NoMoColor: Unified Noise Modulation for Enhanced Diffusion-based Image Colorization (Student Abstract)

Ankan Deria[1,2], **Dwarikanath Mahapatra**[3], **Murari Mondal**[4], **Sudipta Roy**[2]
[1]Mohamed bin Zayed University of Artificial Intelligence, UAE
[2]Jio University, INDIA
[3]Khalifa University, UAE
[4] Kalinga Institute of Industrial Technology, INDIA

Abstract

We present a language-based noise modulation module for diffusion models that improves image color generation under textual guidance. Unlike standard approaches that inject noise uniformly, our method leverages semantic cues from text to selectively control the noise injection process, preserving local details and enhancing color accuracy even when descriptions are ambiguous or incomplete. Applied to language-guided image colorization, this targeted modulation leads to more faithful and visually consistent results. The proposed module is lightweight, generalizable, and can be integrated into existing diffusion pipelines, offering a simple yet effective step toward more controllable text-to-image generation.

Figure 1: Comparison of language-based and automatic image colorization results with varying levels of detail.

Introduction

Colorizing grayscale images from textual descriptions is a challenging task, as users often provide only partial or vague color information. Existing language-based colorization methods (Chang et al. 2022, 2023) improve alignment between text and colorization through feature fusion or object–color decoupling, yet they typically assume detailed descriptions for most objects. This assumption limits performance when descriptions are incomplete, and it fails to preserve background regions that users generally expect to remain unchanged.

To address this issue, we propose a Unified **No**ise **Mo**dulation for Enhanced Diffusion-based Image **Color**ization named as NoMoColor. Standard diffusion injects noise uniformly, which erodes fine details and alters background context. In contrast, our module leverages semantic cues from text to selectively modulate noise during the forward process. By reducing noise in regions not mentioned in the description, the model preserves local structures while recoloring only the specified objects.

As shown in Figure 1, this targeted noise modulation allows accurate color mapping even under minimal or ambiguous descriptions, improving both fidelity and background preservation compared to prior approaches. Our contribution is simple, lightweight, and generalizable to other text-guided diffusion tasks such as editing and style transfer.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Methodology

Diffusion models generate images by gradually adding noise to a latent representation during the forward process and learning to reverse this corruption in the backward process. Standard noise injection is uniform across the image, which often destroys fine details and alters background regions, particularly when textual descriptions are incomplete.

To address this, we propose a **language-based noise modulation module** that selectively adjusts the noise level according to semantic cues in the text. Given an input description, we used SAM segmentation model (Kirillov et al. 2023) to identify regions relevant to the prompt. The resulting mask is encoded into the latent space, where noise is injected differently for masked and unmasked areas: more noise in target regions to enable recoloring, and reduced noise elsewhere to preserve background details (Figure 2). If no mask is available, the model defaults to standard noise addition. At each timestep $t \in \{0, \ldots, T\}$, the noisy latent z_t is defined as:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \epsilon_t, \tag{1}$$

for masked regions, and

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} (\delta \epsilon_t), \tag{2}$$

for unmasked regions, where $\epsilon_t \sim \mathcal{N}(0, 1)$ and $\delta < 1$ controls the reduced noise level. This formulation preserves both local structures and global consistency by limiting corruption in background regions. The denoising network ϵ_θ is

Method	Extended COCO-Stuff				Multi-instance			
	PSNR↑	LPIPS↓	C↑	ΔC↓	PSNR↑	LPIPS↓	C↑	ΔC↓
Text-Guided Colorization								
LBIE (Chen et al. 2018)	22.15	0.263	30.23	5.88	22.05	0.254	29.95	5.92
FILM (Manjunatha et al. 2018)	21.19	0.279	29.78	6.16	20.70	0.291	29.41	6.36
L-CoDe (Weng et al. 2022b)	24.96	0.169	32.94	4.53	23.96	0.172	32.38	4.48
L-CoDer (Chang et al. 2022)	25.50	0.159	33.52	4.13	24.22	0.165	32.92	4.04
L-CoIns (Chang et al. 2023)	25.51	0.157	34.24	3.45	24.81	0.162	33.03	3.61
L-CAD (Weng et al. 2024)	25.97	0.142	36.15	2.26	25.51	0.127	35.74	2.52
NoMoColor (Ours)	**27.92**	**0.133**	**41.25**	**1.06**	**27.46**	**0.118**	**42.34**	**1.02**
Automatic Colorization								
CIC (Zhang, Isola, and Efros 2016)	22.21	0.221	30.52	4.97	22.09	0.233	29.96	5.02
InstColor (Su, Chu, and Huang 2020)	23.79	0.194	25.94	11.62	23.61	0.202	25.43	11.68
ColorFormer (Ji et al. 2022)	24.12	0.188	38.34	1.34	24.04	0.196	37.87	1.37
DISCO (Xia et al. 2022)	20.77	0.208	**43.67**	11.45	20.88	0.202	**43.24**	11.51
ColTran (Kumar, Weissenborn, and Kalchbrenner 2021)	20.68	0.314	35.12	3.61	20.34	0.321	34.56	3.85
GCP (Wu et al. 2021)	23.94	0.186	31.82	5.39	23.78	0.189	31.35	5.44
CT2 (Weng et al. 2022a)	24.16	0.184	39.79	2.11	24.05	0.186	39.48	2.15
NoMoColor (Ours)	**24.42**	**0.168**	37.35	**2.65**	**24.15**	**0.165**	37.96	**2.69**

Table 1: Comparison of selected methods on both tasks and datasets. We report PSNR, LPIPS, Colorful (C) and ΔColorful (ΔC)↓ metrics.

trained in the latent space with conditional inputs y from text, using the LDM loss:

$$\mathcal{L} = \mathbb{E}_{x,y,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2\right], \quad (3)$$

where τ_θ encodes the textual prompt.

Overall, this noise-aware conditioning allows the model to accurately recolor objects while leaving unspecified areas intact, enabling robust performance even under ambiguous or minimal descriptions.

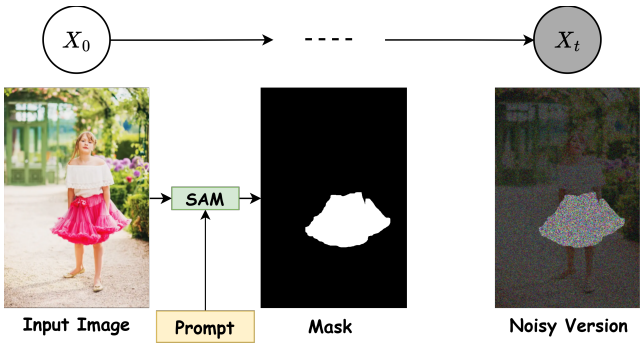

Figure 2: Language-based Noise Addition in the the forward diffusion process.

Experiments

We evaluate our approach on two public datasets for automatic and language-guided colorization: an extended COCO-Stuff dataset (Weng et al. 2022b) containing 59.3K training and 2.5K validation images, and a Multi-instance dataset (Chang et al. 2023) with 65.2K training and 7.2K validation images. Each image is paired with textual descriptions of varying detail.

The model is trained using classifier-free guidance (Ho and Salimans 2022), with 30% of full or partial descriptions randomly replaced by minimal descriptions to improve robustness. The latent space modules are trained for 50 epochs using the Adam optimizer with a learning rate of 5×10^{-5}, while the image space modules are trained separately for 20 epochs with a batch size of 8. Sampling is performed using PLMS (Liu et al. 2022) over 50 steps.

We evaluate performance using PSNR, LPIPS (Zhang et al. 2018), and the colorfulness score (Hasler and Suesstrunk 2003), where ΔColorful measures the absolute difference from the ground truth. Our method is compared with leading language-based techniques (LBIE, FILM, L-CoDe, L-CoDer, L-CoIns, L-CAD) and automatic colorization methods (CIC, InstColor, ColorFormer, Disco, ColTran, GCP, CT2). Results in Table 1 show that our language-based noise modulation achieves the highest PSNR and color accuracy while maintaining lower LPIPS and ΔColorful. Even with minimal descriptions, our model surpasses automatic methods, effectively recoloring target objects while preserving background regions, demonstrating the robustness and controllability of our approach.

Conclusion

We presented a simple yet effective **language-based noise modulation module** for text-guided image colorization. By selectively adjusting noise levels based on textual descriptions, our approach enables accurate recoloring of target regions while preserving background details. Experiments demonstrate that this targeted modulation improves color accuracy and visual fidelity, even with minimal or ambiguous descriptions. This method is lightweight, generalizable, and can be integrated into various diffusion-based image generation tasks, offering a step toward more controllable and precise text-to-image colorization.

References

Chang, Z.; Weng, S.; Li, Y.; Li, S.; and Shi, B. 2022. L-CoDer: Language-based colorization with color-object decoupling transformer. In *European Conference on Computer Vision*, 360–375. Springer.

Chang, Z.; Weng, S.; Zhang, P.; Li, Y.; Li, S.; and Shi, B. 2023. L-CoIns: Language-based colorization with instance awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19221–19230.

Chen, J.; Shen, Y.; Gao, J.; Liu, J.; and Liu, X. 2018. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8721–8729.

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

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ji, X.; Jiang, B.; Luo, D.; Tao, G.; Chu, W.; Xie, Z.; Wang, C.; and Tai, Y. 2022. Colorformer: Image colorization via color memory assisted hybrid-attention transformer. In *European Conference on Computer Vision*, 20–36. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Kumar, M.; Weissenborn, D.; and Kalchbrenner, N. 2021. Colorization transformer. *arXiv preprint arXiv:2102.04432*.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*.

Manjunatha, V.; Iyyer, M.; Boyd-Graber, J.; and Davis, L. 2018. Learning to color from language. *arXiv preprint arXiv:1804.06026*.

Su, J.-W.; Chu, H.-K.; and Huang, J.-B. 2020. Instance-aware image colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7968–7977.

Weng, S.; Sun, J.; Li, Y.; Li, S.; and Shi, B. 2022a. CT 2: Colorization transformer via color tokens. In *European Conference on Computer Vision*, 1–16. Springer.

Weng, S.; Wu, H.; Chang, Z.; Tang, J.; Li, S.; and Shi, B. 2022b. L-code: Language-based colorization using color-object decoupled conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36.

Weng, S.; Zhang, P.; Li, Y.; Li, S.; Shi, B.; et al. 2024. L-cad: Language-based colorization with any-level descriptions using diffusion priors. *Advances in Neural Information Processing Systems*, 36.

Wu, Y.; Wang, X.; Li, Y.; Zhang, H.; Zhao, X.; and Shan, Y. 2021. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14377–14386.

Xia, M.; Hu, W.; Wong, T.-T.; and Wang, J. 2022. Disentangled image colorization via global anchors. *ACM Transactions on Graphics (TOG)*, 41(6): 1–13.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 649–666. Springer.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.