This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ColorFLUX: A Structure-Color Decoupling Framework for Old Photo Colorization

Bingchen Li[1*§] Zhixin Wang[2*†] Fan Li[2‡] Jiaqi Xu[2] Jiaming Guo[2] Renjing Pei[2] Xin Li[1] Zhibo Chen[1†]
[1] University of Science and Technology of China [2] Huawei Noah's Ark Lab
lbc31415926@mail.ustc.edu.cn, {xin.li, chenzhibo}@ustc.edu.cn
{wangzhixin6, lifan61}@huawei.com

Figure 1. ColorFLUX achieves vivid and realistic colorization of old photos by decoupling structure and color restoration, exhibiting strong generalization across diverse scenarios, including portraits, group photos, and landscapes.

Abstract

Old photos preserve invaluable historical memories, making their restoration and colorization highly desirable. While existing restoration models can address some degradation issues like denoising and scratch removal, they often struggle with accurate colorization.This limitation arises from the unique degradation inherent in old photos, such as faded brightness and altered color hues, which are different from modern photo distributions, creating a substantial domain gap during colorization. In this paper, we propose a novel old photo colorization framework based on the generative diffusion model FLUX. Our approach introduces a structure-color decoupling strategy that separates structure preservation from color restoration, enabling accurate colorization of old photos while maintaining structural consistency. We further enhance the model with a progressive Direct Preference Optimization (Pro-DPO) strategy, which allows the model to learn subtle color preferences through coarse-to-fine transitions in color augmentation. Additionally, we address the limitations of text-based prompts by introducing visual semantic prompts, which extract fine-grained semantic information directly from old photos, helping to eliminate the color bias inherent in old photos. Experimental results on both synthetic and real datasets demonstrate that our approach outperforms existing state-of-the-art colorization methods, including closed-source commercial models, producing high-quality and vivid colorization.

1. Introduction

Old photos serve as invaluable visual records of history, capturing people, places, and moments that define the past. However, due to the limitations of early photographic equipment and preservation conditions, most of these images lack vivid colors (*i.e.*, are black-and-white) and often suffer from natural fading (We present some examples in Fig. 2). Recent advances in image restoration models [18, 23, 30, 45, 46] have demonstrated strong capabilities in removing common degradations in old photos, such as scratches, noise, and blur. However, these models are not specifically designed for colorization. A practical workaround is to first restore degradations and then ap-

[*]Equal contribution. [†]Corresponding authors. [‡]Project lead.
[§]Work done during an internship at Huawei Noah's Ark Lab.

Figure 2. llustration of real old photo's fading effect. ColorFLUX brings them back to life (second row).

ply a designated colorization model [41]. But this pipeline faces challenges, as faded grayscale images provide limited color information, requiring the model to infer plausible colors. This involves ensuring semantic consistency, preventing color bleeding, and aligning with modern color preferences. Additionally, the degradation in old photos, like faded brightness and altered color hues, differs from modern photos, creating a difficult training gap.

Existing generative colorization methods [22] typically leverage text-to-image (T2I) models as color priors while maintaining structural consistency between the grayscale inputs and colorized output, but still struggle with issues like semantic consistency and preventing color bleeding. Moreover, these methods heavily rely on text prompts, which often fail to capture the fine details in image - a limitation encapsulated by the adage, "A picture is worth a thousand words."

To address these issues, we propose a framework for old photo colorization based on the generative diffusion model FLUX [19], chosen for its strong ability to generate high-quality colorized images, even from limited or faded inputs, making it ideal for restoring old photos. The proposed framework introduces an effective structure-color decoupling learning strategy to ensure accurate semantic understanding for color restoration of grayscale old photos. Specifically, the framework consists of three training stages: structure consistency training, basic color learning, and fine-color adjustment. In the structure consistency training stage, FLUX is fine-tuned using ControlNet [50], which accepts grayscale images and color visual prompts generated from the color ground truth using Redux[1]. The training loss used is the flow matching loss, enforcing the model to learn the structure rather than the color information. During the basic color learning stage, we fine-tune the pretrained Redux model, allowing it to extract natural semantic embeddings that connect the priors of FLUX, while abandoning the color bias from old photo inputs. This stage combines flow matching loss and feature distillation loss, while freezing both the structure-preserving ControlNet and the pretrained FLUX, to decouple it from the structural training.

After restoring the basic colors, the next step is to correct fading effects such as low saturation or over-/underexposure, which are commonly observed in old photos. We propose fine-color adjustment, a post-training process implemented via Direct Preference Optimization (DPO), enhancing the model's ability to generate vivid colors that align with modern aesthetics and human preferences. Specifically, the preference dataset for DPO training consists of triplets: input, positive, and negative samples. Positive samples are high-quality, realistic images, while negative samples are their corresponding color-augmented versions, and the input is the grayscale version of the color-augmented images. However, learning directly from a large color-augmentation space can overwhelm the model, causing it to focus on obvious color changes rather than more subtle adjustments. To address this, we propose a progressive DPO (Pro-DPO) strategy that begins with larger, more noticeable augmentations and gradually transitions to finer ones, helping the model learn more subtle color preferences.

Experimental results for both real and synthetic datasets demonstrate that our approach outperforms both existing state-of-the-art colorization models and closed-source commercial solutions. We summarize our contributions as:

- We present a framework for old photo colorization based on FLUX and incorporates an effective structure-color decoupling learning strategy.
- We propose a progressive DPO strategy that post-trains the model through coarse-to-fine color augmentation transitions, enabling the model to learn subtle color preferences and generate vivid colors for old photos.
- We conduct extensive experiments to evaluate our colorization framework, demonstrating that it achieves state-of-the-art performance, outperforming both existing methods and closed-source commercial solutions.

2. Related Works

2.1. Old Photo Restoration

The pioneer work [34] performs transfer learning in the latent space to align the distribution of old photos with that of clean images, effectively removing degradations present in old photos. However, it falls short in colorization, as it lacks explicit mechanisms to model vivid colors from the limited information available in grayscale inputs. DeOldify [1] leverages GANs to learn the colorization of old photos, which utilizes self-attention in the U-Net generator to extract color cues from grayscale images. Pik-Fix [41] adopts a divide-and-conquer strategy, restoring old photos through restoration, similarity, and colorization sub-networks, with a reference image providing chromatic guidance.

[1] https://huggingface.co/black-forest-labs/FLUX.1-Redux-dev

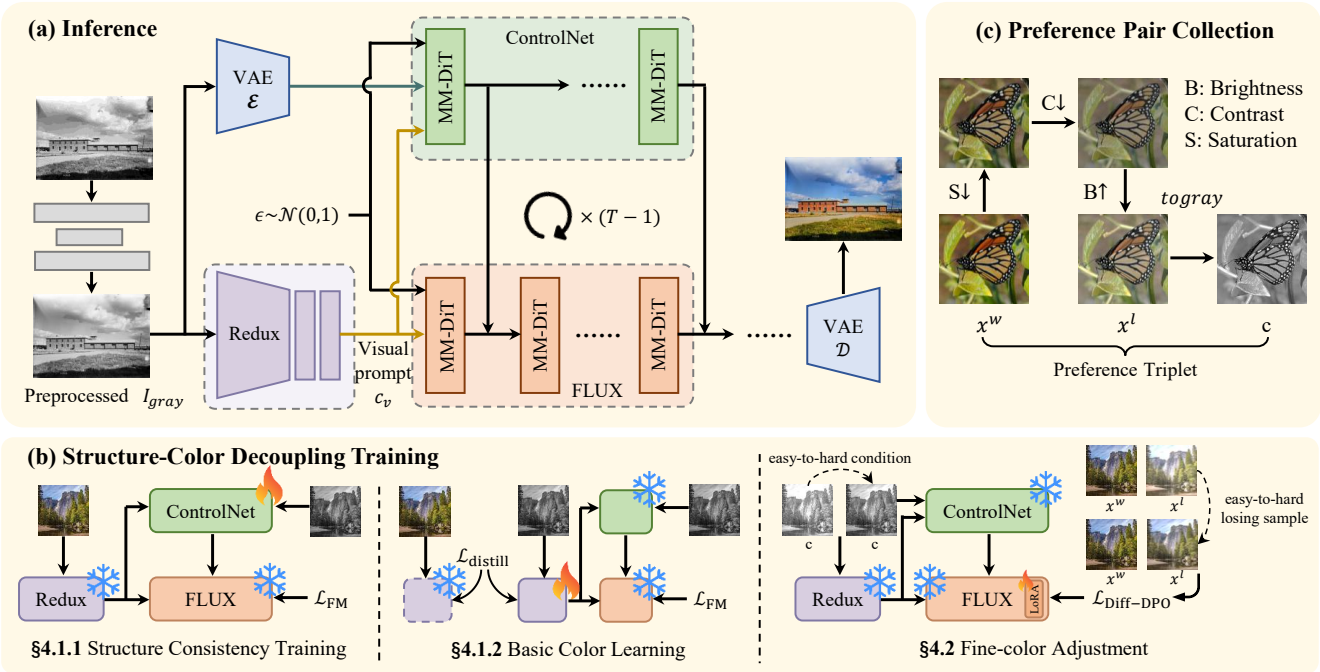

Figure 3. The framework of our proposed method. (a) The inference process of colorization, including a preprocessor to remove degradations in old photos. (b) The training procedure of our method. Notably, we illustrate an example of easy-to-hard losing samples. The model progressively learns preferences from severely augmented samples to milder ones, ensuring precise color perception to effectively counteract the fading effects of old photos. (c) The preference pair collection pipeline. To synthesize x^l, we randomly apply augmentation combinations sampled from ['B', 'C', 'S'].

2.2. Image Colorization

Image colorization typically aims to add color to a grayscale image, restoring the color information that has either been lost or was originally absent. Early works [10, 11, 15] primarily focused on automatic colorization, where networks learn pixel-wise colorization mappings from synthetic color–grayscale image pairs. However, due to the limited chromatic cues available in grayscale images, these automatic methods often struggle to generate rich and realistic colors. Furthermore, they face challenges in accurately understanding the semantics of image instances, resulting in artifacts such as color bleeding [22]. To mitigate these issues, a series of works [22, 24, 25, 35, 38] propose to leverage generative priors to better understand image semantics, while employing various forms of guidance to direct colorization process, such as color hints [9, 47], exemplar images [3, 4, 6], or language descriptions [7, 8, 36–38], *etc*.

However, directly applying these colorization methods to old photos faces a major challenge: the fading effects commonly present in old photos can lead to colorization results that do not align with human preferences, particularly in color attributes such as brightness, contrast, and saturation, which are highly perceptible to the human eye. In this paper, we address this issue by introducing a color-aware progressive Direct Preference Optimization (DPO) learning strategy within the colorization framework.

2.3. Direct Preference Optimization for Diffusion

Direct Preference optimization (DPO) [31] can be applied to pre-trained diffusion models to align them with human preferences. Diffusion-DPO [33] firstly introduces DPO into T2I generation and fine-tunes diffusion models with preference data collected from human-annotated Pick-a-Pic [17] dataset, achieving better prompt alignment and aesthetic scores. Diffusion-RPO [13] improves the construction of the preference datasets by introducing relative samples with similar semantics. MaPO [14] handles reference mismatch by replacing divergence regularization on the reference model with an amplification factor. D3PO [42] further adapts DPO to discrete diffusion models for structured sequence generation tasks, ensuring preference alignment without explicit reward models.

3. Preliminaries

3.1. Rectified Flow-based Diffusion

Modern generative modeling [12, 19] are usually based on Rectified Flow [28], which defines the intermediate noisy data x_t by:

$$x_t = (1-t)x_0 + t\epsilon, \tag{1}$$

where $t \in [0,1]$ is the sampled timestep, $x_0 \sim p(x_0)$ is drawn from real data distribution $p(x_0)$, and $\epsilon \sim \mathcal{N}(0, I)$ is standard Gaussian noise. The predicted velocity field

$v_\theta(x_t, t)$ by the diffusion transformer can be optimized by the flow matching objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\, x_0 \sim p(x_0),\, \epsilon \sim \mathcal{N}(0,I)} \left[\left\| v - v_\theta(x_t, t) \right\|^2 \right], \quad (2)$$

where the target velocity field is $v = \epsilon - x_0$.

3.2. Diffusion Direct Preference Optimization

Preference optimization aims to train models using triplets $(c, x^w, x^l) \sim D$, where the model is encouraged to generate preferred outputs x^w over dispreferred ones x^l under a given condition c (*e.g.*, text prompt). Rafailov *et al*. [31] propose Direct Preference Optimization (DPO), which bypasses explicit reward modeling and instead directly optimizes the model parameters based on preference pairs. To adapt DPO to diffusion models, Wallace *et al*. [33] replace policy log-likelihoods with diffusion ELBO scores and derive $\mathcal{L}_{\text{Diff-DPO}}$ as:

$$-\mathbb{E}\Bigg[\log\sigma\Bigg(-\frac{\beta}{2}\Big(\|\epsilon^w - \epsilon_\theta(x_t^w, t)\|^2 - \|\epsilon^w - \epsilon_{\text{ref}}(x_t^w, t)\|^2 - \Big(\left\|\epsilon^l - \epsilon_\theta(x_t^l, t)\right\|^2 - \left\|\epsilon^l - \epsilon_{\text{ref}}(x_t^l, t)\right\|^2\Big)\Big)\Bigg)\Bigg], \quad (3)$$

where the expectation is taken over preference triplets and the noise schedule t. Consequently, to extend $\mathcal{L}_{\text{Diff-DPO}}$ to advanced Rectified Flow-based diffusion models, Liu *et al*. [26, 27] replace the noise vector ϵ^* with the velocity field v^* in Eq. 2 and form the final loss function (denote as $\mathcal{L}_{\text{Diff-DPO}}$ in the following paper):

$$-\mathbb{E}\Bigg[\log\sigma\Bigg(-\frac{\beta_t}{2}\Big(\|v^w - v_\theta(x_t^w, t)\|^2 - \|v^w - v_{\text{ref}}(x_t^w, t)\|^2 - \Big(\left\|v^l - v_\theta(x_t^l, t)\right\|^2 - \left\|v^l - v_{\text{ref}}(x_t^l, t)\right\|^2\Big)\Big)\Bigg)\Bigg], \quad (4)$$

where $\beta_t = \beta(1-t)^2$. Moreover, Liu *et al*. [26] demonstrate that using a constant $\beta_t = \beta_c$ achieves better results.

4. Method

In this paper, we propose a novel colorization framework for old photos based on FLUX[19], which consists of three training stages: structure consistency training, basic color learning, and fine-color adjustment. Moreover, as shown in Figure 3, we utilize a pretrained low-level image preprocessor to remove degradations such as noise, and blur, thus eliminating potential interference with the subsequent colorization process. For simplicity, we denote the preprocessed image as I_{gray} in the remainder of this paper. Additionally, to comprehensively evaluate our method, we collect 50 real old photos from the Internet and develop an evaluation metric based on a multimodal large language model, as detailed in Sec. 4.3.

4.1. Decoupling Structure-Color Learning

4.1.1. Structure Consistency Training

Following the common practice [22], we utilize ControlNet [50] to inject structural information into FLUX. Unlike previous methods, our ControlNet takes two inputs simultaneously: the grayscale image and a color visual prompt, as shown in Fig. 3 (b). The latter input is obtained by extracting color embeddings from I_{gt} using a pretrained Redux model, providing the model with accurate color priors. This training approach encourages the model to primarily focus on structural consistency rather than color restoration at this stage. Furthermore, we use the flow matching loss defined in Eq. 2 for training, while keeping the pretrained FLUX and Redux models frozen during this stage.

4.1.2. Basic Color Learning

In this stage, we fully leverage the color prior knowledge embedded in the pretrained FLUX model to infer plausible basic colors, aligning them with modern human preferences while preventing color bleeding. Existing approaches relying on textual prompts may fall short in capturing fine-grained details, making them inadequate for preventing color bleeding in old photo colorization tasks. To address this, as illustrated in Fig. 3(b), we fine-tune the pretrained Redux model to extract natural semantic embeddings closely aligned with the color priors of FLUX, while simultaneously reducing the influence of color biases inherent in old photos. Specifically, we fine-tune the Redux model Φ through color-knowledge distillation from the I_{gt} to the grayscale input image I_{gray} as follows:

$$\mathcal{L}_{\text{distill}} = \|\Phi(I_{gray}) - \Phi'(I_{gt})\|^2, \quad (5)$$

where Φ' denotes a copy of pretrained Redux model that is frozen and specifically used to extract ground-truth color information. This distillation loss drives the embeddings extracted from grayscale images to be close to the image embedding extracted from their corresponding color images. To ensure that the generated visual semantic prompt is well aligned with the color priors space of the pre-trained FLUX model, we combine both $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_{\text{distill}}$ to fine-tune Φ.

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \alpha\mathcal{L}_{\text{distill}} \quad (6)$$

where α is a weighting hyperparameter.

4.2. Progressive DPO for Fine-color Adjustment

While basic color restoration provides a foundation, it is insufficient to fully revive old photos. The next step is to subtly correct common fading effects in old photos, such as low saturation, over-/under-exposure. Direct Preference Optimization (DPO) is a widely used post-training method that leverages and enhances a model's inherent ability to generate vivid colors aligned with modern aesthetics and human

Table 1. Quantitative comparisons across different methods on two benchmarks. VQ-R1, CRI, CRA, CCS, SCS, AES, and OA denote VisualQuality-R1, Color RIchness, Color RAtionality, Color ConSistency, Structure ConSistency, AESthetics, and OverAll performance, respectively. Notably, we apply our preprocessor to all methods to ensure a fair comparisons on colorization performance. The Runtime is measured on a single GPU with about 19.5 TFLOPS. Best results are **bolded.**

	NR-IQA			Qwen-score						Runtime
	DeQA	Q-Insight	VQ-R1	CRI	CRA	CCS	SCS	AES	OA	
DIV2K-valid-synthesized										
DeOldify [1]	4.080	3.978	4.625	69.35	**80.95**	83.80	**94.78**	80.70	81.73	2.23s
DDColor [16]	4.162	3.992	4.574	79.80	78.55	77.60	93.22	81.13	81.80	0.19s
CtrlColor [22]	4.144	**4.013**	4.654	75.40	80.25	**84.04**	94.53	81.85	83.28	8.05s
FLUX-Kontext [20]	4.209	4.002	4.661	77.75	80.00	83.61	93.91	82.22	83.48	28.14s
ColorFLUX (Ours)	**4.293**	4.008	**4.739**	**80.80**	79.70	82.23	94.05	**83.12**	**83.73**	7.45s
DIV2K-valid-augmented										
DeOldify [1]	3.841	3.798	4.349	64.70	77.75	81.37	94.29	77.63	79.11	2.23s
DDColor [16]	3.941	3.877	4.410	77.85	76.70	78.66	93.43	80.43	81.54	0.19s
CtrlColor [22]	3.917	3.857	4.432	72.35	78.40	82.99	94.03	80.25	81.71	8.05s
FLUX-Kontext [20]	4.175	**3.981**	4.617	77.25	**79.90**	**84.57**	94.63	81.83	82.55	28.14s
ColorFLUX (Ours)	**4.303**	3.978	**4.728**	**80.15**	79.70	82.66	**94.83**	**82.80**	**83.53**	7.45s
RealOldPhotos										
DeOldify [1]	4.056	3.852	4.419	70.90	80.50	82.66	94.76	79.02	80.86	2.23s
DDColor [16]	4.090	**3.935**	4.500	76.20	81.10	82.36	94.60	82.42	82.70	0.19s
CtrlColor [22]	4.050	3.931	4.486	76.40	80.50	83.24	94.50	81.64	82.68	8.05s
FLUX-Kontext [20]	4.110	3.843	4.452	66.00	76.90	79.74	**94.88**	76.38	77.70	28.14s
Doubao [5]	3.741	3.563	4.097	**83.10**	64.90	75.10	89.00	74.40	75.24	∼29s[2]
ColorFLUX (Ours)	**4.199**	3.869	**4.593**	80.50	**81.30**	**83.36**	94.46	**83.22**	**83.20**	7.45s

preferences. In this paper, we construct a color-preference dataset and introduce a progressive DPO training strategy for fine-color adjustment.

4.2.1. Color-Preference Pairs Collection

The color-preference dataset for DPO training consists of triplets: input, positive, and negative samples. Intuitively, realistic and naturally colored images are defined as the winning (positive) samples denoted as x^w while losing (negative) samples, denoted as x^l are synthesized from x^w by simulating various fading artifacts. Formally, given a colored ground-truth image $x^w = I_{gt}$, we generate the losing sample $x^l = I_{aug}$ with randomly shuffled combinations of *brightness adjustment*, *contrast reduction*, and *saturation reduction*. The condition c is obtained by applying the `RGB2GRAY` algorithm to I_{aug}. The overall pipeline is illustrated in Fig. 3 (c).

4.2.2. Progressive DPO Training

Learning directly from a wide range of color augmentations may cause the model to overfit to simple preference patterns (e.g., avoiding excessively saturated results) while neglecting more subtle but valuable fine-color adjustments, such as slightly increasing saturation to enhance vividness. To mitigate this, we propose a progressive DPO (Pro-DPO) strategy that gradually shifts the training focus from coarse to fine adjustments, enabling the model to capture subtle color preferences over time. Specifically, we design a two-stage progressive DPO training scheme. In the first stage, we apply strong augmentations to construct obvious preference gaps, allowing the model to learn coarse distinctions. In the second stage, we gradually reduce the augmentation strength, introducing more subtle distinctions and encouraging the model to refine its ability to perform fine-grained color adjustments.

4.3. Evaluation of Colorization Framework

Assessing the quality of colorized images is particularly challenging, as color is inherently more subjective and ambiguous compared to image content. Although numerous Image Quality Assessment (IQA) metrics exist [39, 43], there is still no reliable metric specifically tailored to evaluate image colorization. Inspired by evaluation protocols in multimodal large language models (*e.g.*, GPT-Score) [32], we employ the open-source Qwen2.5-VL-72B model [2] to quantitatively assess the perceptual quality of colorized results. The detailed prompts are provided in the Supplementary. Moreover, prior works [34, 41] primarily focus on restoring portrait-style old photos with limited samples (*e.g.*, only 13 old photos in DDColor [16]), while largely overlooking more complex scenarios such as crowds or nat-

[2]Runtime is measured through its official website interface.

Figure 4. Qualitative comparisons between ColorFLUX and other methods on *RealOldPhotos*. Zoom in for better views.

ural landscapes. To enable a more comprehensive evaluation of colorization performance, we additionally collected a set of 50 publicly available old photos from the Internet (*i.e.*, *RealOldPhotos*), covering diverse contents.

5. Experiments

5.1. Implementation Details

Network Architecture. We build our colorization framework based on FLUX.1 [19]. We adopt the official FLUX Redux model [19] as the initialization of our semantic visual prompt extractor, which consists of a SigLip [48] vision encoder and a two-layer MLP to align the features to the text branch of FLUX. For DPO training, we apply LoRA with a rank of 32 to the attention and feed-forward layers within all MM-DiT blocks of FLUX. For preprocessor, we adopt the pre-trained S3Diff [49].

Training Details. The training is conducted on 8×80 GB GPUs and completes in around two days. We provide detailed training descriptions in the supplementary.

Inference. During inference, we resize the input to 1024×1024 (same as training), use the Euler flow-matching noise scheduler, and fix the random seed, guidance scale, and control scale to 0, 3.5, and 1, respectively, with a total of 8 sampling steps. After denoising, we resize the image back to the original resolution.

Metrics. As mentioned in Sec. 4.3, we primarily use Qwen2.5-VL-72B[3] (abbreviated as Qwen-score) to evaluate the performance of colorization. We focus on six aspects: Color RIchness (CRI), Color RAtionality (CRA), Color ConSistency (CCS), Structural ConSistency (SCS), AESthetics (AES), and OverAll (OA). As a supplement to Qwen-score, we further adopt three state-of-the-art NR-IQA metrics, including DeQA [44], Q-Insight [21], and VisualQuality-R1 [40].

[3]https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct

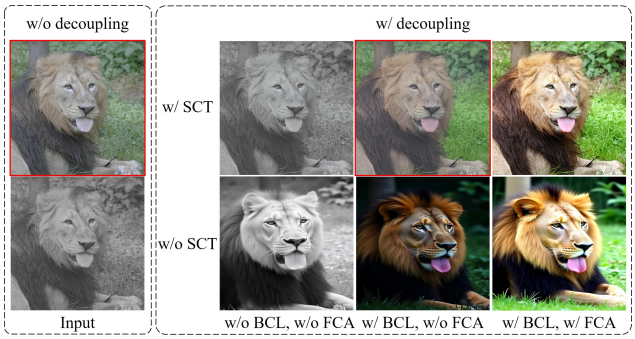

Figure 5. Visual results of decoupling each parts of ColorFLUX. We denote structure consistency training, basic color learning, and fine-color adjustment by SCT, BCL, and FCA, respectively. "W/o" represents that we remove the according network from the trained framework (*e.g.*, w/o SCT denotes for removing ControlNet).

Evaluation Datasets. We evaluate the performance of our method on both synthesized and real datasets. For the former, we adopt two settings: (i) we convert the DIV2K validation set (100 images) to grayscale using the `RGB2GRAY` algorithm (*DIV2K-valid-synthesized*); and (ii) we first apply color augmentations—such as brightness adjustment, saturation reduction, and contrast reduction—to the original images, and then convert them to grayscale (*DIV2K-valid-augmented*). The first setting follows the evaluation protocol adopted by previous image colorization works [16, 29] for fair comparison, whereas the second setting is introduced to mimic the color fading and degradation artifacts typically present in old photos. For the latter, we use the

Table 2. Quantitative comparisons across different semantic prompt strategies. "VS Prompt" refers to visual semantic prompt. Best results are **bolded.**

	DeQA	Q-Insight	VQ-R1	CRI	CRA	CCS	AES
DIV2K-valid-augmented							
Text Prompt	3.758	3.689	4.289	50.00	68.70	78.82	69.74
VS Prompt	**3.932**	**3.752**	**4.344**	**55.80**	**71.85**	**79.44**	**72.24**
RealOldPhotos							
Text Prompt	3.863	3.761	4.248	58.00	71.70	75.00	70.06
VS Prompt	**4.105**	**3.811**	**4.381**	**67.80**	**79.60**	**82.16**	**78.62**

old photos we collected from the Internet (*RealOldPhotos*).

5.2. Colorization Results

For synthesized datasets, we compare our method with three representative colorization approaches: an open-source tool, DeOldify [1]; an encoder–decoder-based method, DDColor [16]; a diffusion-based method, CtrlColor [22]; and a FLUX-based editing model, FLUX-Kontext [20]. For the real dataset, we additionally include a closed-source image editing model, Doubao, to assess ColorFLUX's performance in comparison with a commercial-grade model.

We demonstrate the quantitative results in Tab. 1. As observed, ColorFLUX achieves the best performance on DeQA, VisualQuality-R1 (VQ-R1), Aesthetics (AES), and the overall (OA) score across all three benchmarks. Compared with DDColor and CtrlColor, ColorFLUX not only demonstrates higher color richness but also achieves superior color rationality. This suggests that our method is capable of enriching old photos with vivid colors, while effectively leveraging the pre-trained generative prior through our visual semantic prompts to produce colors that align more naturally with real-world scenes. Notably, although Doubao achieves higher color richness than ColorFLUX, its results tend to be overly saturated, leading to unrealistic appearances as depicted in Fig. 4. This further demonstrates that our method not only enhances color richness but also preserves color rationality.

Referring to Fig. 4, ColorFLUX consistently produces visually appropriate colorization results against different old photos. In contrast, other methods exhibit large variations in brightness (*e.g.*, Kontext), or produce overly dull and monochromatic results due to the fading effect (*e.g.*, DeOldify and CtrlColor). Thanks to the accurate semantics extracted by our visual semantic prompt, ColorFLUX also maintains strong color consistency in complex scenes. In comparison, DDColor and DeOldify suffer from noticeable color bleeding artifacts (*e.g.*, hands and coats).

5.3. Ablation Studies

Effectiveness of Structure-Color Decoupling. We investigate the effectiveness of our structure-color decoupling strategy by visualizing the contributions of each network component to the final colorization results. As shown in the right part of Fig. 5, removing the SCT-optimized ControlNet leads to severe structural inconsistency. While incorporating the BCL-optimized Redux recovers basic colorization, the results still suffer from color degradations inherited from the grayscale input. With the addition of the FCA-optimized LoRA weights, ColorFLUX finally achieves visually pleasing colorization results with enhanced fidelity and vividness.

In the left part of Fig. 5, we show results from jointly training Redux and ControlNet without structure-color decoupling. Comparing the two red-boxed images, the non-decoupled result shows a noticeably duller facial region on the lion, indicating that entangling color learning and structural preservation may interfere with each other, leading to suboptimal results.

Figure 6. Qualitative comparisons for different fine-color post training strategy on *DIV2K-valid-augmented*.

Table 3. Quantitative comparisons across different basic color learning strategies. "$\mathcal{L}_{\text{FM}}$" and "$\mathcal{L}_{\text{distill}}$" denote flow matching loss and distillation loss, respectively. We report results on *DIV2K-valid-synthesized* without augmentations.

	DeQA	Q-Insight	VQ-R1	CRI	CRA	AES	OA
$\mathcal{L}_{\text{FM}}$	3.889	3.868	4.432	46.35	68.45	67.64	69.91
$\mathcal{L}_{\text{distill}}$	4.044	3.895	4.552	**58.55**	70.65	72.04	75.12
$\mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{distill}}$	**4.114**	**3.907**	**4.577**	57.55	**72.45**	**73.32**	**75.39**

Effectiveness of $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_{\text{distill}}$. In this part, we study the effectiveness of utilizing a combination of flow matching loss $\mathcal{L}_{\text{FM}}$ and distillation loss $\mathcal{L}_{\text{distill}}$ to perform the basic color learning of the Redux model. Specifically, we compare our implementation with two more settings: (i) only use $\mathcal{L}_{\text{FM}}$, and (ii) only use $\mathcal{L}_{\text{distill}}$ with out FLUX involving. As demonstrated in Tab. 3, our implementation achieves the best NR-IQA scores and almost the best Qwen-scores. Specifically, using only $\mathcal{L}_{\text{FM}}$ ensures that Redux's output aligns with the pre-trained representation space of FLUX, but may lead to low training efficiency, resulting in suboptimal distillation performance. On the other hand, although distill the Redux model alone achieves better color richness, the relatively lower color rationality suggests a potential distribution shift in Redux's output space without the supervision of $\mathcal{L}_{\text{FM}}$, which impairs FLUX's ability to accurately interpret the visual semantic prompt. This is also validated through visualization in Fig. 7, where our implementation achieves the most vibrant colorization results.

$\mathcal{L}_{\text{FM}}$ $\mathcal{L}_{\text{distill}}$ $\mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{distill}}$

Figure 7. Visual comparisons between different basic color learning strategies on *DIV2K-valid-augmented*.

Table 4. Quantitative comparisons across different fine-color adjusting strategies. "W/o DPO" indicates that no color adjustment training, and "One-stage DPO" refers to DPO training without the progressive strategy. Best results are **bolded.**

	DeQA	Q-Insight	VQ-R1	CRI	CRA	CCS	AES
DIV2K-valid-augmented							
W/o DPO	3.932	3.752	4.344	55.80	71.85	79.44	72.24
SFT	4.236	3.939	4.656	70.95	78.65	84.02	79.80
One-stage DPO	**4.307**	3.972	4.692	77.05	79.35	**84.15**	82.46
Pro-DPO	4.303	**3.978**	**4.728**	**80.15**	**79.70**	82.66	**82.80**
RealOldPhotos							
W/o DPO	4.105	3.811	4.381	67.80	79.60	82.16	78.62
SFT	4.145	3.838	4.456	73.10	81.10	83.14	80.54
One-stage DPO	**4.207**	3.883	4.565	77.80	**82.40**	83.38	82.00
Pro-DPO	4.202	**3.894**	**4.600**	**80.40**	81.60	**83.72**	**82.28**

Effectiveness of Visual Semantic Prompt. We compare our visual semantic prompt design (before DPO training) with the commonly used text prompt approach in Tab. 2. For text prompt training, we utilize Qwen2.5-VL-72B to extract detailed descriptions from images, while filtering out phrases such as "black-and-white image" to avoid incorrect semantic propagation. As demonstrated, leveraging visual semantic prompts consistently outperforms text prompts across all evaluation metrics on both benchmarks. The improvement is especially significant on *RealOldPhotos*, as text prompts often fail to capture the full semantic context—particularly in complex, degraded scenes such as old photos—leading to reduced color richness (CRI) and inaccurate color assignments (low CRA) due to insufficient semantic guidance.

Effectiveness of Progressive DPO Training. We compare our proposed progressive DPO strategy with three settings: (i) the model before DPO training (w/o DPO); (ii) supervised fune-tuning (SFT) using (c, x^w) pairs; and (iii) one-stage DPO, where a wide range of color augmentations is applied throughout training. As demonstrated in Tab. 4, our proposed Pro-DPO achieves almost the best results across all metrics. Compared to one-stage DPO, our progressive strategy enables the model to learn more fine-grained color adjustments, resulting in colorization outputs that better align with modern color preferences. Additionally, as illustrated in Fig. 6, Pro-DPO produces the most vivid and natural colors, particularly in fine-grained regions such as individual fruits. This highlights the advantage of our progressive training strategy in capturing subtle color preferences and enhancing visual realism.

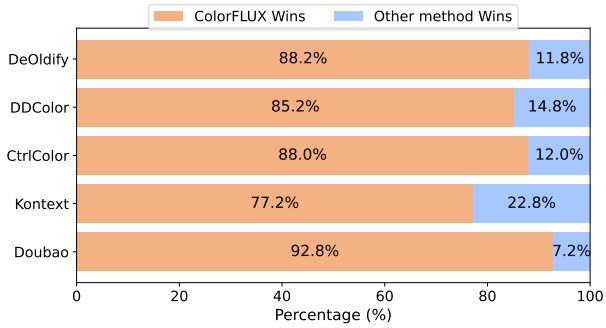

Figure 8. User study of ColorFLUX *vs.* other methods on *RealOldPhotos* among 10 participants.

5.4. User Study

We invited 10 experienced researchers in the image processing field to participate in our user study. Each participant was asked to conduct pairwise comparisons between the results of our method and those of competing methods, and to select the more preferred one. The comparisons were conducted on all images from *RealOldPhotos*. To avoid bias, we randomly shuffled the presentation order of the pairs for each participant. The results are presented in Fig. 8. As observed, our method achieves a significantly higher win rate over all baselines, demonstrating its strong advantage in generating visually preferred colorization results. More details are provided in the supplementary.

6. Conclusion

In this paper, we present a novel old photo colorization framework, dubbed ColorFLUX, that leverages a structure-color decoupling strategy and a progressive Direct Preference Optimization approach to address the unique challenges of colorizing degraded historical images. By incorporating visual semantic prompts, our method bypasses the limitations of text-based guidance and captures fine-grained semantics directly from the input. Extensive experiments on both synthetic and real datasets demonstrate that our approach achieves superior structural consistency and color vividness, surpassing existing state-of-the-art methods. However, our current implementation relies on the FLUX model, which limits its deployment on resource-constrained devices. In future work, we plan to explore lightweight alternatives to enable broader applicability while maintaining colorization performance.

Acknowledgments

This work was supported in part by NSFC under Grant 62371434 and U25B2010, the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20252293, the China Postdoctoral Science Foundation-Anhui Joint Support Program under Grant Number 2024T017AH, China Postdoctoral Science Foundation under Grant Number 2025M783529, Anhui Postdoctoral Scientific Research Program Foundation (No.2025A1015), the Fundamental Research Funds for the Central Universities (No. WK2100250064), and ZGCA Project-C20250302.

References

[1] Jason Antic. jantic/deoldify: A deep learning based project for colorizing and restoring old images (and video!). https://github.com/jantic/DeOldify, 2019. 2, 5, 7

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[3] Vukasin Bozic, Abdelaziz Djelouah, Yang Zhang, Radu Timofte, Markus Gross, and Christopher Schroers. Versatile vision foundation model for image and video colorization. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[4] Aurélie Bugeau, Vinh-Thong Ta, and Nicolas Papadakis. Variational exemplar-based image colorization. *IEEE Transactions on Image Processing*, 23(1):298–307, 2013. 3

[5] ByteDance. Doubao. https://www.doubao.com/, 2024. Accessed: 2025-07-29. 5

[6] Hernan Carrillo, Michaël Clément, and Aurélie Bugeau. Super-attention for exemplar-based image colorization. In *Proceedings of the Asian Conference on Computer Vision*, pages 4548–4564, 2022. 3

[7] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. L-coder: Language-based colorization with color-object decoupling transformer. In *European Conference on Computer Vision*, pages 360–375. Springer, 2022. 3

[8] 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*, pages 19221–19230, 2023. 3

[9] Youngin Cho, Junsoo Lee, Soyoung Yang, Juntae Kim, Yeojeong Park, Haneol Lee, Mohammad Azam Khan, Daesik Kim, and Jaegul Choo. Guiding users to where to give color hints for efficient interactive sketch colorization via unsupervised region prioritization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1818–1827, 2023. 3

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

[11] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE international conference on computer vision*, pages 567–575, 2015. 3

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[13] Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024. 3

[14] Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2024. 3

[15] 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. 3

[16] 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*, pages 328–338, 2023. 5, 6, 7

[17] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 3

[18] Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers, 2025. 1

[19] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3, 4, 6

[20] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 5, 7

[21] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 6

[22] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Control color: Multimodal diffusion-based interactive image colorization: Z. liang et al. *Inter-*

national Journal of Computer Vision, 133(11):7897–7923, 2025. 2, 3, 4, 5, 7

[23] Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, and Xinghao Ding. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22369–22380, 2025. 1

[24] Yunlong Lin, Zixu Lin, Kunjie Lin, Jinbin Bai, Panwang Pan, Chenxin Li, Haoyu Chen, Zhongdao Wang, Xinghao Ding, Wenbo Li, and Shuicheng Yan. Jarvisart: Liberating human artistic creativity via an intelligent photo retouching agent. *arXiv preprint arXiv:2506.17612*, 2025. 3

[25] Yunlong Lin, Linqing Wang, Kunjie Lin, Zixu Lin, Kaixiong Gong, Wenbo Li, Bin Lin, Zhenxi Li, Shiyi Zhang, Yuyang Peng, et al. Jarvisevo: Towards a self-evolving photo editing agent with synergistic editor-evaluator optimization. *arXiv preprint arXiv:2511.23002*, 2025. 3

[26] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 4

[27] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 4

[28] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[29] Gokhan Ozbulak. Image colorization by capsule networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 6

[30] Xinran Qin, Zhixin Wang, Fan Li, Haoyu Chen, RenJing Pei, WenBo Li, and XiaoChun Cao. Camedit: Continuous camera parameter control for photorealistic image editing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 3, 4

[32] Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7419–7427, 2025. 5

[33] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 3, 4

[34] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2747–2757, 2020. 2, 5

[35] Yi Wang, Menghan Xia, Lu Qi, Jing Shao, and Yu Qiao. Palgan: Image colorization with palette generative adversarial networks. In *European Conference on Computer Vision*, pages 271–288. Springer, 2022. 3

[36] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. Ct 2: Colorization transformer via color tokens. In *European conference on computer vision*, pages 1–16. Springer, 2022. 3

[37] 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*, pages 2677–2684, 2022.

[38] 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. 3

[39] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 5

[40] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025. 6

[41] Runsheng Xu, Zhengzhong Tu, Yuanqi Du, Xiaoyu Dong, Jinlong Li, Zibo Meng, Jiaqi Ma, Alan Bovik, and Hongkai Yu. Pik-fix: Restoring and colorizing old photos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1724–1734, 2023. 2, 5

[42] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024. 3

[43] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 5

[44] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14483–14494, 2025. 6

[45] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *ICCV*, pages 25669–25680, 2024. 1

[46] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 2023. 1

[47] Jooyeol Yun, Sanghyeon Lee, Minho Park, and Jaegul Choo. icolorit: Towards propagating local hints to the right region in interactive colorization by leveraging vision transformer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1787–1796, 2023. 3

[48] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 6

[49] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024. 6

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 4