2025 IEEE International Conference on Image Processing (ICIP) | 979-8-3315-2379-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICIP55913.2025.11084307

COLORGPT: AUTOMATIC COLORIZATION WITH GENERATIVE PROMPTS AND TRANSFORMER

*Juntao Qiu**, *Yaping Zhao**,†, *Edmund Y. Lam*†

Department of Electrical and Electronic Engineering, The University of Hong Kong

Fig. 1. The colorization results of our devised **ColorGPT** framework, which is fully automatic without relying on any external references or human-provided guidance. (a) The grayscale input. (b) The ground truth. (c) The photorealistic and contextually accurate colorization results generated by our method.

ABSTRACT

Automatic image colorization, where grayscale images are transformed into plausible color images, is a critical task with applications in film restoration, digital archiving, and artistic creation. Despite progress, challenges persist in achieving realistic and contextually accurate colorization due to the ambiguity in mapping grayscale intensities to colors and the need for semantic understanding. We design **ColorGPT**, a framework integrating generative language models and advanced colorization techniques to address these challenges. Our approach features three key modules: (1) a Prompt Generation Module using large language models to create diverse and context-aware color descriptions, (2) a Score Estimation Module employing CLIP to rank and select the most semantically relevant descriptions, and (3) a Colorization Module with transformer for high-quality image generation. Experiments on Set5 and Set14 demonstrate that ColorGPT outperforms state-of-the-art methods, achieving PSNR improvements of 0.7 dB on Set5 and 1.2 dB on Set14.

Index Terms— Colorization, image enhancement, image restoration, image processing, generative models

1. INTRODUCTION

Automatic image colorization, converting grayscale images to plausible color versions, has gained significant attention in computer vision. This task is crucial for film restoration, digital archiving, and artistic creation. Despite progress, achieving realistic and contextually accurate colorization remains challenging, particularly when precise understanding of image semantics and context is required.

Automatic image colorization poses a complex, ill-posed problem due to ambiguous grayscale-to-color mappings. A single grayscale image may have multiple plausible colorizations based on contextual and semantic interpretations. This challenge intensifies when generating results matching human aesthetic expectations. Developing systems that balance automation with user control remains a critical research objective.

Traditional methods [1, 2] for colorization often rely on optimization-based algorithms, which exhibit limited performance due to their inability to handle complex image semantics. With the advent of deep learning, fully automatic approaches like CT^2 [3] leverage advanced transformer architectures, showing promising results. However, these methods require extensive training on large-scale datasets and often suffer from poor generalization when applied to unseen data. Furthermore, reference-guided colorization methods, such as language-based [4, 5, 6, 7, 8, 9] or exemplar-based [10, 11, 12, 13] approaches, mitigate the limitations of fully automatic systems by providing external guidance.

*Equal contribution. †Corresponding author.

This work is supported by the Research Postgraduate Student Innovation Award (The University of Hong Kong).

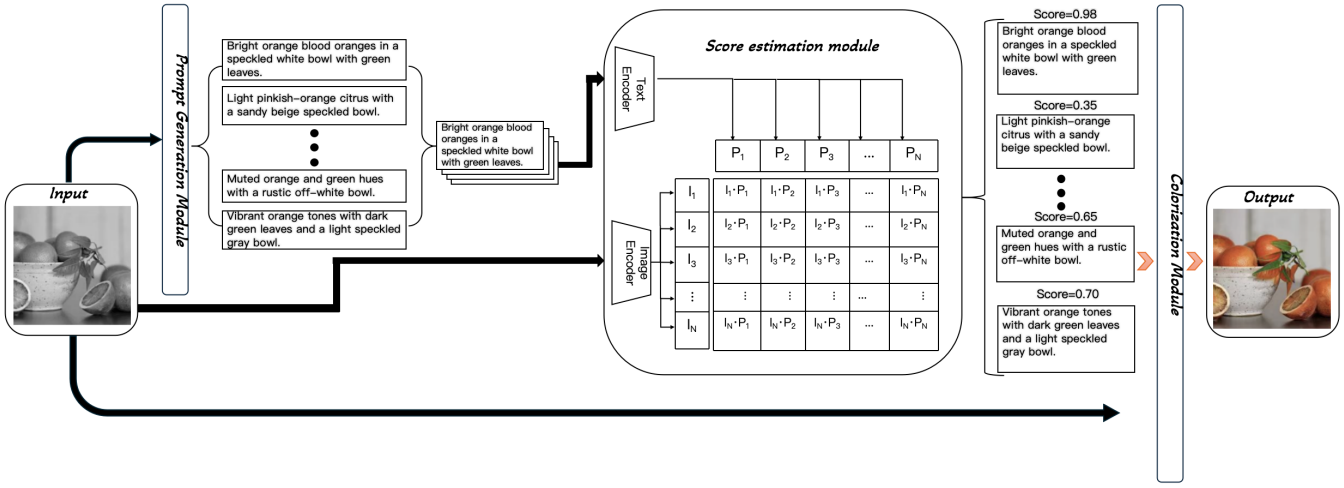

Fig. 2. The framework of **ColorGPT**, which consists of three main modules: the **Prompt Generation Module** generates descriptive prompts based on the grayscale input, the **Score Estimation Module** evaluates the compatibility of prompts with the image features, and the **Colorization Module** selects the best prompt to guide the generation of the final colorized output.

However, these methods rely heavily on additional auxiliary information or even manual inputs, which introduces labor demands.

To overcome these limitations, we propose **ColorGPT**, a novel framework for fully automated image colorization, as shown in Fig. 1 and 2. Our method utilizes a large language model (LLM) [14] to automatically generate descriptive text prompts, eliminating the need for external references or manual guidance. A fully automated scoring module, powered by CLIP [15], ensures the selection of the most contextually relevant prompts to optimize the subsequent colorization process. Finally, a transformer-based colorization module generates high-quality colorized images. The entire process is fully automated, requiring no auxiliary input, yet achieving state-of-the-art results with high fidelity and contextual accuracy on classical Set5 [16] and Set14 [17] datasets.

Our main contributions are as follows:

- We devise ColorGPT, a training-free framework for automatic colorization, integrating generative language models, contrastive learning, and transformers.
- We design a Score Estimation Module powered by CLIP to rank and select the most semantically relevant color description for the input image, which in turn improves the contextual consistency and semantic accuracy of colorization results.
- Extensive experiments on Set5 and Set14 datasets demonstrate the superiority of ColorGPT, achieving an average PSNR improvement of 0.7 dB on Set5 and 1.2 dB on Set14 over SOTA methods.

2. RELATED WORK

2.1. Language-based Colorization

Language-based image colorization techniques use natural language descriptions to assign colors to grayscale images to generate satisfactory color images. Earlier approaches such as FILM [4] incorporates linguistic descriptions through feature transformations, Chen et al [5] use spatial fusion to improve the results, L-CAD [6] uses pre-trained model capabilities and color a priori to generate high-quality results, L-CoIns [7] implements instance-aware colorization, and L-CoDe [8] and L-CoDer [9] use object-color word correspondences to guide colorization. However, these methods require external user-provided descriptions or references to guide the process, introducing significant dependency on manual inputs and reducing the potential for fully automated workflows.

2.2. Automatic Colorization

Automatic image colorization techniques aim to estimate and fill missing color channels from grayscale images without user guidance. Existing approaches are mainly explored based on deep learning techniques. The introduction of generative models such as Generative Adversarial Networks (GAN) [18, 19, 20, 21] and Variational Auto-Encoder (VAE) [22] further enriches the diversity of color generation and improves the realism of colors. In recent years, methods such as CT^2 [3] have effectively enhanced automatic colorization by combining pre-trained generative models and rich color priors. However, these approaches rely heavily on large-scale datasets for training and often fail to generalize well to unseen scenarios, highlighting a lack of robustness and adaptability in their performance.

Grayscale | Ground Truth | CT^2 (ECCV 2022) | L-CAD (NeurIPS 2023) | **ColorGPT (Ours)**

Fig. 3. Qualitative results on the Set5 [16] dataset using different methods. The images above represent ground truth, CT^2 (ECCV 2022) [3], L-CAD (NeurIPS 2023) [6], and ColorGPT (Ours), respectively.

Method	CT^2 (ECCV 2022)		L-CAD (NeurIPS 2023)		Color-Imagination (CVPR 2024)		**ColorGPT (Ours)**	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Baby	20.09	0.8943	20.17	0.9039	22.35	0.8573	23.91	0.9296
Bird	21.07	0.7531	18.49	0.7502	18.45	0.7405	22.73	0.8721
Butterfly	19.42	0.8762	16.65	0.8256	16.76	0.7984	17.58	0.8549
Head	23.03	0.8978	24.68	0.8539	24.59	0.8417	22.43	0.8809
Woman	26.84	0.9467	25.00	0.9031	25.35	0.8818	27.29	0.9336
Average	22.09	0.8736	21.00	0.8473	21.25	0.8239	**22.79**	**0.8942**

Table 1. Quantitative comparison of different methods on the Set5 [16] dataset. The table presents the results of CT^2 [3], L-CAD [6], Color-Imagination [23], and ColorGPT [14] on individual images and their averaged results. Metrics PSNR and SSIM (higher is better) are used for evaluation. The best results are highlighted in bold.

3. METHOD

Given a grayscale image as input, our proposed framework, **ColorGPT**, performs SOTA automatic colorization using a three-stage process: (1) Prompt Generation using an LLM, (2) Score Estimation using CLIP, and (3) Colorization using a language-based colorization model. The following sections describe these modules in detail.

3.1. Prompt Generation Module

The first step in our pipeline is to generate color descriptions for a given grayscale image. We leverage a LLM, specifically ChatGPT [14], to produce a set of candidate color descriptions:

$$\mathcal{T} = \{T_1, T_2, ..., T_k\}, \tag{1}$$

where $\mathcal{T}$ is the set of k different color descriptions, and k is a hyperparameter determined via ablation study. Each prompt describes plausible colorization styles and semantics for the given grayscale image.

3.2. Score Estimation Module

Once the prompts are generated, we utilize the CLIP model [15] to evaluate and rank the candidate color descriptions based on their semantic relevance to the grayscale image. The CLIP model consists of an image encoder f_G and a text encoder f_T, both mapping their respective inputs into a shared embedding space:

$$f_G(G) = I, \tag{2}$$

$$f_T(T) = P, \tag{3}$$

where G is the grayscale image, T is the textual description, I is the image embedding, and P is the text embedding.

The similarity score between an image G and a textual prompt T is computed as:

$$S(G, T) = \frac{f_G(G) \cdot f_T(T)}{\|f_G(G)\| \|f_T(T)\|} = \frac{I \cdot P}{\|I\| \|P\|}, \tag{4}$$

where $\cdot$ denotes the dot product and $\| \cdot \|$ represents the L_2-norm. The final selected prompt is determined by:

$$T^* = \arg\max_{T_i \in \mathcal{T}} S(G, T_i). \tag{5}$$

3.3. Colorization Module

The final stage is the colorization process, where the best-scoring prompt T^* is used to guide the generation of a colorized image. We employ a transformer L-CoDer [9], a language-based colorization model that takes both the grayscale image and the textual description as input:

$$C = \mathcal{F}(G, T^*), \tag{6}$$

where $\mathcal{F}$ represents the L-CoDer model, and C is the resulting colorized image.

Grayscale | Ground Truth | CT^2 (ECCV 2022) | L-CAD (NeurIPS 2023) | **ColorGPT (Ours)**

Fig. 4. Qualitative results on the Set14 [17] dataset using different methods. The images above represent ground truth, CT^2 (ECCV 2022) [3], L-CAD (NeurIPS 2023) [6], and ColorGPT (Ours), respectively.

Method	CT^2 (ECCV 2022)		L-CAD (NeurIPS 2023)		Color-Imagination (CVPR 2024)		**ColorGPT (Ours)**	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Baboon	14.37	0.7554	12.15	0.6887	15.73	0.6416	16.61	0.7396
Barbara	20.27	0.8896	20.87	0.8668	21.12	0.8402	22.24	0.8733
Coastguard	23.95	0.9652	32.51	0.9692	32.26	0.9631	32.40	0.9787
Comic	19.10	0.8690	19.23	0.8298	20.02	0.8100	21.45	0.8885
Flowers	15.27	0.6975	17.11	0.7902	17.65	0.7792	17.67	0.8097
Foreman	27.27	0.9658	27.95	0.9655	27.95	0.9622	28.92	0.9664
Lenna	20.08	0.8947	16.94	0.8692	16.89	0.8581	17.50	0.8672
Monarch	19.52	0.8776	16.73	0.8312	16.76	0.7984	17.74	0.8653
Pepper	15.95	0.6332	15.54	0.6352	19.28	0.7626	18.95	0.7314
PPT3	16.45	0.8784	12.93	0.8499	13.06	0.8391	18.20	0.8923
Head	23.03	0.8978	24.68	0.8539	24.59	0.8417	22.43	0.8809
Zebra	25.60	0.9506	27.16	0.9321	27.01	0.9311	24.58	0.9387
Average	20.07	0.8562	20.32	0.8401	21.02	0.7162	**21.56**	**0.8693**

Table 2. Quantitative comparison of different methods on the Set14 [17] dataset. The table presents the results of various methods (CT^2 [3], L-CAD [6], Color-Imagination [23], and ColorGPT [14]) on individual images (e.g., Baboon, Barbara, etc.) and their averaged results. The metrics PSNR and SSIM are used to evaluate performance, where higher values indicate better performance.

3.4. End-to-End Pipeline

The overall pipeline is summarized as follows:

1. Generate k candidate color descriptions using ChatGPT, as shown on the left of Fig. 2.
2. Compute CLIP similarity scores and select the best description P^*, as shown in the center of Fig. 2.
3. Perform language-based colorization using L-CoDer to colorize the image, as shown on the right of Fig. 2.

This fully automatic framework enables high-quality colorization by leveraging generative language models, multimodal contrastive learning, and advanced deep learning-based colorization techniques.

4. EXPERIMENT

4.1. Experimental Setting

Dataset. We evaluate on standard benchmarks Set5 [16] and Set14 [17], converting RGB images to grayscale for colorization (originals as ground truth). Two inherently grayscale images in Set14 (`bridge`, `man`) are excluded due to absence of color references.

Metric. We adopt peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) for quantitative evaluation.

Baseline. We compare our method with three SOTA approaches: CT² [3], L-CAD [6], and Color-Imagination [23]. For fairness, our method, L-CAD, and Color-Imagination employ the optimal text prompts determined by our scoring module.

4.2. Quantitative and Qualitative Comparisons

Table 1 and Table 2 present the quantitative comparison results of our proposed method with CT² [3], L-CAD [6] and Color-Imagination [23] on the Set5 [16] and Set14 [17] datasets. Our method performs well on both datasets, achieving the highest PSNR and SSIM values of 22.79 dB and 0.8942 on the Set5 dataset, and 21.56 dB and 0.8693 on the Set14 dataset, respectively. This fully demonstrates that our

Grayscale | Ground Truth | $k = 1$ | $k = 2$ | $k = \mathbf{3}$ | $k = 5$

Fig. 5. Demonstration of the effect of varying the number of statements, k, on the colorization results of the image `foreman` of the Set14 dataset. As k increases, the image quality improves up to $k = 3$, closely resembling the Ground Truth. However, for $k > 3$, the quality begins to degrade slightly.

Varying k	Set5		Set14	
	Average PSNR	Average SSIM	Average PSNR	Average SSIM
k = 1	21.89	0.8689	21.03	0.8649
k = 2	22.16	0.8756	21.43	0.8650
k = 3	**22.79**	**0.8942**	**21.56**	**0.8693**
k = 5	21.32	0.8716	20.62	0.8576

Table 3. The impact of different types of statements, k, on the average PSNR and SSIM for the Set5 [16] and Set14 [17] datasets. It is observed that the model achieves the best performance when $k = 3$, with the highest PSNR and SSIM values.

method outperforms existing methods in terms of both pixel similarity and structure preservation.

Table 4 further validates the computational efficiency advantage of our approach, as Color-Imagination requires generating multiple reference images whereas our method only generates several text prompts, achieving significant speed improvements with lightweight processing.

As shown in Figs. 3 and 4, qualitatively, our method is clearly superior to existing methods. On the Set5 dataset, for bird images, our method successfully preserves the true colors and fine details, while the CT² and L-CAD methods are significantly inferior in terms of color consistency and texture preservation. Notably, both CT² and L-CAD incorrectly color the bird beak as green, mistaking it for leaves, while ColorGPT accurately restores the color of the beak.

On the Set14 dataset, facing monarch images with complex textures, our method demonstrated higher color accuracy and stronger detail reproduction. Notably, both CT² and L-CAD fail to accurately colorize the butterfly, resulting in unrealistic hues, while our method achieves more vivid and realistic colors, correctly identifying and colorizing the butterfly.

Both in quantitative and qualitative evaluations, our method demonstrates excellent performance, consistently achieving higher fidelity in color restoration and finer detail preservation compared to existing approaches.

4.3. Ablation Study

In order to determine the optimal number of cue candidates k, we performed ablation experiments to systematically evaluate the effect of different k values on the colorization performance. Specifically, we adjusted the k value and measured the corresponding color image quality.

Method	Color-Imagination (CVPR 2024)	ColorGPT
Time	35.9s	7.5s

Table 4. Running Time Comparison

The experimental results are shown in Table 3 and Fig. 5. When $k = 3$, the model achieves the best performance in evaluation metrics such as PSNR and SSIM, and also significantly outperforms the other k values in terms of color consistency and detail retention. This suggests that k is able to strike an optimal balance between the diversity of cue generation and computational efficiency, resulting in optimal colorization. This highlights the importance of carefully determining the k value, as it has a direct impact on the ability of the model to deliver both visually appealing and computationally efficient colorization results.

5. CONCLUSION

In this paper, we propose an automatic image colorization framework, which combines generative language models and transformer techniques to achieve image colorization through linguistic cues, semantic scoring, and colorization models. Experiments show that the method outperforms existing approaches on Set5 and Set14 datasets and effectively enhances the image colorization results.

6. REFERENCES

[1] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf, "Automatic image colorization via multimodal predictions," in *European Conference on Computer Vision, Part III 10*. Springer, 2008, pp. 126–139.

[2] Anat Levin, Dani Lischinski, and Yair Weiss, "Colorization using optimization," in *ACM SIGGRAPH*. 2004.

[3] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi, "CT^2: Colorization transformer via color tokens," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–16.

[4] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis, "Learning to color from language," *arXiv preprint arXiv:1804.06026*, 2018.

[5] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu, "Language-based image editing with recurrent attentive models," in *IEEE conference on computer vision and pattern recognition*, 2018.

[6] 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*, vol. 36, 2023.

[7] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi, "L-CoIns: Language-based colorization with instance awareness," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19221–19230.

[8] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi, "L-CoDe: Language-based colorization using color-object decoupled conditions," in *AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 2677–2684.

[9] 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*, 2022, pp. 360–375.

[10] Yaping Zhao, Haitian Zheng, Mengqi Ji, and Ruqi Huang, "Cross-camera deep colorization," in *CAAI International Conference on Artificial Intelligence*. Springer, 2022, pp. 3–17.

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

[12] Yaping Zhao, Haitian Zheng, Jiebo Luo, and Edmund Y Lam, "Improving video colorization by test-time tuning," in *2023 IEEE International Conference on Image Processing*. IEEE, 2023, pp. 166–170.

[13] Yaping Zhao and Edmund Y Lam, "LUCK: Lighting up colors in the dark," in *IEEE Optoelectronics Global Conference*, 2024, pp. 51–55.

[14] OpenAI, "ChatGPT: An ai language model," Online, 2023, Accessed:2025-01-01, URL: https://chat.openai.com.

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

[16] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *British Machine Vision Conference*, 2012.

[17] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, 2010.

[18] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu, "Unsupervised diverse colorization via generative adversarial networks," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD, Part I 10*. Springer, 2017, pp. 151–166.

[19] Zhitong Huang, Nanxuan Zhao, and Jing Liao, "Unicolor: A unified framework for multi-modal colorization with transformer," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 1–16, 2022.

[20] Patricia Vitoria, Lara Raad, and Coloma Ballester, "ChromaGAN: Adversarial picture colorization with semantic class distribution," in *IEEE/CVF winter conference on applications of computer vision*, 2020.

[21] 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*. Springer, 2022, pp. 271–288.

[22] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth, "Learning diverse image colorization," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 6837–6845.

[23] 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*, 2024, pp. 2609–2619.