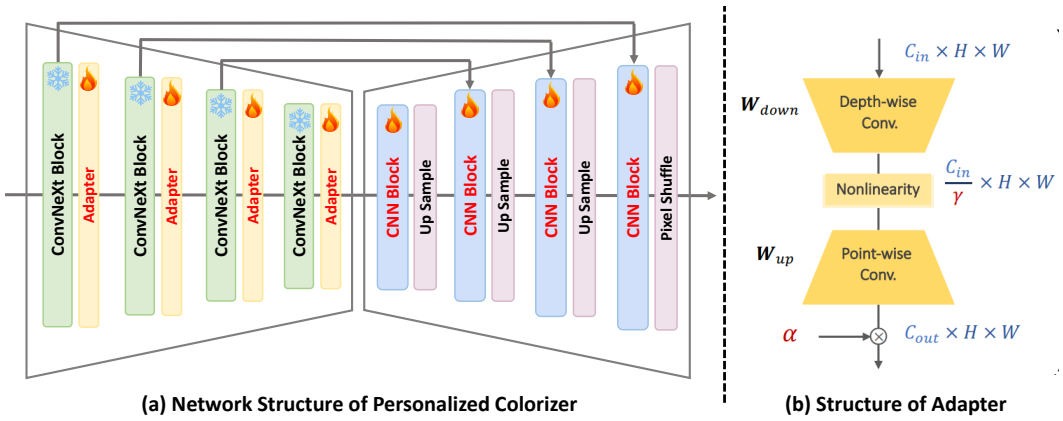
4D Gaussian Splatting (4DGS) extends 3DGS by introducing a learnable deformation field, allowing dynamic scenes to be modeled via a canonical set of 3D Gaussians. Temporal dynamics are captured by predicting time-dependent offsets in position, rotation, and scale, formulated as:

$$g_i = (\mu_i + \delta\mu_i, q_i + \delta q_i, s_i + \delta s_i, \alpha_i, c_i), \quad (\delta\mu_i, \delta q_i, \delta s_i) = \mathcal{F}(\mu_i, t), \tag{9}$$

where $\mathcal{F}$ indicates deformation prediction module.

A.2 More Method Details and Analysis

Structure details of pensorlized colorizer. As shown in Fig. 8(a), our personalized colorizer adopts an encoder-decoder structure, augmented with lightweight adapter modules to enable efficient per-scene fine-tuning. The encoder comprises several pre-trained ConvNeXt blocks, each equipped with an inserted adapter. During fine-tuning, the backbone ConvNeXt blocks from pre-trained DDColor Kang et al. (2023) are frozen, and only the adapters and decoder are updated. Each adapter (Fig. 8(b)) follows a bottleneck design consisting of a depth-wise convolution, nonlinearity, and point-wise convolution. Given an input feature map, the adapter first reduces its channel dimension, applies a nonlinear transformation (ReLU), and projects it back to the original size. A residual connection modulated by an attention weight further refines the adaptation. The decoder is composed of a stack of simple and lightweight CNN blocks, each consisting of two convolutional layers, one instance normalization layer, and LeakyReLU serves as activation function. Our structure design effectively allows the colorizer to specialize in learning scene-specific color mappings from given views while maintaining the semantic representation capacity gained from the pre-trained image colorization model.



(a) Network Structure of Personalized Colorizer (b) Structure of Adapter

Figure 8: (a) Overall architecture of the proposed personalized colorizer. Only the adapters and decoder are updated during fine-tuning. (b) Structure of the adapter module Chen et al. (2024a), which efficiently injects scene-specific colorization capabilities.

Why can the personalized colorizer achieve consistent color propagation? Here, we intuitively explain why a colorizer trained on augmented samples from a single colorized view can maintain color consistency when generalized to novel viewpoints and video frames. We illustrate this by comparing a standard image colorization model with our per-scene personalized colorizer.

- **Standard image colorization model.** Typically, training a general-purpose image colorization model requires large-scale luminance-color image pairs. However, the same semantic object may appear in different colors across images, for example, a door might be white in one image but black in another. Consequently, the model learns an uncertain one-to-many color mapping and tend to infer colors based on contextual cues rather than object identity, making it prone to inconsistent color predictions when facing slight viewpoint or motion changes.
- **Our per-scene personalized colorizer.** In contrast, our personalized colorizer is trained on augmented samples derived from a single key view, where the same objects maintain consistent colors across all variations. These augmentations simulate changes in deformation, context, viewpoint, motion, unseen content, etc. This setup encourages the model to learn a scene-specific, variation-agnostic, one-to-one color mapping. Leveraging the inherent inductive bias of deep neural networks Neyshabur et al. (2017); Battaglia et al. (2018), the trained colorizer can predict consistent colors for novel views and video frames by mapping the same content to the corresponding colors it learned during training.

A.3 More Experiments

Dataset details.

LLFF Dataset Mildenhall et al. (2019) consists of 8 forward-facing real-world scenes. Following prior work, we select every eighth image for testing and use the remaining images for training. All images are processed at a resolution of 1008×756 for our experiments.

Mip-NeRF 360 Dataset Barron et al. (2022) comprises 9 real-world scenes, including 5 outdoor and 4 indoor environments. Following the same protocol as LLFF, we use one-eighth of the views for testing and the remaining views for training. All experiments are conducted on images downsampled by a factor of 4 (approximately 1K resolution).

DyNeRF Dataset Li et al. (2022) includes six 10-second video sequences captured at 30 fps by 15 to 20 cameras with face forward perspective, involving extended periods and intricate camera motions.

More implementation details of pensorlized colorizer tuning. For the single view augmentation strategy, we first generate nine augmented samples based on the original view. Specifically, four samples are produced using the outpainting model Stable Diffusion Rombach et al. (2022), one sample is extracted from the final frame generated by the image-to-video model Stable Video Diffusion Blattmann et al. (2023), and three samples are obtained by sparsely sampling along an orbit using the novel view synthesis model Stable Virtual Camera Zhou et al. (2025), together with the original view itself. The nine samples were then randomized to apply traditional augmentation operations. During training, the samples input to the personalized colorizer are randomly cropped to 320×320. We train the personalized colorizer for 4K iterations using the Adam optimizer. During training, only one sample is used per iteration, i.e., $batchsize = 1$, which allows the model to focus on capturing the most distinctive and rich color information from the given sample. The learning rate is initialized to 1×10^4, which is steadily decreased to 1×10^6 using the cosine annealing strategy.

Image and video colorizers vs. per-scene colorizer. We evaluate the effectiveness of our per-scene colorizer design by comparing it with two representative baselines: the image colorization model DDColor Kang et al. (2023) and the video colorization method ColorMNet Yang et al. (2024a), as illustrated in Fig. 9. It can be observed that DDColor suffer from severe cross-view color inconsistencies due to its per-image one-to-many color mapping nature. Treating multi-view images as a video sequence and applying video colorization similarly fails to resolve this issue, since the viewpoint changes inherent in 3D capture far exceed the temporal variations commonly seen in video data. In contrast, our personalized colorizer learns a scene-specific, one-to-one color mapping from the reference view and leverages its inherent inductive bias to consistently project colors to corresponding content in novel views and time steps. This design effectively addresses the challenge of color consistency in both static and dynamic 3D scenes.

Figure 9: Visual ablation study illustrating the effectiveness of the per-scene colorizer design.

Table 3: Quantitative comparisons of language-guided 3D scene colorization on Mip-NeRF 360 and DyNeRF datasets.

Method	Language-Guided Colorization		
	FID↓	CLIP Score↑	ME↓
Mip-NeRF 360 (Static 3D Scenes)			
NeRF+2DColorizer	104.85	0.6153	0.225
3DGS+2DColorizer	112.33	0.6148	0.202
Color3D-NeRF	76.68	0.6213	0.065
Color3D (Ours)	**68.23**	**0.6246**	**0.058**
DyNeRF (Dynamic 3D Scenes)			
NeRFPlayer+2DColorizer	86.77	0.6148	0.133
4DGS+2DColorizer	89.39	0.6159	0.124
Color3D-NeRFPlayer	63.29	0.6243	0.046
Color3D (Ours)	**58.62**	**0.6271**	**0.041**

Effort of 3D representations. We further investigate how different 3D representations impact colorization performance. As shown in Tab.3, compared to 3DGS, combining NeRFMildenhall et al. (2021) or NeRFPlayer Song et al. (2023) with 2D colorizers yields suboptimal consistency, reflected by an increase of 0.023 in ME. This is mainly because the explicit modeling of 3DGS gives it an inherent resistance and geometric priors, which can naturally alleviate color inconsistency to some extent. In addition, we replace the Lab Gaussian representation in Color3D with NeRF and NeRFPlayer, respectively. The results underscore the effectiveness of our proposed Lab Gaussian representation while also demonstrating that the Color3D framework can be adapted to various 3D reconstruction backbones with reasonable performance.

Can editing or stylization methods work for colorization? In Tab. 1 of the main text, we present quantitative comparisons with a representative 3D editing method (GaussianEditor Chen et al. (2024b)) and a 3D stylization method (Ref-NPR Zhang et al. (2023)), both of which yield suboptimal performance on colorization tasks. To demonstrate their visual limitations, we further provide qualitative comparisons in Fig. 10.

It is observed that GaussianEditor struggles to comply with color editing instructions when applied to monochromatic radiation fields, primarily prioritizing semantic fidelity over chromatic accuracy. As a result, the edited outputs, although containing colors, often lack chromatic realism and diversity, and may also suffer from structural artifacts. Similarly, Ref-NPR primarily focuses on transferring global appearance by relying on perceptual constraints, but fails to accurately capture local color values. In contrast, colorization aims to produce plausible, detailed, and spatially coherent colors that align with the underlying scene content. Consequently, editing and stylization methods are inherently ill-suited for faithful 3D scene colorization. Moreover, their controllability remains limited: editing is typically driven by text prompts, while stylization relies on reference images to guide appearance transfer.

Universal color manipulation framework. Beyond colorizing 3D scenes from monochromatic inputs, our proposed scheme can also be naturally extended to 3D scene recoloring. Specifically, we apply InstructPix2Pix Brooks et al. (2023) to edit the colors of a single key view and leverage SAM

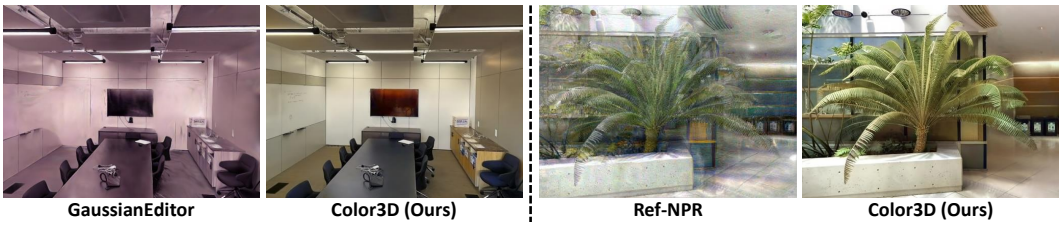


Figure 10: Visual comparisons with the 3D editing method GaussianEditor Chen et al. (2024b), and the 3D stylization method Ref-NPR Zhang et al. (2023).

Kirillov et al. (2023) to constrain the editing within specified regions. For the personalized colorizer, the input is the original image, and the output is its recolored version. As illustrated in Fig. 11, our method not only enables colorization from monochromatic inputs, but also supports high-fidelity and consistent color manipulations, demonstrating its robustness and versatility.

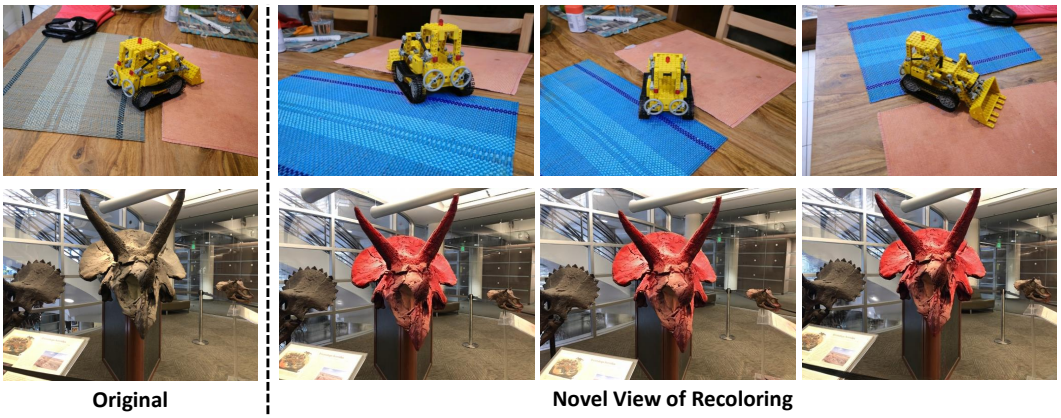


Figure 11: Visual results of 3D recoloring.

User study. To further assess the perceptual quality of the generated results, we conduct a comprehensive user study, acknowledging that subjective human perception remains a critical benchmark in the field of colorization. Specifically, we design 5 evaluation questions covering multiple aspects, including color richness, consistency, aesthetic preference, image quality, and the fidelity of alignment between the synthesized views and the provided conditions (text, exemplars). Participants are asked to rate each sample on a scale of 0 to 5 for each aspect. For a fair comparison, we recruit 30 participants, each evaluating 10 sets of comparisons across different methods. Aggregated results are shown in Fig. 12. The naive combination of 3DGS with off-the-shelf 2D colorizers yields reasonably rich colors but suffers from significant inconsistency and reduced visual quality. While ColorNeRF demonstrates better consistency, it often lacks color diversity and visual appeal. In contrast, our method consistently achieves higher user ratings across all aspects, highlighting its advantage in terms of perceptual quality and user preference.

More visual results. We further present additional colorization renderings of our Color3D in Fig. 13, Fig. 14, and Fig. 15. Our approach exhibits strong consistency across both spatial and temporal dimensions, delivering vibrant and semantically coherent colorizations that faithfully adhere to user-specified intent.

A.4 Limitations and Future Work

One potential limitation, shared with other single-view methods, is that our approach tends to favor assigning plausible colors consistent with the observed input rather than hallucinating entirely novel chromatic content when confronted with highly out-of-domain views (e.g., reconstructing a large room with multiple unseen bedrooms). In future work, we plan to further explore the integration of generative priors to enrich color diversity under drastically unseen viewpoints while maintaining color consistency. Moreover, the per-scene personalization strategy introduced in this work is highly general and carries substantial potential beyond colorization. Extending this paradigm to

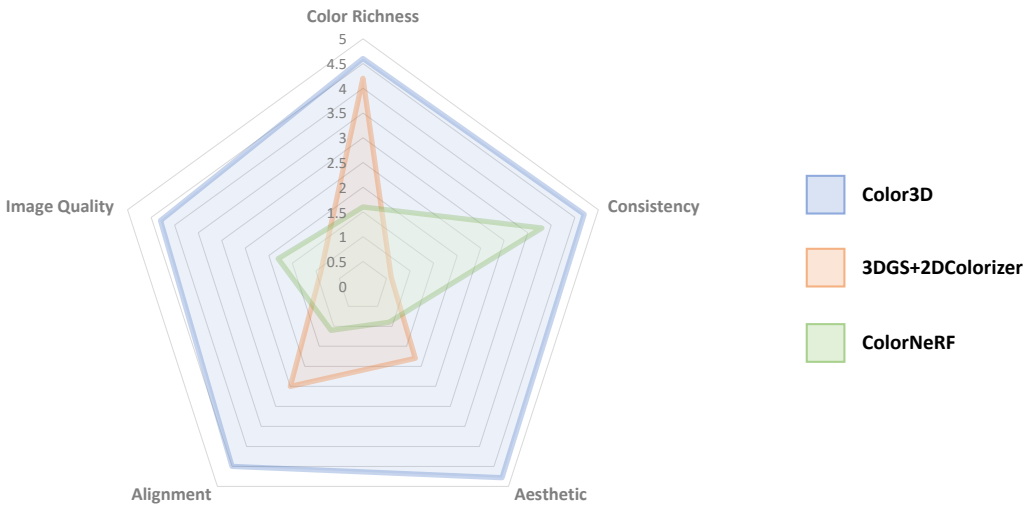


Figure 12: User study results. Our method demonstrates superior performance across all evaluated aspects.

other scene attributes—such as illumination enhancement, white balance adjustment, and style transfer—represents a promising avenue toward broader and more versatile scene-level editing capabilities.

Figure 13: Visual results of language-guided 3D scene colorization. From top to bottom are samples from the LLFF, Mip-NeRF 360, and DyNeRF datasets.

Figure 14: Visual results of automatic 3D scene colorization. From top to bottom are samples from the LLFF, Mip-NeRF 360, and DyNeRF datasets.

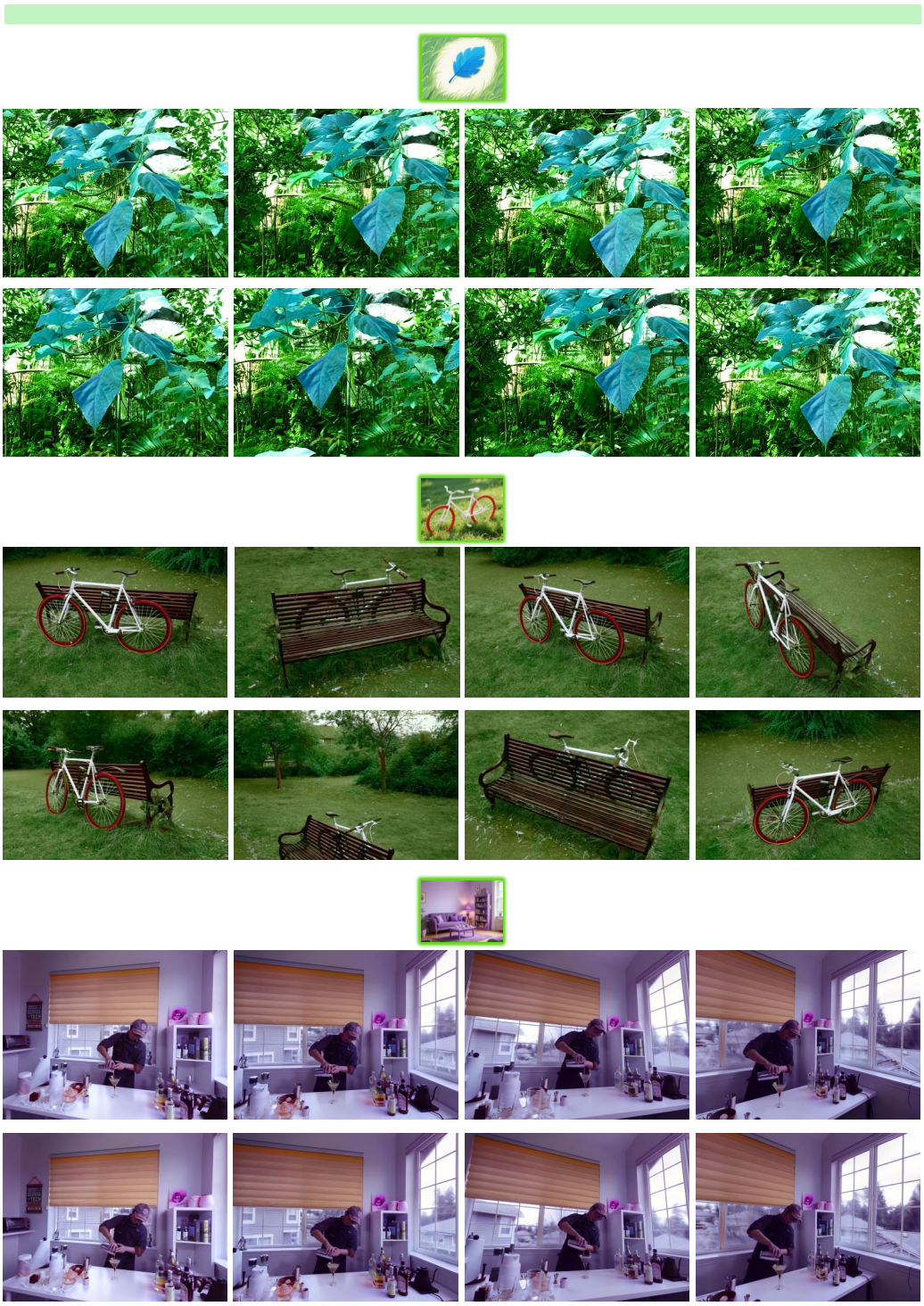


Figure 15: Visual results of reference-based 3D scene colorization. From top to bottom are samples from the LLFF, Mip-NeRF 360, and DyNeRF datasets.