Unleashing the Semantic Adaptability of Controlled Diffusion Model for Image Colorization

Xiangcheng Du[1] , **Zhao Zhou**[1] , **Yanlong Wang**[2] , **Yingbin Zheng**[3] ,
Xingjiao Wu[4] , **Peizhu Gong**[1] , **Cheng Jin**[1,5*]
[1]Fudan University, Shanghai, China
[2]Shanxi University, Taiyuan, China
[3]Videt Lab, Shanghai, China
[4]East China Normal University, Shanghai, China
[5]Innovation Center of Calligraphy and Painting Creation Technology, MCT, China

Abstract

Recent data-driven image colorization methods have leveraged pre-trained Text-to-Image (T2I) diffusion models as generative prior, while still suffering from unsatisfactory and inaccurate semantic-level color control. To address these issues, we propose a **Se**mantic **Ada**ptation method (**SeAda**) that enhances the prior while considering the semantic discrepancy between color and grayscale image pairs. The SeAda employs a semantic adapter to produce refined semantic embeddings and a controlled T2I diffusion model to create reasonably colored images. Specifically, the semantic adapter transfers the embedding from grayscale to color domain, while the diffusion model utilizes the refined embedding and prior knowledge to achieve realistic and diverse results. We also design a three-staged training strategy to improve semantic comprehension and prior integration for further performance improvement. Extensive experiments on public datasets demonstrate that our method outperforms existing state-of-the-art techniques, yielding superior performance in image colorization.

1 Introduction

Image colorization aims to assign plausible color information to grayscale images and produce visually realistic colored versions [Kang *et al.*, 2023; Zabari *et al.*, 2023]. This technique relies on the inference of appropriate colors for diverse objects, principally guided by semantic understanding of the scene [Weng *et al.*, 2024; Liang *et al.*, 2024].

The emergence of deep learning has drawn significant attention and shows encouraging advancements in image colorization. Previous methods use CNNs to predict per-pixel color distribution [Cheng *et al.*, 2015; Zhang *et al.*, 2016]. Despite their early promise, these approaches may produce suboptimal results due to the lack of image semantic understanding. To overcome the limitations, recent techniques [Kim *et al.*, 2022; Wu *et al.*, 2021] resort to GANs, exploiting their rich representations as generative priors for more accurate colorization. Nevertheless, the constrained representation space leads to unsaturated results or unpleasant artifacts. Moreover, some methods integrate transformer [Vaswani *et al.*, 2017] for colorization, but they may cause noticeable color bleeding in complex scenarios [Weng *et al.*, 2022; Ji *et al.*, 2022; Kumar *et al.*, 2021].

Figure 1: Semantic descriptions of grayscale and color images are generated by pre-trained captioner. The semantic description of color images includes additional attributes, such as color and style, which are absent in the descriptions of grayscale images.

In recent studies, denoising diffusion probabilistic models (DDPMs) [Ho *et al.*, 2020] possess remarkable image generation efficacy, showing their ability to produce high-quality visuals. Researchers have leveraged pre-trained DDPMs to solve image restoration tasks. However, their applicability to various scenarios is constrained by semantic understanding. In the development of the field, large-scale pre-trained text-to-image (T2I) models [Rombach *et al.*, 2022; Saharia *et al.*, 2022a] have been trained on surpassing 5 billion image-text pairs, and have risen as formidable tools in generating diverse visual outputs. These models excel at embedding rich semantic information into generated images, thereby enhancing the relationship between semantic and visual information.

While T2I models have harnessed substantial semantic knowledge from image-text datasets, their training is mostly on color images, which constrains semantic inference for grayscale images in colorization task. There are discrepancy

*Corresponding author. E-mail: jc@fudan.edu.cn

in semantic comprehension between color and grayscale images. As shown in Figure 1, we utilize a pre-trained captioner [Yu *et al.*, 2022] to derive semantic descriptions from both grayscale and color images. The semantics extracted from grayscale images tend to focus on structural elements and overall content. The color images provide additional color and style related information. When grayscale image semantics are insufficient for accurate color inference, the additional semantics can serve as a powerful supplement. We believe that refined semantic information can yield more pleasing visualizations.

Based on the aforementioned observations, we present a novel method named **SeAda**, **Se**mantic **Ada**ptation for image colorization. SeAda leverages refined semantic embedding to strengthen the generative capabilities of pretrained T2I models. SeAda contains two main components: semantic adapter and controlled T2I diffusion model. The semantic adapter focuses on refining initial semantic embeddings of grayscale images, transferring the embedding from grayscale to color domain and capturing more detailed semantic representation. Subsequently, the controlled T2I diffusion model exploits the refined embeddings to achieve superior colorization. In terms of training methodology, we design a staged training strategy: we first focus on optimizing the initial semantic embedding to procure refined embedding. Subsequently, we fine-tune the pre-trained controlled T2I diffusion model, which receives the refined embeddings to recover image colors. Finally, the semantic adapter and controlled T2I diffusion model are jointly trained to establish the relation between semantic embedding and generative prior.

Overall, our contributions can be summarized as follows:

- We introduce semantic-visual information into an image colorization method based on T2I diffusion models. The refined semantic embedding can capture comprehensive semantic information compared with initial embedding through our carefully designed semantic adapter.
- We present a staged training strategy that incrementally refines the capability of the model. Each component is specifically optimized for more accurate colorizations.
- Extensive experimental results on public datasets show that our method achieves state-of-the-art performance on both colorful metrics and perceptual quality.

2 Related Work

2.1 Image Colorization

Colorization methods are generally classified into two categories: automatic colorization methods and conditional image colorization methods.

Automatic colorization. Early methods treat image colorization as a classification task [Zhang *et al.*, 2016]. To incorporate semantic information into colorization, researchers have integrated class labels [Kim *et al.*, 2022] or instance bounding boxes [Su *et al.*, 2020] into the colorization networks. Subsequently, GANs [Goodfellow *et al.*, 2014] have shown promising results in this task. Notably, methods like GCPColor [Wu *et al.*, 2021] and BigColor [Kim *et al.*, 2022] harness the generative priors of pre-trained GANs to enhance performance. Furthermore, leveraging the extensive receptive field of Transformers [Vaswani *et al.*, 2017], recent advancements involve the prediction of color tokens, thereby improving the contextual relevance of the colorization results [Kumar *et al.*, 2021; Huang *et al.*, 2022]. DDColor [Kang *et al.*, 2023] includes a multi-scale image decoder and a transformer-based color decoder. The two decoder aims to learn semantic-aware color embedding and optimize color queries. MultiColor [Du *et al.*, 2024] automatically colorize grayscale images that combines clues from multiple color spaces.

Conditional colorization introduces user-defined controls into colorization. The technique can be divided into three categories: stroke-based, reference-based, and prompt-based colorization. Stroke-based colorization utilizes similarity metrics, such as spatial offsets or neural network-learned features [Levin *et al.*, 2004; Endo *et al.*, 2016] to disseminate localized color hints throughout the image. Reference-based colorization leverages a pre-trained network to perform feature matching between the grayscale image and a color reference image [He *et al.*, 2018; Huang *et al.*, 2022]. This colorization not only maintains the structural integrity of the grayscale image but also closely aligns with the color distribution of the reference. Prompt-based colorization integrates textual descriptions to guide the colorization process. Techniques of this category have evolved to include the fusion of textual and visual features [Chen *et al.*, 2018]. More recent works [Chang *et al.*, 2023; Zabari *et al.*, 2023] use diffusion prior to achieve prompt control, such as L-CAD [Chang *et al.*, 2023], Diffusing Colors [Zabari *et al.*, 2023], Control Color [Liang *et al.*, 2024] and GoLoColor [Yue *et al.*, 2025].

2.2 Diffusion Models in Image Restoration

Diffusion models have been applied to image restoration. Based on generation space, diffusion-based image restoration can be divided into image and latent space based methods.

The image space-based methods directly synthesize structures and textures. SR3 [Saharia *et al.*, 2022b] leverages diffusion model to generate conditional images and achieves super-resolution through a stochastic denoising process. Whang et al. [Whang *et al.*, 2022] present a novel framework for blind image deblurring, utilizing conditional diffusion models for this application. For image inpainting task, Repaint [Lugmayr *et al.*, 2022] leverages pre-trained unconditional DDPM [Ho *et al.*, 2020] as the generative basis and modifies the reverse diffusion process to incorporate samples from the unmasked regions of the available image. Additionally, diffusion models are also applied in various image restoration tasks, such as image denoising [Feng *et al.*, 2023], low-light enhancement [Wang *et al.*, 2023], and shadow removal [Guo *et al.*, 2023].

The latent space-based methods utilize a well-designed encoder to convert images into a latent representation, thereby enhancing the efficiency of generation processes. DiffBIR [Lin *et al.*, 2024] achieves realistic image restoration by leveraging the generative capacities of the pre-trained Stable Diffusion. Besides, text-to-image diffusion models encode text inputs into latent vectors using pre-trained language models [Radford *et al.*, 2021] and achieve state-of-the-art re-

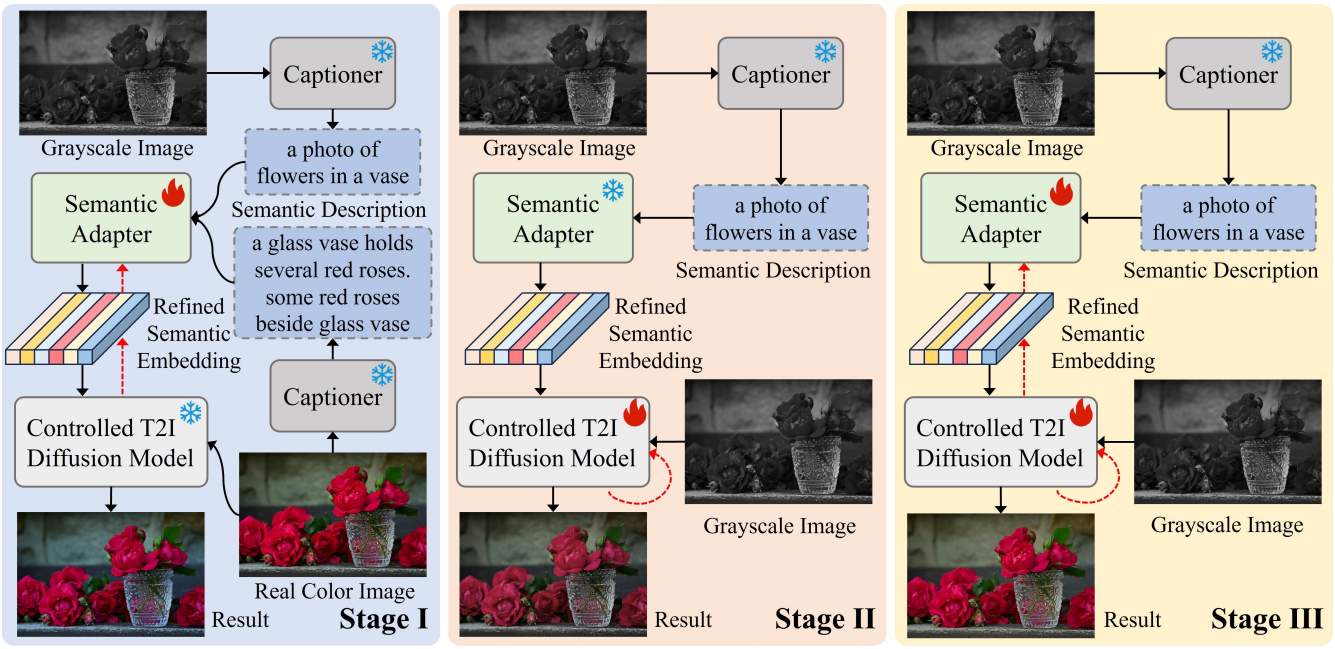

Figure 2: The overall framework. **Stage I**: The semantic embedding of grayscale image is optimized using semantic adapter. The semantic description of grayscale and real color image is produced by pre-trained captioner. **Stage II**: The semantic adapter is frozen to preserve the refined semantic embeddings. We then fine-tune the controlled T2I diffusion model, conditioned with grayscale image to predict colored result. **Stage III**: The semantic adapter and controlled T2I diffusion model are jointly trained to achieve the desired colorized result. The red arrow indicates model updates.

sults in image restoration. SeeSR [Wu *et al.*, 2024] explores the impact of different styles of text prompts on the generated super-resolution image. BFRffusion [Chen *et al.*, 2024] leverages generative priors encapsulated in pretrained Stable Diffusion for blind face restoration.

3 Semantic Adaptation Image Colorization

3.1 Preliminaries

We begin with a brief review of Stable Diffusion (SD) [Rombach *et al.*, 2022], which is the basis of the proposed approach. SD is structured as a two-stage diffusion model, which comprises an autoencoder and an UNet denoiser. In the first stage, the autoencoder is trained to encode images, denoted as $\mathbf{x}_0$, into a latent representation $\mathbf{z}_0$ and reconstruct them. In the second stage, the UNet denoiser executes the denoising operation directly within the latent space. The optimization process can be defined as follows:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_t,\mathbf{c},\boldsymbol{\epsilon},t}(||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})||_2^2), \tag{1}$$

where $\mathbf{z}_t = \sqrt{\overline{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}$ represents the noised feature map at step t, and $\boldsymbol{\epsilon} \in \mathcal{N}(0, \mathbf{I})$. $\mathbf{c}$ represents the conditional information. $\boldsymbol{\epsilon}_\theta$ refers to the function of UNet denoiser. During inference, the input latent map $\mathbf{z}_t$ is synthesized from a random Gaussian distribution. Given $\mathbf{z}_t$, ϵ_θ estimates the noise at each step t, conditioned on $\mathbf{c}$. Through iterative subtraction of the estimated noise, the noised feature map is progressively clarified. After T iterations, the refined latent feature $\hat{\mathbf{z}}_0$ is decoded by the autoencoder to generate the final image. In the conditional part, SD utilized the pre-trained CLIP [Radford *et al.*, 2021] text encoder to map text inputs to sequence representation. Subsequently, cross-attention mechanism integrates the representation into the denoising process.

3.2 Framework

Given a grayscale image $\mathbf{x}_g \in \mathbb{R}^{H\times W\times 1}$, our goal is to create colorized version $\mathbf{x}_c \in \mathbb{R}^{H\times W\times 3}$ that exhibits semantic correctness and visual fidelity. To achieve this, we explicitly modeling semantic information for controlled T2I diffusion model to facilitate colorization process. Distinct from previous diffusion-based colorization approaches that rely on predefined text prompts, SeAda leverages a semantic adapter to produce refined semantic embedding. When the controlled T2I diffusion model is equipped with the refined semantic embeddings, it is capable of generating colored images that align well with the semantic information. As depicted in Figure 2, SeAda employs a three-stage training strategy and we will describe each stage in detail.

Stage I: Semantic embedding optimization. We start to refine initial semantic embedding of grayscale image through semantic adapter. Before the refinement process, both the grayscale image and corresponding real color image are first processed through individual fixed captioner, which produce semantic descriptions denoted as Des_g and Des_c. Here we leverage CoCa [Yu *et al.*, 2022] to generate these descriptions. Then the Des_g and Des_c are fed into the semantic adapter and mapped into semantic embedding by individual semantic encoder. The purpose of the semantic adapter is to transfer the semantic embedding of grayscale image into em-

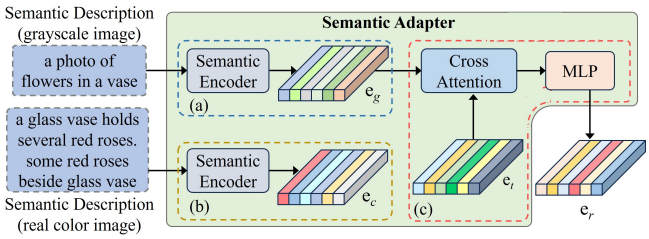

Figure 3: Architecture of the semantic adapter.

bedding of color image description. The refined semantic embedding provides deeper understanding of image content. We freeze the parameters of the controlled T2I diffusion model and optimize the initial semantic embedding to produce colorized image. Note that, the conditional image is real color image which enables refined semantic embedding to match the color image as closely as possible.

Stage II: Controlled T2I diffusion model fine-tuning. This stage differentiates from Stage I by focusing on fine-tuning the controlled T2I diffusion model rather than the operation of the semantic adapter. Here, the adapter only processes the semantic description of grayscale image and generates its corresponding semantic embedding. Following the Stage I, the refined semantic embedding is enriched with more comprehensive information. We freeze the semantic adapter to maintain the quality of the embeddings and finetune controlled T2I diffusion model. This operation allows the model to effectively incorporate the refined semantic embeddings into the colorization process and enhances the overall quality of colorized results. In this stage, the conditional image of the controlled T2I diffusion model is grayscale image, ensuring colorization aligns with the original content and structure.

Stage III: Jointly training. We finally jointly train the semantic adapter and controlled T2I diffusion model to achieve the desired colorized result. Although Stage II established robust color recovery model, the feature distribution between different components is still biased. To address the limitation, we focus on optimization of the whole model to enhance its performance and adaptability. In this stage, the parameters of both the semantic adapter and the controlled T2I diffusion model are fully trainable. The joint training manner allows semantic adapter and T2I diffusion model to further learn from each other, which can fuse information of different modalities effectively.

3.3 SeAda Design

Semantic adapter. The semantic adapter is segmented into three parts as shown in Figure 3. In part (a), the semantic description of grayscale image Des_g is processed through a semantic encoder [Radford *et al.*, 2021]. The encoder transforms Des_g into its corresponding semantic embedding $\mathbf{e}_g \in \mathbb{R}^{L\times d}$, where L is the length of the embedding, and d is the embedding dimension. Meanwhile, semantic descriptions of real color images are also encoded into $\mathbf{e}_c$ in part (b). Although the semantic embedding of grayscale images provides approximative semantic information, it struggles to capture color attributes present in the color images. In order

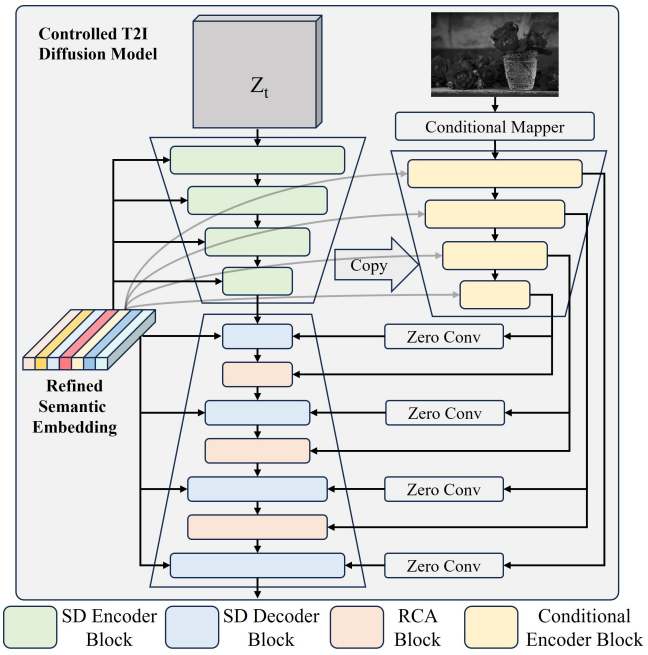

Figure 4: Architecture of Controlled T2I diffusion model.

to enhance semantic understanding effectively, we integrate a trainable semantic embedding, denoted as $\mathbf{e}_t$ in part (c). The embedding is designed to synergize with $\mathbf{e}_g$ to guide the colorization process. Specifically, the trainable semantic embedding $\mathbf{e}_t \in \mathbb{R}^{L\times d}$ matches the size of the output of semantic encoder. To further refine the semantic embedding, a cross-attention layer is introduced to facilitate interaction between $\mathbf{e}_g$ and $\mathbf{e}_t$. The overall process is defined as:

$$\mathbf{e}_g = \texttt{SeEnc}(\text{Des}_g) \tag{2}$$

$$\mathbf{e}_r = \texttt{MLP}(\texttt{softmax}(\mathbf{e}_t \cdot \mathbf{e}_g^T/\alpha) \cdot \mathbf{e}_g + \mathbf{e}_g) \tag{3}$$

where $\texttt{SeEnc}$ is semantic encoder, $\texttt{MLP}$ is MLP layers, and α is scaling parameter to control the magnitude of the dot product before applying the $\texttt{softmax}$. $\mathbf{e}_r$ is refined semantic embedding and can provide valuable semantic guidance for the following colorization process.

Controlled T2I diffusion model. Figure 4 illustrates the detailed structure of the controlled T2I diffusion model. Benefit from the success of ControlNet [Zhang *et al.*, 2023] in conditional image generation, we employ it as the controller within our T2I diffusion model for image colorization. Following ControlNet [Zhang *et al.*, 2023], we make a trainable copy of the pre-trained SD encoder as conditional encoder, designated as the conditional encoder (E_{con}). The encoder processes the conditional image to generate control information. The copy strategy offers an advantageous weight initialization for the network. Initially, the input conditional image with dimensions of 512×512 is downscaled to 64×64 latent space vector $\mathbf{c}_f$ that matches the size of Stable Diffusion. The transformation is implemented by a conditional mapper comprising four convolutional layers, each utilizing 4×4 kernels and 2×2 strides. Then, the concatenation of $\mathbf{c}_f$ and the noisy latent $\mathbf{z}_t$ at time t forms the input to E_{con}. Additionally, we

incorporate zero convolutions within $\mathtt{E}_{con}$ to avoid random noise as gradients during the early training stage and enhance the stability of the model.

In contrast to ControlNet which adds the conditional representation directly to SD decoder, we incorporate Representation Cross-Attention (RCA) [Yang *et al.*, 2024; Wu *et al.*, 2024] in Stage II and Stage III, to better preserve the structural integrity of the image. The RCA block are placed after the SD decoder block.

3.4 Training Objectives

Stage I. We fixed the parameters of pre-trained controlled T2I diffusion model, and focus on refining the semantic embedding $\mathbf{e}_r$ using the denoising diffusion objective:

$$\mathcal{L}_s = \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{e}_r,\mathbf{c}_f,\theta}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}_f, \mathbf{e}_r)\|_2^2\right] \tag{4}$$

where t represents a randomly selected timestep between 1 and T, $\mathbf{z}_t$ is a noisy version of $\mathbf{z}_0$ (the latent representation of real color image $\mathbf{x}_c$) by using Gaussian noise $\boldsymbol{\epsilon} \in \mathcal{N}(0, \mathbf{I})$, and θ is the diffusion model weights. Note that during this stage, part (b) of the semantic adapter is frozen, whereas parts (a) and (c) remain trainable.

To enhance the robustness of the semantic embedding, we impose a constraint that encourages the semantic embedding $\mathbf{e}_g$ derived from the grayscale image, to closely align with the embedding of the real color image $\mathbf{e}_c$ by MSE loss:

$$\mathcal{L}_r = \mathtt{MSE}(\mathbf{e}_c, \mathbf{e}_g) \tag{5}$$

The objective is to ensure the semantic embedding accurately reflects color image. The overall objective for learning the semantic embedding is formulated as:

$$\mathcal{L}_e = \mathcal{L}_r + \lambda_s \mathcal{L}_s \tag{6}$$

where λ_s is set to 0.5 in our experiments.

Stage II. The conditional image at this stage is grayscale image, we specifically utilize parts (a) and (c) of the semantic adapter to predict the refined semantic embedding $\mathbf{e}_r$. Since the $\mathbf{e}_r$ has been refined, we freeze the semantic adapter. This allows us to focus on optimizing the parameters θ of the controlled T2I diffusion model. The optimization of these parameters is guided by the same loss function presented in Eq. (4). The $\mathbf{z}_t$ is obtained by latent representation of grayscale image $\mathbf{x}_g$ and Gaussian noise $\boldsymbol{\epsilon} \in \mathcal{N}(0, \mathbf{I})$.

Stage III. In this stage, we jointly train semantic adapter and conditioned T2I diffusion model using the loss function presented in Eq. (4). Consistent with Stage II, the semantic adapter continues to utilize parts (a) and (c). The distinction in final stage is that the semantic adapter is fully trainable, unlike in Stage II where it was fixed. This adjustment allows the semantic adapter to dynamically refine its performance based on the feedback received from the training process.

4 Experiments

Dataset. To keep fairness with previous methods, our experiments are conducted on ImageNet [Russakovsky *et al.*, 2015]. We utilize the training part of ImageNet to train our model, assessing its performance on the validation set. Besides, in order to demonstrate the generalization capability of our method, we further evaluate it on the validation sets of COCO-Stuff [Caesar *et al.*, 2018] and ADE20K [Zhou *et al.*, 2017] without any fine-tuning.

Evaluation metrics. The principal evaluation criteria for image colorization include perceptual realism and color vividness. To measure perceptual realism, we utilize the Fréchet Inception Score (FID) [Heusel *et al.*, 2017] to quantify the distribution similarity between the predictions and ground truth images. For assessing color vividness, we employ the absolute colorfulness score difference ΔCF [Hasler and Suesstrunk, 2003]. This metric compares the colorfulness between the real color and recolored images. Additionally, we include the Peak Signal-to-Noise Ratio (PSNR) metric commonly used standard in prior works [Vitoria *et al.*, 2020; Su *et al.*, 2020; Wu *et al.*, 2021; Ji *et al.*, 2022]. While PSNR provides a measure of error relative to the ground truth, plausible colorization outcomes may vary significantly in hue and saturation from the original images.

Implementation details. Stable Diffusion 1.5 is adopted as foundational denoising diffusion network. Following ControlNet [Zhang *et al.*, 2023], we pretrain the controlled T2I diffusion model of Stage I with grayscale conditional image. We use AdamW optimizer [Kingma and Ba, 2015] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 10^{-5}. We train our model for 80K steps with a batch size of 32. The first 20K training aims to optimize semantic adapter and the later 50K steps add semantic embedding for fine-tuning T2I diffusion model. The final 10K is for joint training. All experiments are conducted on 4 NVIDIA Tesla A100 GPUs.

4.1 Comparisons with Previous Methods

We compare performance with previous colorization methods, including CNN-based methods, GAN-based methods, transformer-based methods, and diffusion-based methods.

Quantitative comparison. We compare quantitative performance with recent colorization methods as reported in Table 1. Our method achieves sota performance in most metrics. Specifically, SeAda achieves 0.06dB (24.53→24.47) and 0.33 (3.69→3.36) improvements in terms of PSNR and FID on the ImageNet-5K test set compared with the second-best method Diffusing Colors [Zabari *et al.*, 2023] and L-CAD [Chang *et al.*, 2023], respectively. On the COCO-Stuff test set, SeAda obtains FID and PSNR of 5.07 and 24.37dB that greatly outperform the second-best approaches, i.e., DDColor [Kang *et al.*, 2023] and DeOldify [Antic, 2019], by 0.11 (5.18→5.07) and 0.18db (24.37→24.19). On the ADE20K test set, our method improves the two metrics by at least 1.93 (=7.64→5.71) and 0.47db (24.87→24.40). In terms of ΔCF, the lower ΔCF values imply more precise colorization, with our method achieving comparable ΔCF values, highlighting its efficacy in generating natural and lifelike colorization results.

Qualitative comparison. Figure 5 presents visualization of colorization results. We display comparisons of images in different scenes from the ImageNet validation dataset. Note

Method	ImageNet (val5k)				COCO-Stuff				ADE20K			
	FID↓	CF↑	ΔCF↓	PSNR↑	FID↓	CF↑	ΔCF↓	PSNR↑	FID↓	CF↑	ΔCF↓	PSNR↑
CIC [Zhang *et al.*, 2016]	8.72	31.60	6.61	22.64	27.88	33.84	4.40	22.73	15.31	31.92	3.12	23.14
DeOldify [Antic, 2019]	6.59	21.29	16.92	24.11	13.86	24.99	13.25	24.19	12.41	17.98	17.06	24.40
InstColor [Su *et al.*, 2020]	8.06	24.87	13.34	23.28	13.09	27.45	10.79	23.38	15.44	23.54	11.50	24.27
GCPColor [Wu *et al.*, 2021]	5.95	32.98	5.23	21.68	13.97	28.41	9.83	24.03	13.27	27.57	7.47	22.03
CT2 [Weng *et al.*, 2022]	5.51	38.48	0.27	23.50	13.15	36.22	2.02	23.67	11.42	**35.95**	0.91	23.90
BigColor [Kim *et al.*, 2022]	5.36	39.74	1.53	21.24	12.58	36.43	1.81	21.51	11.23	35.85	0.81	21.33
ColorFormer [Ji *et al.*, 2022]	4.91	38.00	0.21	23.10	8.68	36.34	1.90	23.91	8.83	32.27	2.77	23.97
Unicolor [Huang *et al.*, 2022]	4.36	41.38	3.17	22.83	10.62	**41.91**	3.67	23.56	10.87	32.92	2.12	23.85
Diffusing Colors [Zabari *et al.*, 2023]	3.69	38.42	0.21	-	8.05	36.56	1.68	-	-	-	-	-
DDColor [Kang *et al.*, 2023]	3.92	38.26	**0.05**	23.85	5.18	38.48	**0.24**	22.85	8.21	34.80	**0.24**	24.13
L-CAD [Chang *et al.*, 2023]	4.36	34.53	3.68	24.47	7.36	40.78	2.54	23.87	7.64	35.55	0.51	24.07
Control Color [Liang *et al.*, 2024]	4.29	**44.72**	6.51	-	10.26	33.00	5.24	-	-	-	-	-
SeAda [Ours]	**3.36**	36.98	1.23	**24.53**	**5.07**	38.99	0.75	**24.37**	**5.71**	35.41	0.37	**24.87**

Table 1: Quantitative comparison of different methods on benchmark datasets. Best and second best results are in **bold** and underlined respectively. ↑ (↓) indicates higher (lower) is better.

Figure 5: Visual comparison of previous methods on image colorization.

that the GT images are provided for reference only but the evaluation criterion should not be color similarity. A noticeable trend is that our results exhibit a more vivid appearance. We can see that the booth colorization (Row 1) of previous methods looks unnatural in contrast to our consistent color. Meanwhile, our method produces more saturated colorization results. InstColor [Su *et al.*, 2020] employs a pre-trained detector to detect objects and cannot color the whole image well (Column 2). GCPColor [Wu *et al.*, 2021] and ColorFormer [Ji *et al.*, 2022] usually lead to incorrect semantic colors and low color richness. The Control color [Liang *et al.*, 2024] correctly estimated realistic colors, but the shapes are warped and the fine-grained details in the image have been roughened. Instead, our method maintains the consistent color and captures the details as shown in row 2 of Figure 5. Furthermore, our method can yield more diverse and lively colors for the whole image as shown in the last row.

Stage I	Stage II	Stage III	FID↓	ΔCF↓	PSNR↑
	✓		6.93	4.31	22.67
		✓	5.37	3.42	23.16
✓	✓		3.72	2.06	23.74
✓		✓	3.47	2.11	23.53
✓	✓	✓	**3.36**	**1.23**	**24.53**

Table 2: Ablation studies of training strategy.

4.2 Ablation Study

In this section, we explore the effects of different designs of our method. We conduct all experiments on ImageNet val-5k.

Staged training strategy. Table 2 shows the impact of staged training strategy for colorization. We observe that staged training significantly improves colorization performance. The goal of stage I is to optimize semantic embedding and the conditional image of T2I diffusion model is real

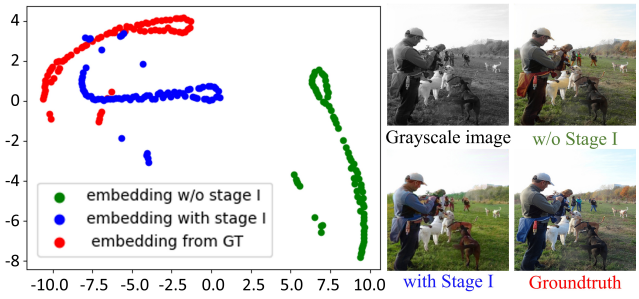

Figure 6: Visualization for the semantic embeddings and colorization results. Compared with initial semantic embedding of grayscale image, the refined semantic embeddings are more closely with the embedding of real color images. The refined semantic embedding can guide the model to restore more realistic colors.

Adapter	FID↓	ΔCF↓	PSNR↑
(a)	6.47	3.77	22.89
(a)+(b)	4.12	1.82	24.26
(a)+(b)+(c)	**3.36**	**1.23**	**24.53**

Table 3: Effects of semantic adapter.

color image, thus the results of this stage are not reported. Row 1 shows the results of training the model only using Stage II, where the controlled T2I diffusion model is updated under the grayscale semantic embedding. In Row 3, the semantic adapter is optimized using Stage I, and the model exhibits a significantly improved ability (PSNR=6.93→3.72, ΔCF=4.31→2.06).

For semantic embedding visualization, we utilize t-SNE technology [Van der Maaten and Hinton, 2008] for dimensionality reduction. The visualization of the embedding (with & w/o Stage I training) and corresponding colorization results are shown in Figure 6. Stage III trains semantic adapter and T2I diffusion model at the same time, and the performance is better than Stage II training. Moreover, staged training strategy iteratively refines the parameters, leading to improvements in the colorization results.

Semantic adapter. Recall that different parts of the semantic adapter as shown in Figure 3. To assess the effectiveness of semantic adapter, we conduct experiments by selectively removing parts from the full setting. The quantitative results are detailed in Table 3. As shown in row 1, only part (a) results in decreases in all metrics, as the generative capacity of T2I diffusion models relies heavily on semantic embedding and part (a) only produces semantic embedding of the grayscale image. The part (b) is designed to learn robust representation embedding, thus combining part (b) can boost the performance. The incorporation of trainable embedding (part (c)) further yields noticeable improvements, which can be improved by 0.76 and 0.59 for FID and ΔCF, respectively.

Controlled T2I diffusion model. The default diffusion model contains a series of RCA modules. To test the importance of the RCA modules, we remove them, which caused a significant drop in performance. Table 4 shows that RCA modules yield 1.22 dB PSNR gain than without RCA.

Configure	FID↓	ΔCF↓	PSNR↑
w/o RCA	4.57	2.85	23.31
with RCA	3.26	1.23	24.53

Table 4: Effects of RCA module.

Captioner	FID↓	ΔCF↓	PSNR↑
ClipCap [Mokady *et al.*, 2021]	3.47	1.25	23.95
BLIP [Li *et al.*, 2022]	3.51	1.66	23.85
BLIP2 [Li *et al.*, 2023]	3.45	1.46	**24.53**
CoCa [Yu *et al.*, 2022]	**3.36**	**1.23**	24.48

Table 5: Ablation studies of captioner.

Captioner. We explore the influence of different captioners for our image colorization model: ClipCap [Mokady *et al.*, 2021], BLIP [Li *et al.*, 2022], BLIP2 [Li *et al.*, 2023], and CoCa [Yu *et al.*, 2022]. The results of the ablation experiments are shown in Table 5. We can observe that CoCa captioner for our model achieves the most satisfactory results. Moreover, while the CoCa captioner leads in performance, replacing it with other captioners does not result in a significant degradation of performance. Our model requires the captioner to generate semantic information for each image which may decelerate the colorization speed.

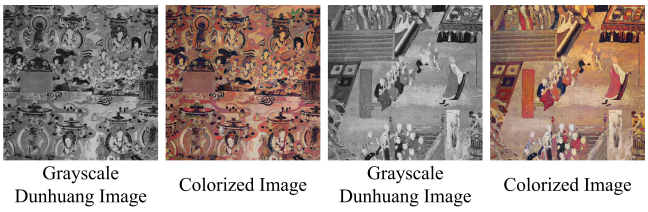

Figure 7: Colorizing real historical grayscale photos.

4.3 Real-world Applications

We collect some real historical photos, such as Dunhuang grayscale images, to demonstrate the capability of our method in real-world scenarios (as shown in Figure 7).

5 Conclusion

We propose a diffusion-based framework named SeAda for efficient image colorization. The technical core of our approach lies in a controlled T2I diffusion model, which leverages the refined semantic embedding combined with the generative prowess of diffusion models to yield visually satisfactory results. To enhance the semantic embedding of grayscale images, we implement a semantic adapter that transfers the embeddings from grayscale to color domain. Furthermore, we design a staged training strategy to improve semantic understanding and generative priors for further performance improvements. Experimental results on publicly available benchmarks demonstrate that our method outperforms previous methods both quantitatively and qualitatively.

This work was supported by AI for Science Foundation of Fudan University (FudanX24AI028) and National Archives Administration of China Research Program (2024-X-013).

References

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

[Caesar *et al.*, 2018] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018.

[Chang *et al.*, 2023] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-cad: Language-based colorization with any-level descriptions. *NeurIPS*, 2023.

[Chen *et al.*, 2018] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *CVPR*, 2018.

[Chen *et al.*, 2024] Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaochun Cao. Towards real-world blind face restoration with generative diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Cheng *et al.*, 2015] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *ICCV*, pages 415–423, 2015.

[Du *et al.*, 2024] Xiangcheng Du, Zhao Zhou, Xingjiao Wu, Yanlong Wang, Zhuoyao Wang, Yingbin Zheng, and Cheng Jin. Multicolor: Image colorization by learning from multiple color spaces. In *ACM Multimedia*, 2024.

[Endo *et al.*, 2016] Yuki Endo, Satoshi Iizuka, Yoshihiro Kanamori, and Jun Mitani. Deepprop: Extracting deep features from a single image for edit propagation. In *Computer Graphics Forum*, 2016.

[Feng *et al.*, 2023] Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *ICCV*, pages 10520–10531, 2023.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, 2014.

[Guo *et al.*, 2023] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *CVPR*, 2023.

[Hasler and Suesstrunk, 2003] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, volume 5007, pages 87–95. SPIE, 2003.

[He *et al.*, 2018] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics*, 2018.

[Heusel *et al.*, 2017] 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. In *NeurIPS*, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020.

[Huang *et al.*, 2022] Zhitong Huang, Nanxuan Zhao, and Jing Liao. Unicolor: A unified framework for multimodal colorization with transformer. *ACM Transactions on Graphics*, 2022.

[Ji *et al.*, 2022] 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 *ECCV*, pages 20–36. Springer, 2022.

[Kang *et al.*, 2023] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. Ddcolor: Towards photo-realistic image colorization via dual decoders. In *ICCV*, pages 328–338, 2023.

[Kim *et al.*, 2022] 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 *ECCV*, pages 350–366. Springer, 2022.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Kumar *et al.*, 2021] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *ICLR*, 2021.

[Levin *et al.*, 2004] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH*. 2004.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023.

[Liang *et al.*, 2024] 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.

[Lin *et al.*, 2024] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, et al. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*. Springer, 2024.

[Lugmayr *et al.*, 2022] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

[Mokady *et al.*, 2021] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

[Saharia *et al.*, 2022a] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, volume 35, pages 36479–36494, 2022.

[Saharia *et al.*, 2022b] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[Su *et al.*, 2020] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *CVPR*, pages 7968–7977, 2020.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Vitoria *et al.*, 2020] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *WACV*, pages 2445–2454, 2020.

[Wang *et al.*, 2023] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. In *ICCV*, 2023.

[Weng *et al.*, 2022] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. Ct2: Colorization transformer via color tokens. In *ECCV*, pages 1–16. Springer, 2022.

[Weng *et al.*, 2024] Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, Boxin Shi, et al. L-cad: Language-based colorization with any-level descriptions using diffusion priors. In *NeurIPS*, volume 36, 2024.

[Whang *et al.*, 2022] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *CVPR*, pages 16293–16303, 2022.

[Wu *et al.*, 2021] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *CVPR*, pages 14377–14386, 2021.

[Wu *et al.*, 2024] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024.

[Yang *et al.*, 2024] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*. Springer, 2024.

[Yu *et al.*, 2022] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[Yue *et al.*, 2025] Tianai Yue, Xiangcheng Du, Jing Liu, and Zhongli Fang. Golocolor: Towards global-local semantic aware image colorization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2025.

[Zabari *et al.*, 2023] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image colorization with text guided diffusion. In *ACM SIGGRAPH Asia*, 2023.

[Zhang *et al.*, 2016] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.