

Beyond Fixed Luminance: Towards Panchromatic and Orthochromatic Image Colorization

Swarnim Maheshwari¹, Syed Imam Ali¹, and Vineeth N. Balasubramanian^{1,2}

¹ Indian Institute of Technology Hyderabad, Kandi, Telangana, India
{cs25mtech02006, ai24mtech14005}@iiith.ac.in

² Microsoft Research India, Bengaluru, Karnataka, India
vineethnb@cse.iiith.ac.in, vineeth.nb@microsoft.com

Abstract. Most image colorization systems operate in Lab space by predicting chroma (ab) while preserving an input-derived luminance channel (L). While effective on standard benchmarks, this fixed-luminance design restricts brightness changes and becomes unreliable when grayscale formation deviates from natural-image luminance, as in historical orthochromatic photography. We propose a luminance-agnostic colorization framework that formulates colorization as full-RGB image editing using a foundation image-editing model. To bridge modern panchromatic and historical orthochromatic conditions, we introduce a mixed grayscale objective that trains the model under both standard luminance grayscale and a red-insensitive grayscale formation. Experiments on COCO, ImageNet, and a multi-instance benchmark show that our method is competitive on standard grayscale inputs and substantially more robust under orthochromatic inputs, with qualitative comparisons and a human study indicating fewer visible color artifacts.

Keywords: Image Restoration · Image Colorization · Diffusion models

1 Introduction

The paradigm across modern approaches for image colorization follows a consistent formulation: convert an RGB image to the Lab color space [8], preserve the luminance channel L , and predict chrominance components ab conditioned on that fixed luminance. This design reduces the learning problem to chroma estimation while treating luminance as ground truth. While effective for reconstruction benchmarks, this formulation imposes a strict and often overlooked constraint: the model is mathematically incapable of altering brightness (Figure 1). Because L is treated as immutable, the network can only produce colors that are consistent with the original grayscale intensity. This restriction fundamentally limits creative recolorization. For example, mapping a dark grayscale region to a vibrant yellow or red surface is physically infeasible under fixed- L constraints, since bright colors require high luminance energy in RGB space (see the d. Rose and e. Fire Ext. columns of Figure 3). Thus, current methods operate under a constrained prediction regime rather than generative restoration.

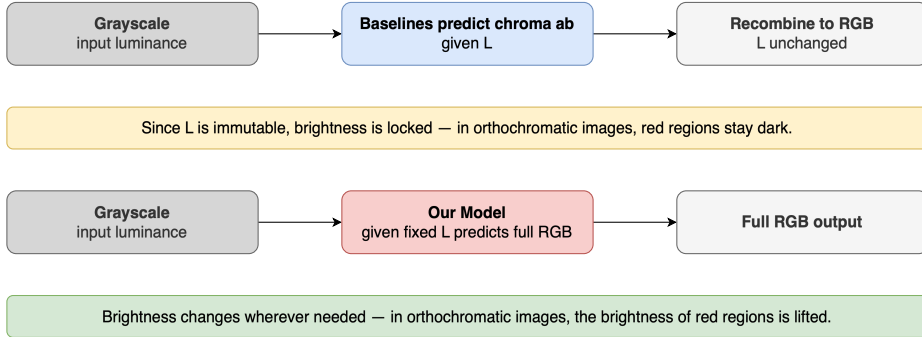


Fig. 1: Comparison of the conventional colorization paradigm (**top**) with our luminance-agnostic approach (**bottom**).

The issue becomes more severe when applied to historical photography. Early orthochromatic film stock was insensitive to red wavelengths, producing grayscale images whose luminance distribution does not correspond to human visual perception. Modern colorization models, trained on panchromatic data and conditioned on preserved luminance, inherit a structural domain mismatch. They are forced to assign chroma to brightness values that were never perceptually accurate, leading to suppressed reds. In this sense, current systems fail by design: when the input grayscale image is darker than it should be, the model is incapable of brightening those regions (Figure 3).

Most contemporary approaches that adapt pretrained text-to-image diffusion models do so by changing the architecture to condition them on grayscale [2, 5, 27–29, 42]. While somewhat effective, these approaches inherit a fundamental mismatch: text-to-image models are not pretrained for image-conditioned tasks. As a result, significant architectural engineering is required to “force” grayscale structure into a generative pipeline originally designed for synthesis from text alone. Recent foundation models natively trained on image editing provide a superior alternative for colorization without architectural hacking [24, 25, 37].

In this work, we propose a luminance-agnostic framework for image colorization. Instead of predicting chroma conditioned on fixed L , we fine-tune a foundation image-editing model to predict the full RGB output directly from grayscale input—without luminance swapping. Our approach leverages the strong generative and editing prior of the FLUX.2-klein (4B) model. We choose FLUX.2-klein because it supports high-quality few-step generation, reducing inference time to a few seconds per image. Our key contributions are three-fold: (i) we identify the fixed-luminance bottleneck in Lab-based colorization and show why it is especially problematic for orthochromatic imagery; (ii) we propose a luminance-agnostic colorization framework that formulates colorization as full-RGB image editing using a foundation image-editing model; and (iii) we perform a comprehensive suite of experiments which show that the proposed method, while remaining competitive on standard grayscale inputs, is substantially more robust

on orthochromatic inputs and also improves prompt-following behavior compared to baselines.

2 Related Work

2.1 Image colorization

Among early **CNN colorizers**, CIC [43] framed colorization as classification in quantized CIELAB to avoid desaturated outputs; InstColor [34] colorizes instance crops with a fusion module (but can suffer from erroneous external priors and color overflow); and DISCO [39] uses a coarse-to-fine, anchor-based scheme to learn global affinities and reduce color ambiguity.

GAN-based colorization methods trade realism, speed, and cost: ChromaGAN [35] adopts a PatchGAN discriminator with a WGAN loss for realism, HistoryNet [17] augments generation with classification/segmentation modules and a large old-movie dataset, DeOldify [1] speeds training via asynchronous generator/discriminator updates, ToVivid [38] uses pretrained BigGAN [3] for inversion but suffers from inversion inaccuracies on grayscale inputs, and BigColor [20] embeds BigGAN’s [3] generator/discriminator into an encoder-generator model. Time-Travel Rephotography [30] projects old portrait photographs into the manifold of modern high-resolution face images using a StyleGAN2 prior, jointly performing restoration, enhancement, and recolorization; notably, it also considers different historical film types, including orthochromatic negatives, although it is specialized to face portrait imagery.

Transformer-based colorization has advanced quickly: ColTran [23] proposed a multi-stage transformer colorizer but suffers from limited CNN-like inductive bias; ColorFormer [16] adds a global-local hybrid self-attention and a color memory for efficient semantic-color mapping; ct2 [36] derives 313 meaningful color tokens from color-space statistics and uses adaptive attention to link them to luminance; DDColor [19] introduces learnable color tokens with cross-attention fusion of grayscale and color features; and MultiColor [10] builds on DDColor with a multi-branch design to better capture colors using different color spaces.

Diffusion-based colorization from text-to-image backbones. Recent work has adapted pretrained text-to-image diffusion models for colorization by injecting grayscale structure into the generative process. Existing strategies include ControlNet-style conditioning branches [27–29, 42], mid-layer feature injection through extended convolutional modules [5], and formulations that replace the initial noise with grayscale image structure [40]. These methods demonstrate that large diffusion priors can improve semantic colorization quality, but they typically require additional architectural modifications to make text-to-image models usable for image-conditioned restoration tasks.

2.2 Image context diffusion models

More recently, a new class of foundation generative models has been developed specifically for image editing and in-context image generation. Examples include

Qwen-Image [37], FLUX.1 Kontext [25], and FLUX.2-klein [24]. Such models are trained to preserve and transform image content directly, making them a more natural backbone for colorization than text-to-image models that must be retrofitted with grayscale conditioning. Our method follows this direction by formulating colorization as an image-editing problem rather than as chroma prediction under a fixed luminance constraint.

3 Methodology

Our overall objective: A core contribution of our work is the Mixed Grayscale Objective, designed to bridge the gap between modern digital intensity and historical film sensitivity [41].

To this end, given a ground-truth colorful image $I \in \mathbb{R}^{H \times W \times 3}$ with channels I_R, I_G, I_B , we generate two distinct types of grayscale conditioning images, denoted as I_{gray} .

For the **Standard Mode**, we simulate modern panchromatic film using the standard luminance formulation, conditioned with the text prompt $c_{pan} = \text{“colorize”} + \text{Caption}$:

$$I_{gray}^{pan} = 0.299 \cdot I_R + 0.587 \cdot I_G + 0.114 \cdot I_B \quad (1)$$

For the **Orthochromatic Mode**, we simulate historical red-insensitive film using a simplified approximation motivated by the fact that orthochromatic materials are sensitive primarily to blue and green wavelengths and insensitive to red [11, 12, 21, 22]. To simulate orthochromatic images from modern-day color images, we take inspiration from prior work on historical photo restoration that explicitly models different negative film types [30]. We therefore exclude the red channel and define

$$I_{gray}^{ortho} = \frac{I_B + I_G}{2}, \quad (2)$$

conditioned with the text prompt $c_{ortho} = \text{“colorize ortho”} + \text{Caption}$. Here, “colorize” acts as a trigger word for the action of colorization, while “ortho” acts as a trigger word to notify the model that the input is an orthochromatic image. During training, we encode the ground-truth target image into the latent space using the frozen VAE, yielding the data latent $z_1 = \mathcal{E}(I)$. Because the Flux architecture is based on Rectified Flow (a formulation of continuous-time diffusion models), the forward process constructs noisy latents z_t by linearly interpolating between standard Gaussian noise $z_0 \sim \mathcal{N}(0, I)$ and the target latent z_1 over timestep $t \in [0, 1]$:

$$z_t = tz_1 + (1 - t)z_0 \quad (3)$$

The DiT is trained to predict the vector field (velocity) v_θ that transports the noise distribution to the data distribution. The network is conditioned on the text embedding $c_{text} \in \{c_{pan}, c_{ortho}\}$ and the spatially concatenated grayscale

image $I_{gray} \in \{I_{gray}^{pan}, I_{gray}^{ortho}\}$. The generative diffusion objective $\mathcal{L}$ is minimized as follows:

$$\mathcal{L} = \mathbb{E}_{z_0, z_1, t} [\|v_\theta(z_t, t, c_{text}, I_{gray}) - (z_1 - z_0)\|_2^2] \quad (4)$$

By training on this objective, the model learns to interpret luminance values differently based on the provided text prompt. Instead of replacing the predicted luminance with a constrained L channel during inference, the network maps the conditional inputs directly to a full RGB representation. This allows it to “lift” the brightness of dark regions in orthochromatic inputs where a standard L -constrained model would otherwise fail.

Architecture and Training Configuration: We employ FLUX.2-klein-4B as our foundational generative prior, leveraging its robust semantic understanding for high-fidelity color synthesis. To maintain computational efficiency while adapting the model to the colorization task, we utilize Low-Rank Adaptation (LoRA) applied exclusively to the Diffusion Transformer (DiT) blocks. Both the VAE (Image Encoder/Decoder) and the Qwen3-4B Text Encoder remain frozen throughout the process. Our LoRA configuration utilizes a rank of 16 and an alpha of 16, resulting in approximately 23.10 million trainable parameters, representing a mere 0.5% of the total 4.3B parameter count.

Optimization and Training Procedure: The model was trained using the AdamW optimizer with a constant learning rate of 1×10^{-4} and a 100-step warmup period. We utilize bfloat16 mixed-precision and enable gradient checkpointing. Training was executed across a distributed setup of three NVIDIA RTX 6000 Ada GPUs with an effective batch size of 12. The model was trained for 10 epochs on a subset of ImageNet from Kaggle [18], totaling 34,745 images.

4 Experiments

4.1 Datasets

To evaluate our proposed method across different visual domains, we utilize subsets of three widely adopted benchmarks: **ImageNet** [33]: We sample the first 5,000 images from the validation set. **COCO** [4]: We use the first 5,000 images from the test set. **Multi-Instance** [6]: The complete validation set, consisting of 7,213 images. For ImageNet and COCO, we used BLIP2 [26] to generate captions, while for Multi-Instance, human-annotated captions were provided with the dataset.

4.2 Evaluation Metrics

We evaluate on both perceptual realism and the statistical distribution of the generated colors.

Perceptual Quality Metrics: We utilize the PyTorch Image Quality Assessment [7] framework to compute the Fréchet Inception Distance (FID) [14], spatial FID (sFID) [9], and FID-DINO, which is the same as FID but uses

DINOv2 [32] features instead of InceptionNet. These metrics measure the distributional shift between the synthesized and ground-truth images.

Chroma and Colorfulness Metrics: To evaluate both the perceptual vibrancy and the statistical distribution of the generated colors independent of luminance, we define a unified set of metrics.

First, we quantify human-perceived vibrancy using the standard Colorfulness metric [13] widely used in image colorization work, which is computed using the mean (μ) and standard deviation (σ) of the opponent color spaces $rg = R - G$ and $yb = \frac{1}{2}(R + G) - B$:

$$C = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2} \quad (5)$$

For a dataset of N images, let μ_a, μ_b and σ_a, σ_b represent the mean and standard deviation of the a and b color channels in the LAB color space for a given image I_k :

- **Col-diverse (Chroma Diversity):** Measures the amount of chromatic variance generated within an image, calculated as the Euclidean combination of the channel standard deviations:

$$\text{Col-diverse} = \frac{1}{N} \sum_{k=1}^N \sqrt{\sigma_{a,k}^2 + \sigma_{b,k}^2} \quad (6)$$

- **Saturation (Mean LAB Chroma):** Calculates the average per-pixel chroma in the LAB space. Let a_p and b_p be the channel values for pixel p , and $|I_k|$ be the total number of pixels in image I_k :

$$\text{Saturation} = \frac{1}{N} \sum_{k=1}^N \left(\frac{1}{|I_k|} \sum_{p \in I_k} \sqrt{a_p^2 + b_p^2} \right) \quad (7)$$

Colornet (Perceptual Colorfulness) [41]: While traditional metrics rely on heuristic formulations in opponent color spaces, we also evaluate the perceptual vibrancy of our generated outputs using ColorNet. Developed by Zerman et al., ColorNet is a Convolutional Neural Network (CNN) specifically trained on subjective human datasets to predict Mean Opinion Scores (MOS) of image colorfulness. By leveraging deep feature representations rather than simple statistical variances, this learned metric provides an assessment of chromatic intensity that aligns with the human visual system.

Baselines: We use the following baselines: BigColor [20], COCO-LC [27], DDColor [19], DISCO [39], and UniColor [15]. We use BLIP2 to generate captions for both COCO-LC and our method on the ImageNet and COCO datasets, and use the provided human-labeled captions for the Multi-Instance dataset. We evaluate all baselines by converting each image to grayscale and feeding it into the model. For the ortho setting we used I_{gray}^{ortho} , while for the pan setting we used the standard grayscale formulation I_{gray}^{pan} .

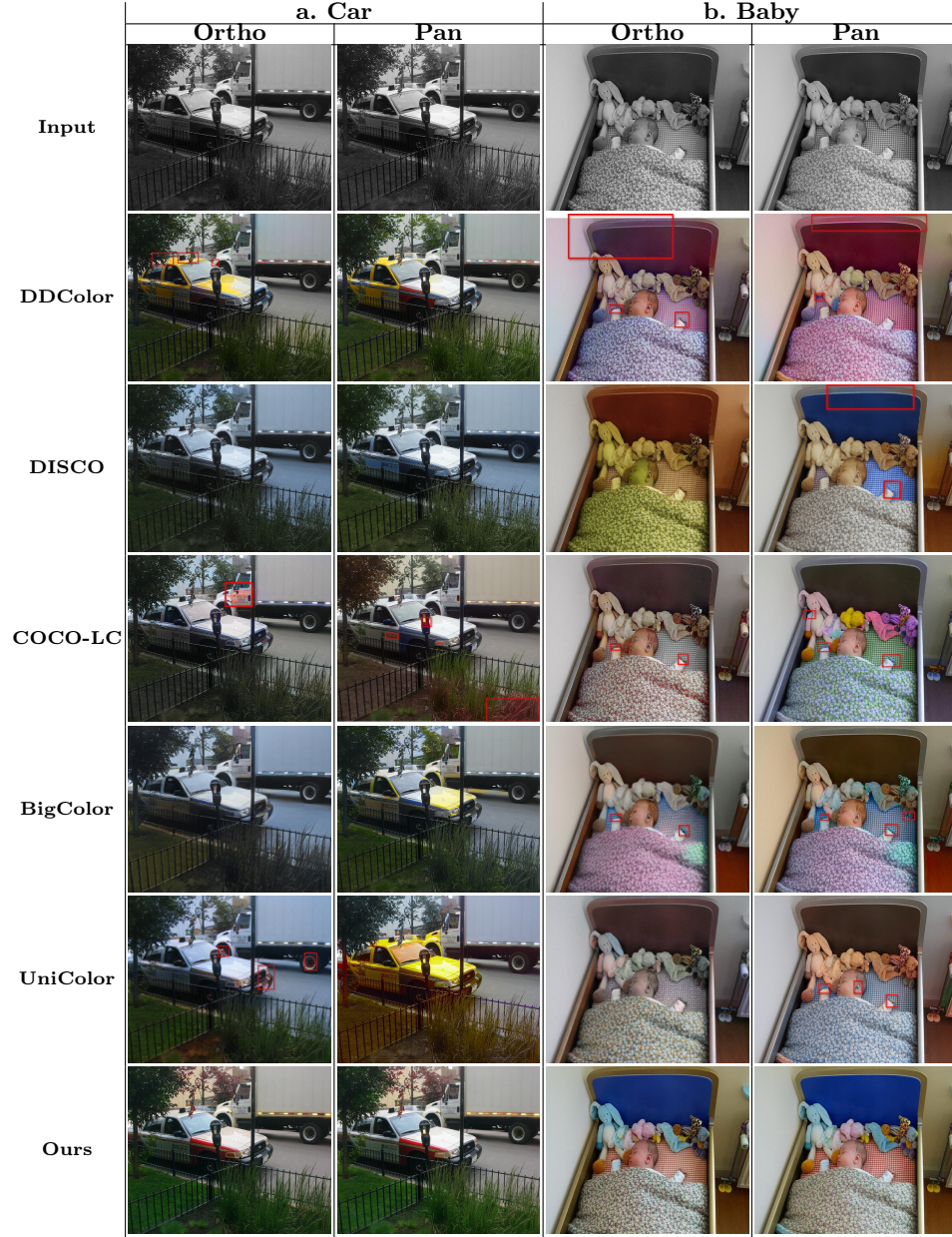


Fig. 2: Notice in the b. Baby images that for most methods (except ours), the color and pattern from the bedsheets bleed into the hands. (**Zoom-in for best view**)

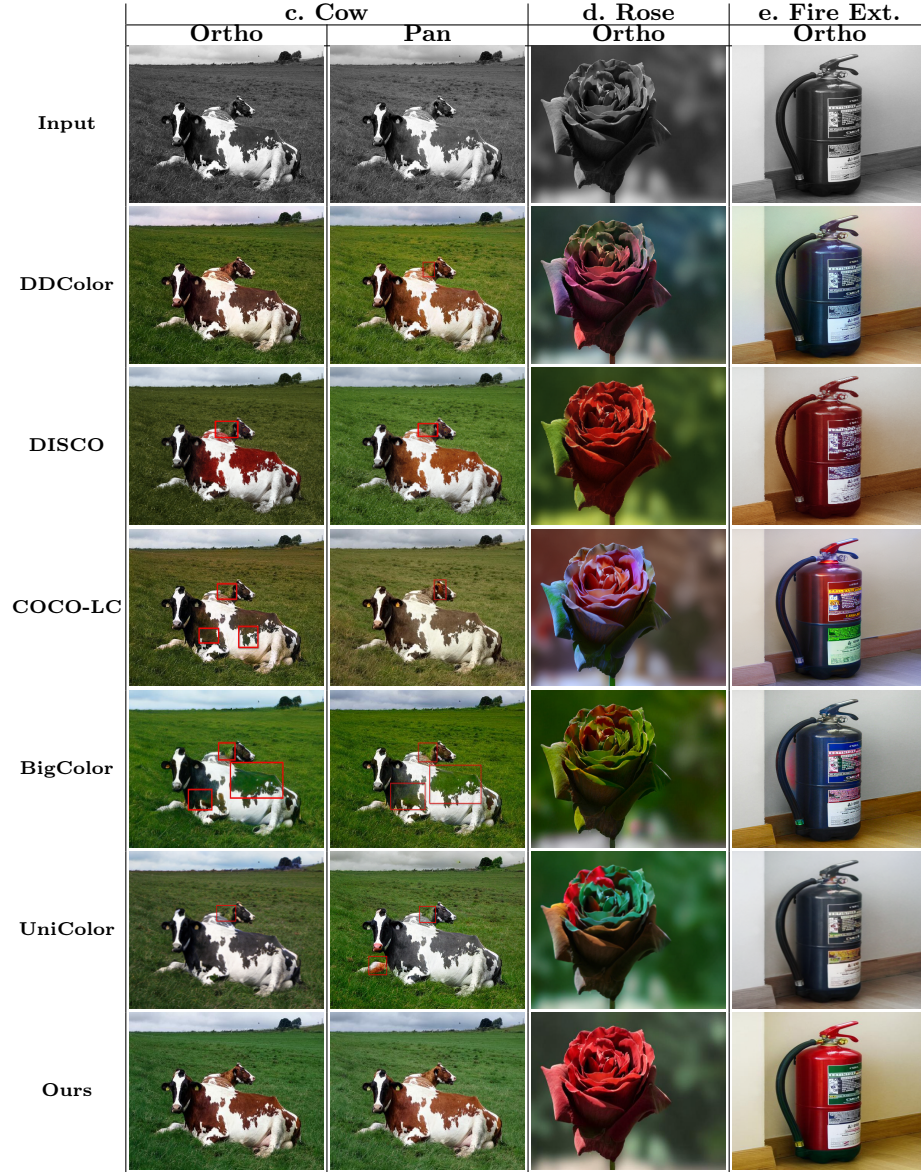


Fig. 3: Notice that for all methods except ours, the ortho result of DDColor, and the pan result of COCO-LC, the neck of the farther-away cow is confused with the grass field. In COCO-LC and BigColor, the green color from the grass bleeds into the dark spots on the cow. Also notice that the dark colors in d. Rose and e. Fire Ext., which are supposed to be bright, are not lifted by any method other than ours. This is because the brightness of a color depends on the L channel, which is fixed for the other methods. (Zoom-in for best view)



Fig. 4: Qualitative colorization of real historical orthochromatic glass-plate photographs. Each row shows the grayscale input followed by the output of each method, and each column is a distinct archival scene: **a.** mounted cavalry, **b.** a family on a porch, **c.** a studio baby portrait, and **d.** a group of young men. Unlike our simulated-orthochromatic benchmark inputs, these are genuine red-insensitive captures, so the brightness of red regions is not preserved; our method lifts and recolors these regions more naturally than the fixed-luminance baselines. Images taken from [31]. (**Zoom-in for best view**)

4.3 Results

We find that existing colorization methods degrade substantially when the input grayscale image does not preserve human-perceived luminance, as in orthochromatic imagery. In contrast, our method is substantially more robust under this shift while remaining competitive on standard grayscale inputs. This trend is reflected both quantitatively across ImageNet, COCO, and the Multi-Instance benchmark (Tables 2, 1, and 3) and qualitatively, where our outputs exhibit fewer local artifacts and more consistent color assignments (Figures 3 and 2). We also evaluate on real historical orthochromatic camera images taken from [31] (Figure 4). More examples are in supplementary.

Tables 1, 2, and 3 show a consistent pattern across all three benchmarks. On standard panchromatic grayscale inputs, our method remains competitive with strong recent baselines: DDColor typically achieves the best FID-family metrics, while our approach is often second-best on FID, sFID, and FID-DINO and achieves stronger perceptual color statistics such as Col-diverse and ColorNet. The advantage of our method becomes more pronounced under simulated orthochromatic inputs, where baseline performance degrades more substantially. On COCO, our method achieves the best FID-DINO and the strongest scores on all reported color metrics, while trailing DDColor only slightly on FID and sFID. On ImageNet, our method is the strongest overall in the ortho setting, achieving the best FID, sFID, FID-DINO, Colorfulness, Col-diverse, and ColorNet scores. On the Multi-Instance benchmark, our method again leads on most ortho metrics, including sFID, FID-DINO, Colorfulness, and Col-diverse, while remaining second-best on FID, Saturation, and ColorNet. Overall, these results support our central claim: the proposed luminance-agnostic formulation preserves strong performance on standard grayscale inputs and yields substantially greater robustness when grayscale formation deviates from modern panchromatic luminance.

Preserving the structure on the image during colorization: A potential concern with adapting a generative foundation model is that it may hallucinate content rather than faithfully recolor the input. To verify that our model preserves the structure of the grayscale input, we measure *structural fidelity* via grayscale reprojection: we convert each colorized RGB output back to grayscale using the corresponding grayscale formation and compare it against the input grayscale image using L1, L2, and SSIM. We compare three variants of FLUX.2-klein-4B: the non-fine-tuned model (*Not-FT*), a model fine-tuned only on standard panchromatic grayscale (*Normal-FT*), and our full model trained with the Mixed Grayscale Objective (*Mixed-FT*). Tables 4, 5, and 6 report results on COCO, ImageNet, and Multi-Instance, respectively. The non-fine-tuned model exhibits substantially weaker structure fidelity (SSIM around 0.36), since it tends to alter image content, whereas both fine-tuned variants preserve structure well (SSIM around 0.81–0.88). This confirms that our luminance-agnostic formulation maintains the grayscale structure while still enabling full-RGB colorization.

4.4 Human Evaluation

To complement the automatic metrics, we conducted a human study focused on visible color artifacts. We sampled the first five images from each of the three evaluation datasets (ImageNet, COCO, and Multi-Instance) under both panchromatic and orthochromatic grayscale settings, resulting in 30 evaluation inputs per method. Each input was colorized by all compared methods and evaluated by 25 human participants.

Participants were asked a binary question: whether the displayed colorized image contained noticeable artifacts such as color bleeding, unnatural hue shifts, or spatially inconsistent colorization. We summarize the results using the *Artifact-Free Rate*, defined as

$$\text{Artifact-Free Rate}(\%) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[r_i = \text{No Artifact}] \times 100, \quad (8)$$

where r_i denotes an individual response and N is the total number of valid responses.

As shown in Figures 5b and 5a, our method attains the highest Artifact-Free Rate among all compared approaches. This suggests that its outputs are perceived as more stable and less artifact-prone.

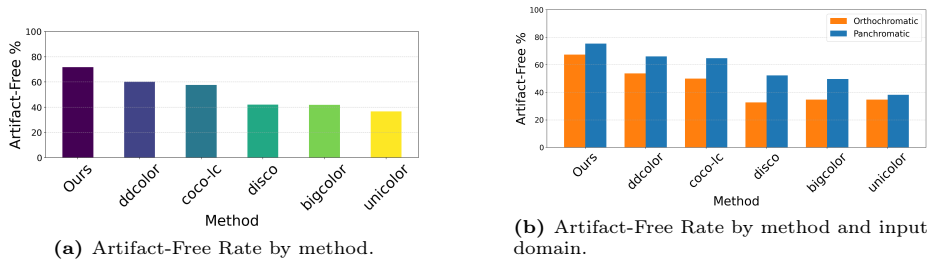


Fig. 5: Human evaluation of visible color artifacts. We report the *Artifact-Free Rate*, defined as the percentage of responses in which participants judged a colorized image to contain no noticeable artifacts. Higher values indicate that the outputs were perceived as cleaner and more visually consistent.

Complex prompt following: We compare our method against COCO-LC for prompt following. We find that our method is much better at following text prompts, even very difficult ones (Figure 6). When dealing with multiple color words in a text prompt, e.g., “the blue cup, the yellow cup, and the cyan cup”, COCO-LC incorrectly colors the middle cup red, while our method correctly colors all three cups. More examples are given in supplementary.

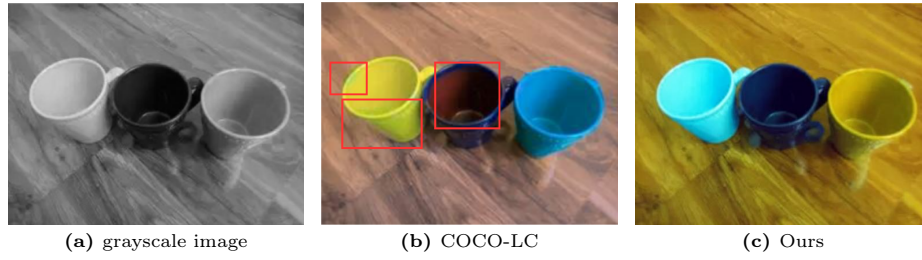


Fig. 6: a) Input grayscale image that needs to be colorized. b) The colorized output from COCO-LC. c) The colorized output from our method. When dealing with multiple color words in a text prompt, e.g., “the blue cup, the yellow cup, and the cyan cup”, COCO-LC incorrectly colors the middle cup red, while our method correctly colors all three cups. Moreover, there are unnatural green spots in the yellow cup for the COCO-LC output, compared to our output, which is clean and more realistic. (**Zoom-in for best view**)

Table 1: Quantitative comparison on the COCO benchmark under standard panchromatic (*pan*) and simulated orthochromatic (*ortho*) grayscale inputs. The best result within each input setting (pan/*ortho*) is shown in bold, and the second-best is underlined.

		FID ↓	sFID ↓	FID-DINO ↓	Colorfulness ↑	Col-diverse ↑	Saturation ↑	ColorNet ↑
BigColor	pan	9.93	10.28	55.59	42.53	15.16	17.01	4.87
	ortho	14.78	15.55	127.73	39.4	<u>14.17</u>	15.55	4.81
COCO-LC	pan	8.94	11.12	37.83	37.48	14.51	13.44	4.87
	ortho	<u>10.45</u>	14.18	53.61	36.66	14.16	13.27	<u>4.84</u>
DDColor	pan	7.06	8.26	32.34	44.00	15.50	17.32	5.06
	ortho	10.15	10.28	<u>47.25</u>	35.86	13.02	14.34	4.70
DISCO	pan	15.66	11.20	75.58	52.18	<u>16.81</u>	23.77	4.79
	ortho	17.77	12.31	91.67	<u>41.20</u>	13.80	<u>18.11</u>	4.37
UniColor	pan	11.10	10.02	55.70	42.13	14.63	17.00	4.78
	ortho	15.22	13.72	91.91	39.50	13.81	16.05	4.71
Ours	pan	8.73	9.50	37.33	<u>49.31</u>	17.25	<u>19.35</u>	5.14
	ortho	10.67	<u>10.61</u>	43.32	46.33	16.36	18.38	5.06

Table 2: Quantitative comparison on the ImageNet benchmark under standard panchromatic (*pan*) and simulated orthochromatic (*ortho*) grayscale inputs. The best result within each input setting (pan/*ortho*) is shown in bold, and the second-best is underlined.

		FID ↓	sFID ↓	FID-DINO ↓	Colorfulness ↑	Col-diverse ↑	Saturation ↑	ColorNet ↑
BigColor	pan	7.76	8.98	43.34	42.45	14.60	18.04	4.76
	ortho	11.62	13.38	101.79	39.15	13.56	16.36	<u>4.68</u>
COCO-LC	pan	9.73	10.95	51.22	35.48	13.98	12.84	4.72
	ortho	11.88	14.02	79.16	<u>42.54</u>	<u>13.91</u>	12.74	4.17
DDColor	pan	5.51	7.44	27.50	44.51	15.31	18.50	4.92
	ortho	<u>8.74</u>	<u>9.52</u>	<u>42.36</u>	36.34	12.87	15.27	4.56
DISCO	pan	11.62	9.52	68.07	51.05	<u>16.17</u>	24.25	4.65
	ortho	14.06	10.83	84.05	41.07	13.47	18.87	4.27
UniColor	pan	9.64	8.76	44.50	42.18	14.55	17.82	4.70
	ortho	13.01	12.12	78.86	39.83	13.75	17.05	4.61
Ours	pan	7.03	8.48	32.55	<u>48.25</u>	16.88	<u>19.54</u>	5.03
	ortho	7.91	9.35	38.48	45.59	16.06	<u>18.74</u>	4.98

Table 3: Quantitative comparison on the Multi-Instance benchmark under standard grayscale (*pan*) and simulated orthochromatic (*ortho*) inputs. The best result within each input setting (pan/ortho) is shown in bold, and the second-best is underlined.

		FID ↓	sFID ↓	FID-DINO ↓	Colorfulness ↑	Col-diverse ↑	Saturation ↑	ColorNet ↑
BigColor	pan	8.37	7.97	51.48	44.17	16.13	16.88	5.05
	ortho	13.11	12.36	118.92	40.08	14.71	15.32	4.91
COCO-LC	pan	6.69	10.12	47.86	39.46	15.37	13.52	5.07
	ortho	7.79	11.15	45.38	40.60	<u>15.99</u>	13.72	5.21
DDColor	pan	5.09	6.22	26.76	45.46	16.31	17.15	5.26
	ortho	8.33	<u>7.93</u>	<u>40.46</u>	36.36	13.48	14.00	4.84
DISCO	pan	13.15	8.56	63.81	52.94	17.55	23.01	4.87
	ortho	15.07	9.49	77.98	<u>41.39</u>	14.28	17.42	4.42
UniColor	pan	8.86	7.55	47.17	43.57	15.33	16.93	4.92
	ortho	12.46	10.72	77.46	40.49	14.43	15.87	4.81
Ours	pan	7.14	7.02	30.94	<u>48.64</u>	<u>17.45</u>	<u>18.26</u>	<u>5.21</u>
	ortho	<u>8.02</u>	7.88	35.71	45.22	16.37	<u>17.16</u>	<u>5.11</u>

Table 4: Structural fidelity on **COCO** via grayscale reprojection: each colored output is converted back to grayscale and compared to the input grayscale image. L1 and L2: lower is better; SSIM: higher is better. Best in bold.

Method	Ortho			Pan		
	L1 ↓	L2 ↓	SSIM ↑	L1 ↓	L2 ↓	SSIM ↑
Not-FT	0.136	0.301	0.360	0.139	0.308	0.354
Normal-FT	0.050	0.164	0.828	0.028	0.181	0.838
Mixed-FT (Ours)	0.063	0.170	0.811	0.026	0.179	0.841

5 Conclusion

We revisited image colorization through the lens of luminance mismatch. Most existing systems predict chroma while preserving an input-derived luminance channel, which works well when grayscale intensities align but can break under domain shifts such as historical orthochromatic imagery. To address this, we proposed a luminance-agnostic colorization framework that predicts full RGB outputs directly using a foundation image-editing model, rather than relying on luminance swapping.

Central to our approach is a mixed grayscale objective that trains the model under both standard panchromatic and simulated orthochromatic conditioning, allowing it to reinterpret brightness when the grayscale formation deviates from human-perceived luminance. Across COCO, ImageNet, and a multi-instance benchmark, we find that while our method remains competitive on standard grayscale inputs, it is substantially more robust on orthochromatic inputs, where prior methods degrade more strongly. Qualitative comparisons further indicate cleaner outputs with fewer local color artifacts.

Limitations: We adapt the foundation model using LoRA only; full fine-tuning with larger datasets, longer schedules, and more compute may further improve fidelity and robustness. Moreover, our orthochromatic simulation is a simplified approximation of historical film response, while real archival imagery

Table 5: Structural fidelity on **ImageNet** via grayscale reprojection (see Table 4 for the metric definition). Best in bold.

Method	Ortho			Pan		
	L1 ↓	L2 ↓	SSIM ↑	L1 ↓	L2 ↓	SSIM ↑
Not-FT	0.126	0.270	0.373	0.128	0.271	0.370
Normal-FT	0.050	0.133	0.857	0.026	0.136	0.875
Mixed-FT (Ours)	0.061	0.138	0.842	0.024	0.135	0.878

Table 6: Structural fidelity on **Multi-Instance** via grayscale reprojection (see Table 4 for the metric definition). Best in bold.

Method	Ortho			Pan		
	L1 ↓	L2 ↓	SSIM ↑	L1 ↓	L2 ↓	SSIM ↑
Not-FT	0.154	0.330	0.342	0.142	0.308	0.345
Normal-FT	0.050	0.157	0.839	0.027	0.165	0.855
Mixed-FT (Ours)	0.060	0.161	0.826	0.025	0.163	0.857

exhibits diverse capture and aging effects (e.g., blue-biased spectral sensitivity, color fading, and chemical degradation) as well as material-dependent appearance changes. For instance, facial skin can appear harsher under red-insensitive imaging, plausibly due to reduced contributions from red-wavelength subsurface light transport. Future work includes incorporating more physically grounded spectral sensitivity models, training with a broader family of grayscale/color distortions, and extending the framework to jointly address additional restoration factors beyond grayscale formation.

6 Acknowledgment

The authors sincerely thank the conference reviewers for their valuable feedback. We also thank Rajat Maheshwari, Panshul Jindal, and Krishn Vishwas Kher for their valuable support. The first author, Swarnim Maheshwari, gratefully acknowledges support from the Reliance Foundation Fellowship.

References

1. Antic, J.: Deoldify: A deep learning based project for colorizing and restoring old images (and video!) (2019), <https://github.com/jantic/DeOldify>, accessed: 2026-02-15
2. Bozic, V., Djelouah, A., Zhang, Y., Timofte, R., Gross, M., Schroers, C.: Versatile vision foundation model for image and video colorization. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
4. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)

5. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-cad: Language-based colorization with any-level descriptions using diffusion priors. In: NeurIPS (2023)
6. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-coins: Language-based colorization with instance awareness. In: CVPR (2023)
7. Chen, C., Mo, J.: IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: <https://github.com/chaofengc/IQA-PyTorch> (2022), accessed: 2026-02-15
8. Commission Internationale de l’Éclairage (CIE): Colorimetry—part 4: Cie 1976 l*a*b* colour space (2007), <https://cie.co.at/publications/colorimetry-part-4-cie-1976-lab-colour-space-1>, standard CIE S 014-4/E:2007. Accessed: 2026-02-15
9. Ding, X., Wang, Y., Xu, Z., Welch, W.J., Wang, Z.J.: Continuous conditional generative adversarial networks for image generation: Novel losses and label input mechanisms. arXiv preprint arXiv:2011.07466 (2020)
10. Du, X., Zhou, Z., Wang, Y., Wang, Z., Zheng, Y., Jin, C.: Multicolor: Image colorization by learning from multiple color spaces. arXiv preprint arXiv:2408.04172 (2024)
11. Encyclopaedia Britannica: Orthochromatic film (nd), <https://www.britannica.com/technology/orthochromatic-film>, describes orthochromatic films as sensitive to violet/blue/green/yellow but not to red. Accessed: 2026-02-15.
12. Film Colors: Orthochromatic stock (nd), <https://filmcolors.org/timeline-entry/1346/>, historical note: sensitivity extended to record green as well as blue (Vogel, 1873). Accessed: 2026-02-15.
13. Hasler, D., Suesstrunk, S.E.: Measuring colorfulness in natural images. In: Human vision and electronic imaging VIII. vol. 5007, pp. 87–95. SPIE (2003)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
15. Huang, Z., Zhao, N., Liao, J.: Unicolor: A unified framework for multi-modal colorization with transformer. ACM TOG (2022)
16. Ji, X., Jiang, B., Luo, D., Tao, G., Chu, W., Xie, Z., Wang, C., Tai, Y.: Color-former: Image colorization via color memory assisted hybrid-attention transformer. In: ECCV (2022)
17. Jin, X., Li, Z., Liu, K., Zou, D., Li, X., Zhu, X., Zhou, Z., Sun, Q., Liu, Q.: Focusing on persons: Colorizing old images learning from modern historical movies. In: ACM MM (2021)
18. Kaggle: Imagenet object localization challenge (2020), <https://www.kaggle.com/competitions/imagenet-object-localization-challenge/data>, accessed: 2026-02-15
19. Kang, X., Yang, T., Ouyang, W., Ren, P., Li, L., Xie, X.: Ddcolor: Towards photo-realistic image colorization via dual decoders. In: ICCV (2023)
20. Kim, G., Kang, K., Kim, S., Lee, H., Kim, S., Kim, J., Baek, S.H., Cho, S.: Bigcolor: Colorization using a generative color prior for natural images. In: ECCV (2022)
21. Kodak: Basic sensitometry and characteristics of film (film essentials, module 6) (nd), https://www.kodak.com/uploadedfiles/motion/US_plugins_acrobat_en_motion_newsletters_filmEss_06_Characteristics_of_Film.pdf, states orthochromatic films are sensitive mainly to the blue-green portions of the visible spectrum. Accessed: 2026-02-15.
22. Kodak: Glossary of motion picture terms (nd), <https://www.kodak.com/en/motion/page/glossary-of-motion-picture-terms/>, entry: “Orthochromatic (Ortho) Film” (sensitive to only blue and green light). Accessed: 2026-02-15.

23. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer. In: ICLR (2021)
24. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. <https://bfl.ai/blog/flux-2> (2025), accessed: 2026-02-15
25. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), <https://arxiv.org/abs/2506.15742>
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
27. Li, Y., Bai, Y., Yang, S., Liu, J.: Coco-ic: Colorfulness controllable language-based colorization. In: ACM MM (2024)
28. Liang, Z., Li, Z., Zhou, S., Li, C., Loy, C.C.: Control color: Multimodal diffusion-based interactive image colorization. arXiv:2402.10855 (2024)
29. Liu, H., Xing, J., Xie, M., Li, C., Wong, T.T.: Improved diffusion-based image colorization via piggybacked models. arXiv preprint arXiv:2304.11105 (2023)
30. Luo, X., Zhang, X., Yoo, P., Martin-Brualla, R., Lawrence, J., Seitz, S.M.: Time-travel rephotography. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2021) **40**(6) (12 2021). <https://doi.org/https://doi.org/10.1145/3478513.3480485>
31. Minnesota Digital Library: Glass plate negatives collection. Minnesota Digital Library (2024), https://collection.mndigital.org/?f%5Bphysical_format_ssi%5D%5B%5D=Glass+plate+negatives, accessed: 2026-02-15
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
34. Su, J.W., Chu, H.K., Huang, J.B.: Instance-aware image colorization. In: CVPR (2020)
35. Vitoria, P., Raad, L., Ballester, C.: Chromagan: Adversarial picture colorization with semantic class distribution. In: WACV (March 2020)
36. Weng, S., Sun, J., Li, Y., Li, S., Shi, B.: Ct2: Colorization transformer via color tokens. In: ECCV (2022)
37. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), <https://arxiv.org/abs/2508.02324>
38. Wu, Y., Wang, X., Li, Y., Zhang, H., Zhao, X., Shan, Y.: Towards vivid and diverse image colorization with generative color prior. In: CVPR (2021)
39. Xia, M., Hu, W., Wong, T.T., Wang, J.: Disentangled image colorization via global anchors. ACM TOG (2022)

- 40. Zabari, N., Azulay, A., Gorkor, A., Halperin, T., Fried, O.: Diffusing colors: Image colorization with text guided diffusion. In: SIGGRAPH Asia 2023 Conference Papers (2023)
- 41. Zerman, E., Rana, A., Smolic, A.: Colornet-estimating colorfulness in natural images. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 3791–3795. IEEE (2019)
- 42. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: CVPR (2023)
- 43. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)