Supplmentary Material:
L-CoIns: Language-based Colorization with Instance Awareness

Zheng Chang[#1] Shuchen Weng[#2,3] Peixuan Zhang[1] Yu Li[4] Si Li[*1] Boxin Shi[2,3]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications

[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[4]International Digital Economy Academy

{zhengchang98,pxzhang,lisi}@bupt.edu.cn, {shuchenweng,shiboxin}@pku.edu.cn, liyu@idea.edu.cn

7. Appendix

7.1. Parameter Setting

We explore three different parameters settings for the grouping transformer, including the group token number N_{G}, the grouping block number L, the channel number C_{z}, the attention head number N_{H}, and the parameter number, as shown in Tab. 3. We further show the quantitative results for each setting in Tab. 4, which demonstrate that the performance improves as the parameter number increases. In the main paper, we report the results of L-CoIns (Large).

To further investigate the effect of the number of group tokens, we change N_{G} while maintaining other parameters of L-CoIns (Large). As shown in Tab. 5, increasing N_{G} leads to improve qualitative performance. Since more than 80 group tokens could only provide minor improvements, we select 80 as N_{G}.

Table 3. Different parameter settings.

Model	N_{G}	L	C_z	N_{H}	Params
L-CoIns (Small)	20	4	768	12	47M
L-CoIns (Base)	40	8	768	12	75M
L-CoIns (Large)	80	12	1024	16	177M

Table 4. Quantitative experiment results of different parameter settings. Throughout the paper, ↑ (↓) means higher (lower) is better. Best performances are highlighted in **bold**.

Method	Extended COCO-Stuff			Multi-instance		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
L-CoIns (Small)	25.280	0.91287	0.167	24.389	0.90238	0.175
L-CoIns (Base)	25.409	0.91405	0.164	24.574	0.91105	0.165
L-CoIns (Large)	**25.511**	**0.92104**	**0.157**	**24.823**	**0.91717**	**0.162**

# Equal contributions. * Corresponding author.

Table 5. Quantitative experiment about numbers of group tokens.

N_{G}	Extended COCO-Stuff			Multi-instance		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
20	25.314	0.91723	0.169	24.544	0.91077	0.170
40	25.447	0.91957	0.164	24.677	0.91413	0.165
80	25.511	0.92104	**0.157**	24.823	0.91717	0.162
100	**25.539**	**0.92116**	**0.157**	**24.855**	**0.91809**	**0.160**

7.2. Visualization of Statistical Correlation

We visualize the statistical correlation between luminance and colors by drawing a bar chart that presents the proportion of three typical colors (*i.e.*, red, green, and blue) in different luminance intervals. Specifically, we randomly select 10000 images from the training set and convert them into HSV color space. After defining the ranges of red, green, and blue colors in Tab. 6, we calculate the pixel number of each color belonging to different luminance intervals. As shown in Fig. 7 top, colors and luminance are statistically correlated. To break down this statistical correlation and drive the model towards understanding language descriptions, we propose the luminance augmentation. After performing this strategy, we redraw the bar chart with the augmented luminance and show it in Fig. 7 bottom, which demonstrates independence between luminance and colors. We show more augmented grayscale images in Fig. 8.

Table 6. Division ranges of typical colors.

	red	green	blue
hue	$[0^\circ, 20^\circ] \cup [340^\circ, 360^\circ]$	$[100^\circ, 140^\circ]$	$[220^\circ, 260^\circ]$
saturation	$[50, 255]$	$[50, 255]$	$[50, 255]$
brightness	$[50, 255]$	$[50, 255]$	$[50, 255]$

7.3. Visualization of Grouping Results

To demonstrate the effectiveness of the grouping transformer that aggregates similar image patches for correctly identifying corresponding regions to be colorized, we visualize the grouping results in Fig. 9.

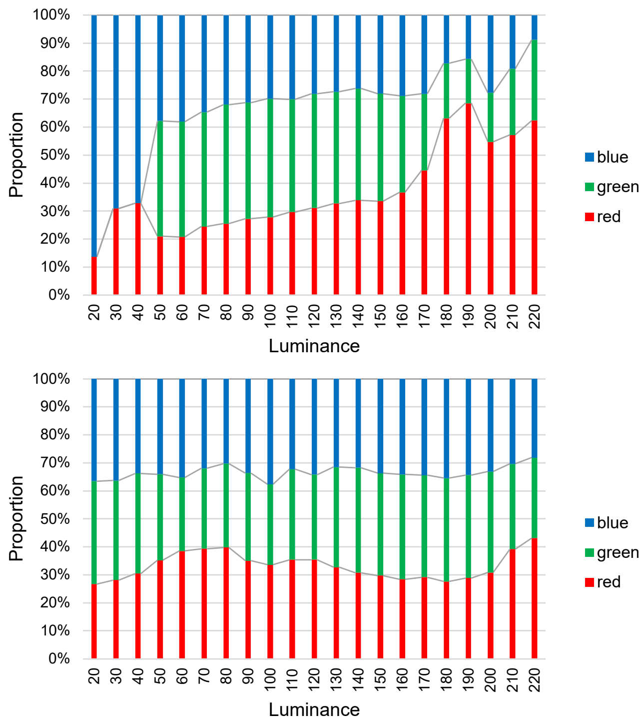


Figure 7. Visualization of statistical correlation. **Top:** Before performing the luminance augmentation, with the luminance increasing, the dominant color gradually changes from blue to green, and then red. **Bottom:** After performing the luminance augmentation, all colors have almost the same proportion regardless of luminance. In this figure, we only consider correlations with luminances between 20 and 220, since brighter and deeper luminances are often perceived as white and black, respectively.

7.4. Comparisons with Automatic Methods

We make additional comparisons with existing automatic colorization methods (*e.g.*, CIC [11], ChromaGAN [6], InstColor [5], and CT^2 [7]) to demonstrate the advantage of the language condition as supervisory signal of colorization task. The additional quantitative and qualitative comparisons with automatic colorization methods are shown in Tab. 7 and Fig. 10. With the provided language description, our method could better colorize the specified instance according to the preference of the user.

Table 7. Quantitative comparisons with automatic colorization methods.

Method	Extended COCO-Stuff			Multi-instance		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
CIC [11]	22.156	0.89705	0.224	22.219	0.89623	0.222
ChromaGAN [6]	22.085	0.84161	0.275	22.411	0.85848	0.248
InstColor [5]	23.914	0.90618	0.194	22.661	0.89838	0.218
CT^2 [7]	24.217	0.89612	0.187	23.041	0.90257	0.195
Ours	**25.511**	**0.92104**	**0.157**	**24.823**	**0.91717**	**0.162**

7.5. Additional Comparison Results

We show additional qualitative and quantitative comparison results with state-of-the-art language-based colorization methods, *e.g.*, LBIE [2], ML2018 [4], Xie2018 [9], L-CoDe [8] and L-CoDer [1]. In Fig. 11, we present more qualitative comparison results to demonstrate the advantages of our method for the four typical language descriptions, as illustrated in Sec. 5.1 of the main paper. In Tab. 8, we show two more quantitative metrics to measure the distance between the generated images and original images (Fréchet inception distance, FID [3]) and whether colorized images are well conditioned on the given language condition (R-precision [10]). As the table shows, our method performs best on both metrics.

Table 8. More quantitative comparisons with language-based methods.

Method	Extended COCO-Stuff		Multi-instance	
	FID↓	R-precision↑	FID↓	R-precision↑
LBIE [2]	32.594	42.276%	27.373	33.571%
ML2018 [4]	33.908	43.443%	29.831	33.214%
Xie2018 [9]	33.137	41.954%	27.796	32.582%
L-CoDe [8]	30.718	44.046%	26.993	34.995%
L-CoDer [1]	30.097	47.103%	27.280	35.769%
Ours	**29.506**	**48.154%**	**25.151**	**36.605%**

7.6. Additional Ablation Results

We present more ablation results in Fig. 12 to study the impact of our proposed modules. The ablation details are described in Sec. 5.3 of the main paper.

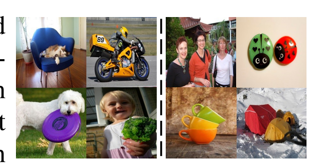
7.7. Additional Application Results

We present diverse results with various language descriptions to demonstrate the controllability of our method in Fig. 13. Moreover, we demonstrate our generalization capability by showing colorization results on legacy black-and-white photos, as shown in Fig. 14.

7.8. Failure cases

As illustrated in limitation (Sec. 6 of the main paper), our model still has difficulty capturing regions of small objects corresponding to color words in a long caption containing detailed information. Failure cases are shown in Fig. 15.

7.9. Necessity of building multi-instance dataset.



Although existing extended COCO-Stuff dataset [42] provides various scenarios with abundant object categories (left image), it lacks samples with distinctive visual characteristics and detailed language descriptions for multiple instances in image (right image). Therefore, we build the new dataset with these miscellaneous cases to train the model to learn inter-instance relationships and assign distinct colors to each instance.

References

[1] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. L-CoDer: Language-based colorization with color-object decoupling transformer. In *ECCV*, 2022. 2

[2] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *CVPR*, 2018. 2

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a nash equilibrium. In *NIPS*, 2017. 2

[4] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. Learning to color from language. In *NAACL*, 2018. 2

[5] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *CVPR*, 2020. 2

[6] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chroma-GAN: Adversarial picture colorization with semantic class distribution. In *WACV*, 2020. 2

[7] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. CT^2: Colorization transformer via color tokens. In *ECCV*, 2022. 2

[8] Shuchen Weng, Hao Wu, Zheng Chang Chang, Jiajun Tang, Si Li, and Boxin Shi. L-CoDe: Language-based colorization using color-object decoupled conditions. In *AAAI*, 2022. 2

[9] Yanping Xie. Language-guided image colorization. Master's thesis, ETH Zurich, Departement of Computer Science, 2018. 2

[10] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 2

[11] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

Figure 8. More examples of luminance augmentation. **Top left:** Enlarging the relative luminance. **Top right:** Reversing the relative luminance. **Bottom left:** Increasing the global luminance. **Bottom right:** Decreasing the global luminance.

Figure 9. Visualization of grouping results. Image patches assigned to the same group are represented by the same color.

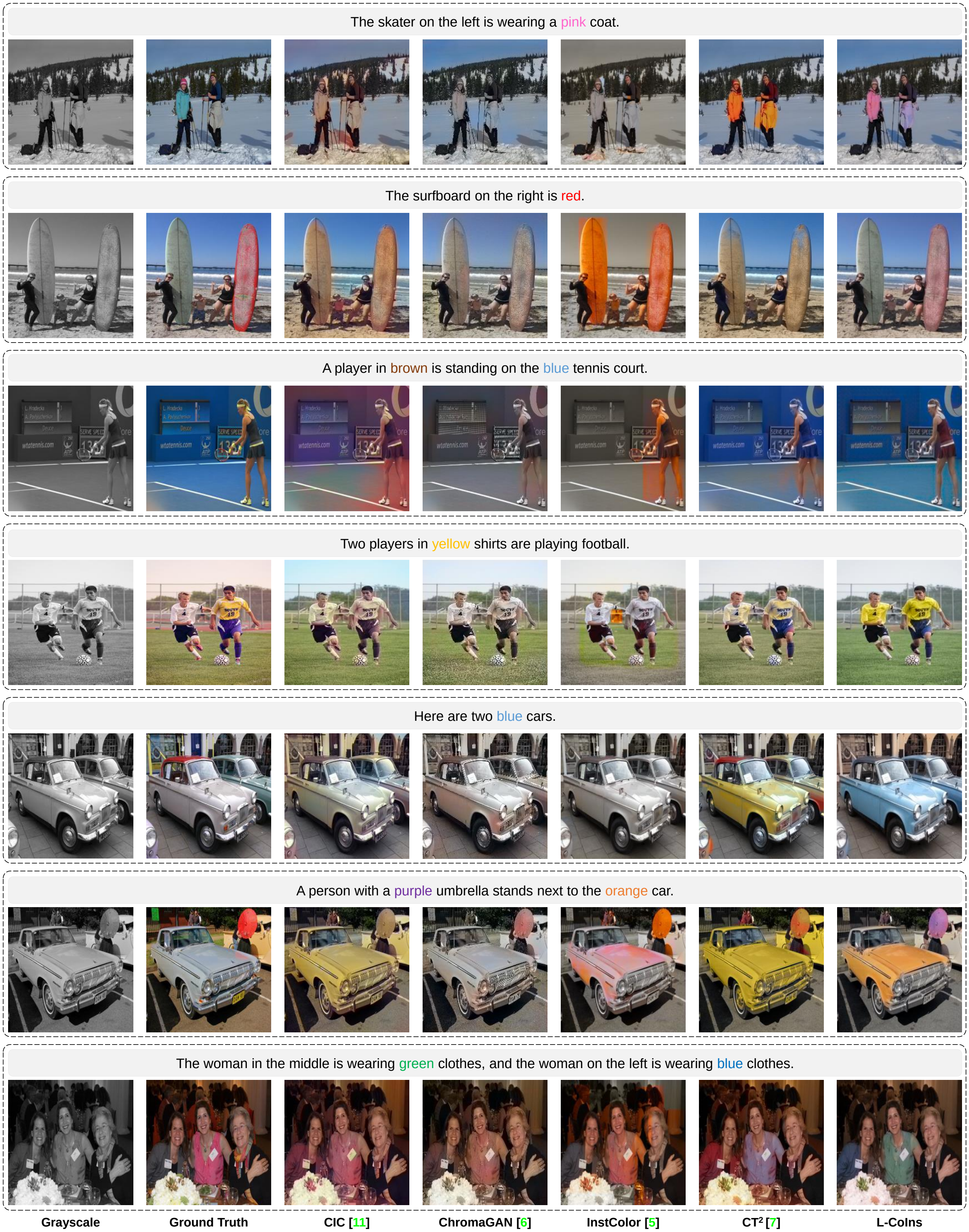


Figure 10. Comparisons with automatic colorization methods. With language descriptions, our model meets the specific requests of users.

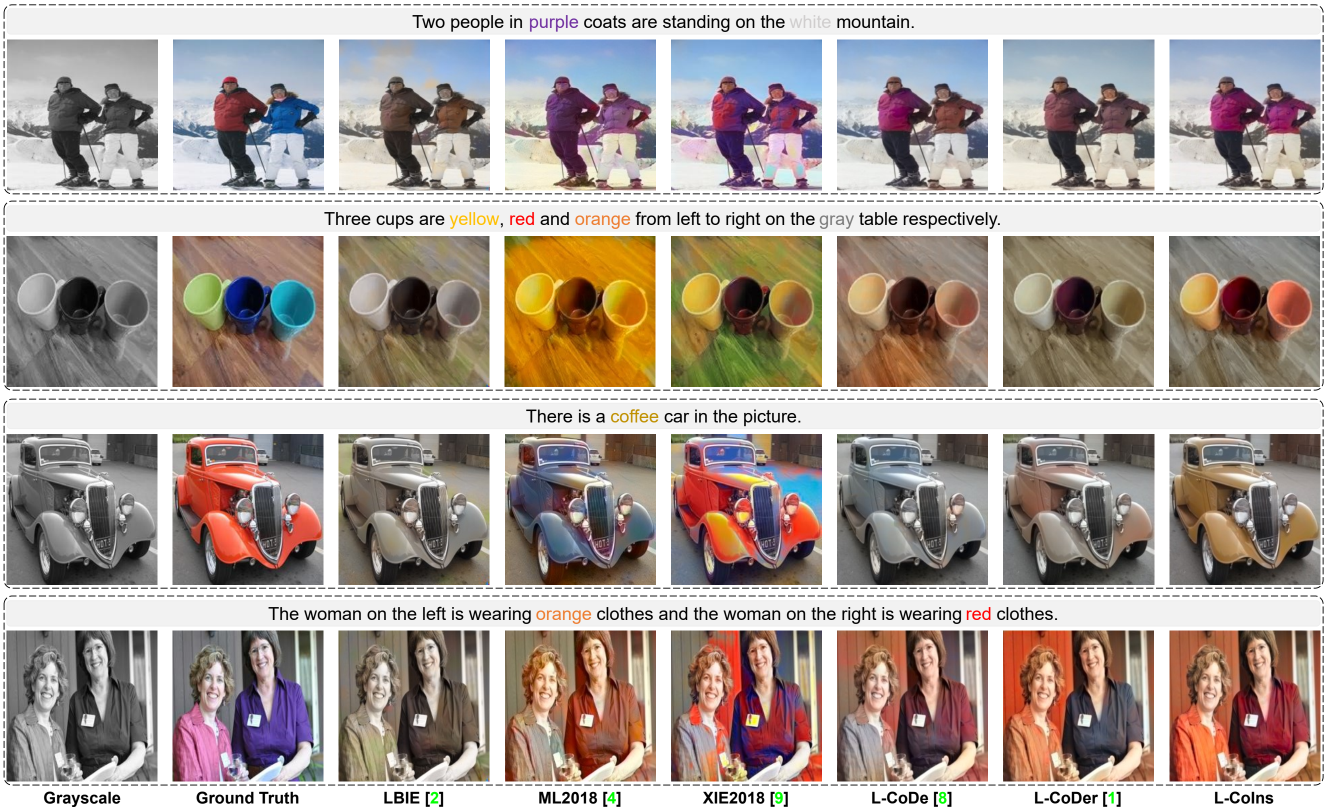


Figure 11. Comparison with language-based colorization. **First row:** Our method correctly colorizes all corresponding regions (two purple coats).**Second row:** Our method assigns the distinct color to each corresponding instance (right orange cup) **Third row:** Our method exactly understands the unobserved correspondence (coffee car) **Fourth row:** Our method shows robustness for the luminance (red colorizes the woman's region that has an extremely dark luminance)

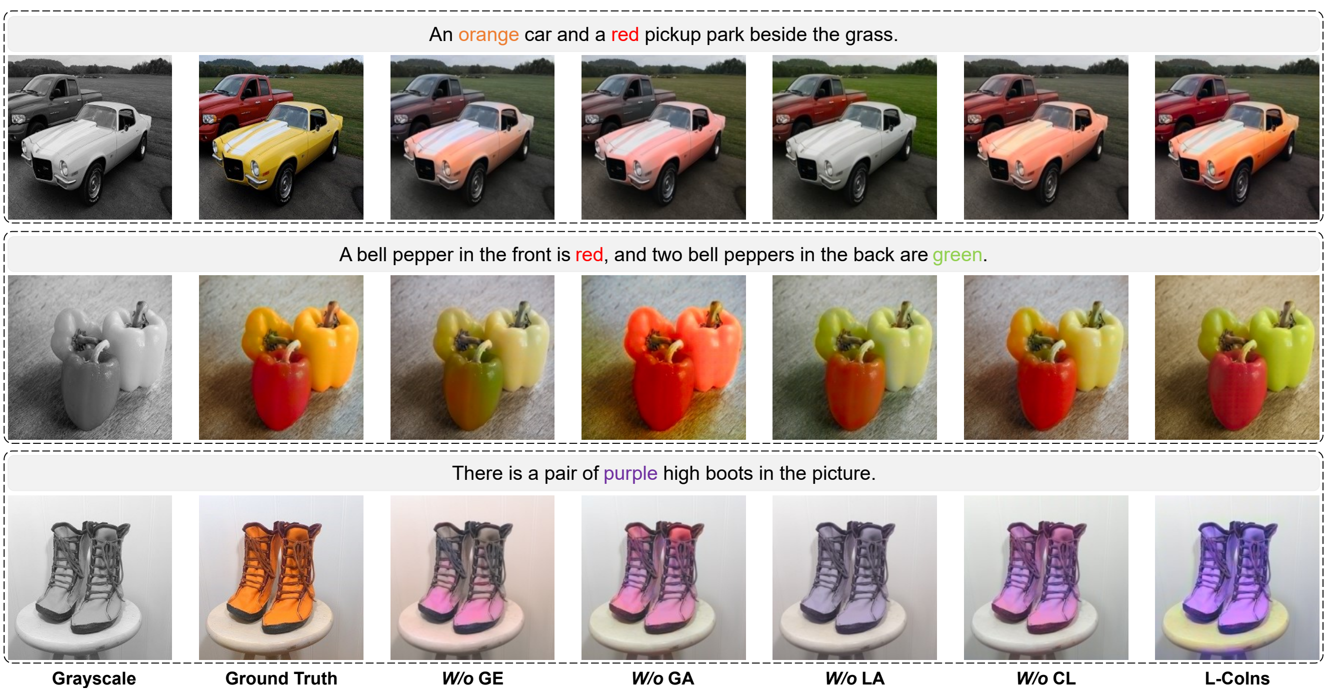


Figure 12. Additional ablation results. Disabling some parts of our proposed modules degrades the colorization quality.

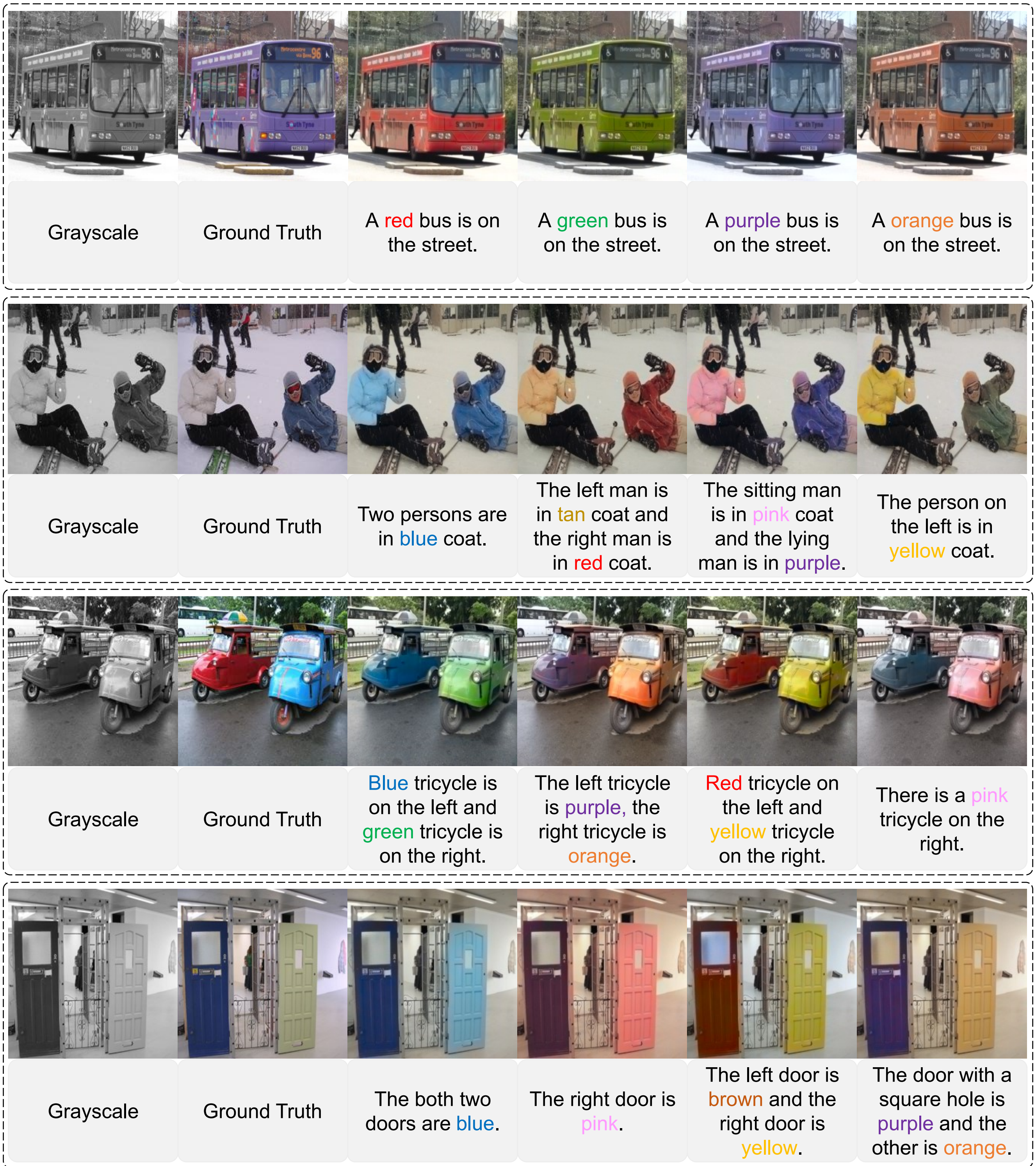


Figure 13. Colorization results under the guidance of various language descriptions.

Figure 14. More colorization results of legacy black-and-white photos.

Figure 15. Failure cases of our method. **Left:** Our model has difficulty identifying all skirt regions and recognizing the person in the pig costume. **Middle:** It is difficult to determine which girl is the tallest. **Right:** It is difficult to locate all of the people who are coming in.