

# Supplementary Materials for "COCO-LC: Colorfulness controllable language-based colorization"

Anonymous Authors

## 1 IMPLEMENTATION DETAILS

We implement our COCO-LC on Stable Diffusion v1.5 [7]. We adopt DDIM-50 sampler during inference to reduce time complexity. We can generate colorized results in 5 seconds on a single RTX 4090 with batch size=1 and float16 precision.

In order to fairly compare the FID scores of L-CoDe [8], L-CoDer [1], L-CoIns [3], L-CAD [2], we upsampled the RGB images generated by each method to 512 resolution, converted them to the LAB color space, concatenate the *ab* channel with the *L* channel of the gray image, and finally converted them back to the RGB color space to get higher resolution results. Compared with not performing the above operations, the FID score of those methods have been improved to a certain extent. As Unicolor [5] can generate results with the same resolution as the original image, so we directly resize it to 512 resolution for fair comparison.

## 2 ABLATION STUDY OF COCO-DECODER

We present our COlorfulness COntrollable Decoder (COCO-Decoder) to generate diverse colorized results with different color style varying from fantastic, realistic to vintage. However, a straightforward method is change the saturation of images in HSL color space. Thus, we conduct an ablation study to demonstrate that our method can generate more plausible and flexible results compared with adjusting saturation directly. We convert colorized results into HSL color space and modify the saturation by a linear scaling factor  $k$  to generate modified results with approximately same colorfulness with our results. The bigger  $k$  is, the lower saturation. We test our COCO-Decoder with different  $\alpha \in \{1.0, 0.9, 0.8, 0.7, 0.6\}$ . As shown at Fig.1 pointed by a red arrow, the colors of blue shirt will change to white, leading to inconsistency with text description. Also, red pots on the hat of the man are reduced by our COCO-Decoder gradually to maintain a harmonious global color style, but simply decrease saturation will keep those red spots even the image becomes gray.

## 3 MORE QUALITATIVE COMPARISON

As shown at Fig.2, 3, 4, we provide more qualitative results and comparison with previous language-based methods: UniColor [5], L-CoDe [8], L-CoDer [1], L-CoIns [3], L-CAD [2]. It can be seen that our method achieves more natural and plausible results in diverse scenarios, and produces more semantically consistent colors for a variety of objects. We also provide a clearer and comprehensive file (supp\_figures.pdf) to shown our results in advance.

## 4 DETAILS OF SPATIAL SEMANTIC HIGH-LEVEL CONDITION

In our main text, we present a multi-level condition to reduce color overflow. As for high-level condition, we utilize Mask2Former [4] to extract semantic maps of grayscale images and use CLIP [6] to convert text labels into a semantic feature. We modify some of

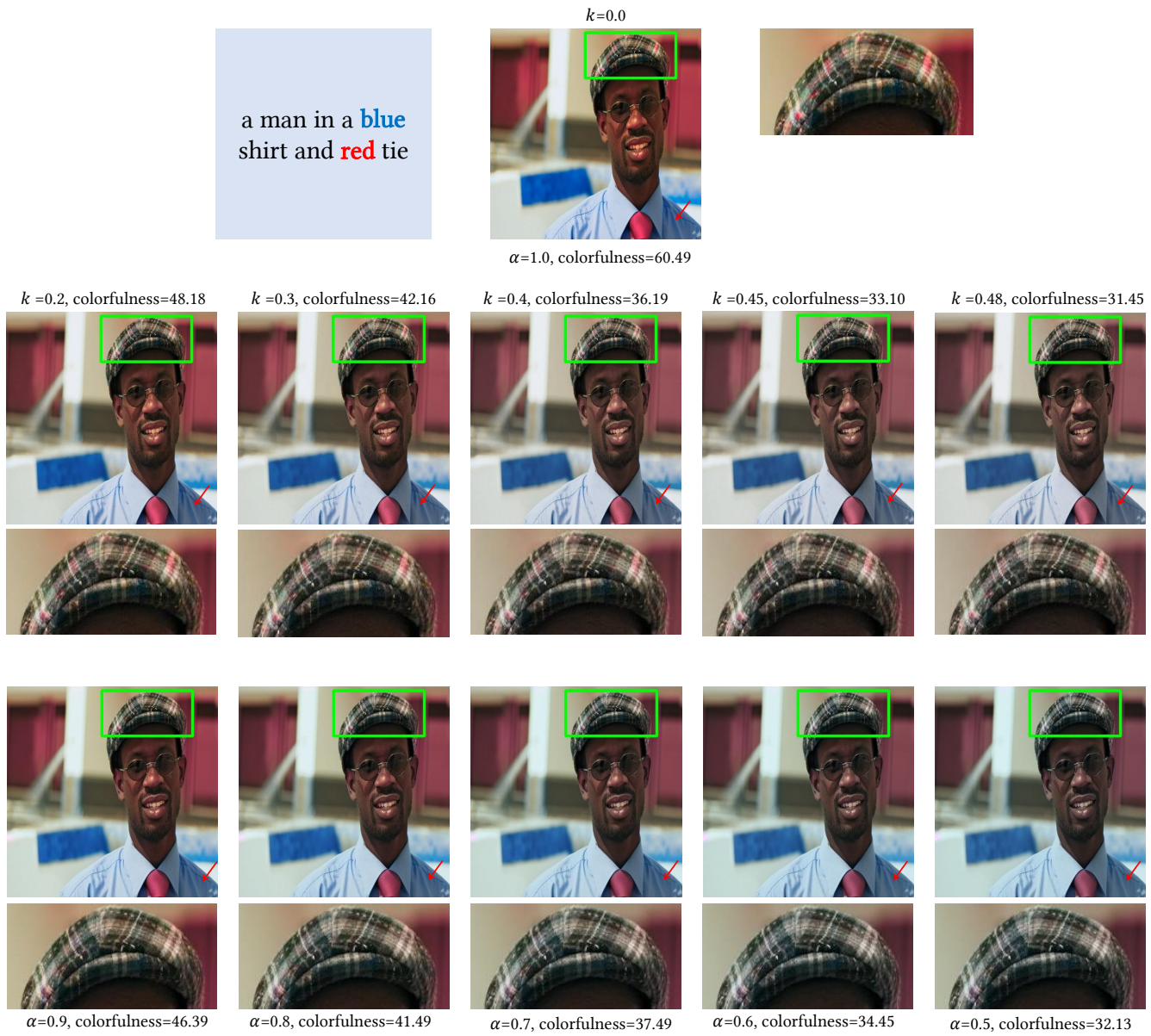
those labels into natural language, in order to insert more accurate semantic condition. All labels we use are showed as below:

'person', 'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train', 'truck', 'boat', 'traffic light', 'fire hydrant', 'stop sign', 'parking meter', 'bench', 'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra', 'giraffe', 'backpack', 'umbrella', 'handbag', 'tie', 'suitcase', 'frisbee', 'skis', 'snowboard', 'sports ball', 'kite', 'baseball bat', 'baseball glove', 'skateboard', 'surfboard', 'tennis racket', 'bottle', 'wine glass', 'cup', 'fork', 'knife', 'spoon', 'bowl', 'banana', 'apple', 'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut', 'cake', 'chair', 'couch', 'potted plant', 'bed', 'dining table', 'toilet', 'tv', 'laptop', 'mouse', 'remote', 'keyboard', 'cell phone', 'microwave', 'oven', 'toaster', 'sink', 'refrigerator', 'book', 'clock', 'vase', 'scissors', 'teddy bear', 'hair drier', 'toothbrush', 'banner', 'blanket', 'bridge', 'cardboard', 'counter', 'curtain', 'door', 'wood floor', 'flower', 'fruit', 'gravel', 'house', 'light', 'mirror', 'net', 'pillow', 'platform', 'playing-field', 'railroad', 'river', 'road', 'roof', 'sand', 'sea', 'shelf', 'snow', 'stairs', 'tent', 'towel', 'brick wall', 'stone wall', 'tile wall', 'wood wall', 'water', 'closed window', 'window', 'tree', 'fence', 'ceiling', 'sky', 'cabinet', 'table', 'floor', 'pavement', 'mountain', 'grass', 'dirt', 'paper', 'food', 'building', 'rock', 'wall', 'rug'.

Although above labels cannot cover all class in natural world, we find it is enough to guide colorization as high-level compensation during diffusion process.

## REFERENCES

- [1] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. 2022. L-CoDer: Language-based Colorization with Color-object Decoupling Transformer. In *Proc. European Conf. Computer Vision*.
- [2] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. 2023. L-CAD: Language-based Colorization with Any-level Descriptions. *arXiv preprint arXiv:2305.15217* (2023).
- [3] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. 2023. L-CoIns: Language-based Colorization with Instance Awareness. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.
- [4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.
- [5] Zhitong Huang, Nanxuan Zhao, and Jing Liao. 2022. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics* (2022).
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. IEEE Int'l Conf. Machine Learning*. PMLR.
- [7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.
- [8] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. 2022. L-CoDe: Language-based colorization using color-object decoupled conditions. In *Proc. AAAI Conf. Artificial Intelligence*.



**Figure 1: Ablation results of COCO-Decoder. Compared with grayish and unreasonable results generated by changing saturation directly, our method can produce more plausible and flexible results with different scaling factor. Zoom in for better visualization.**



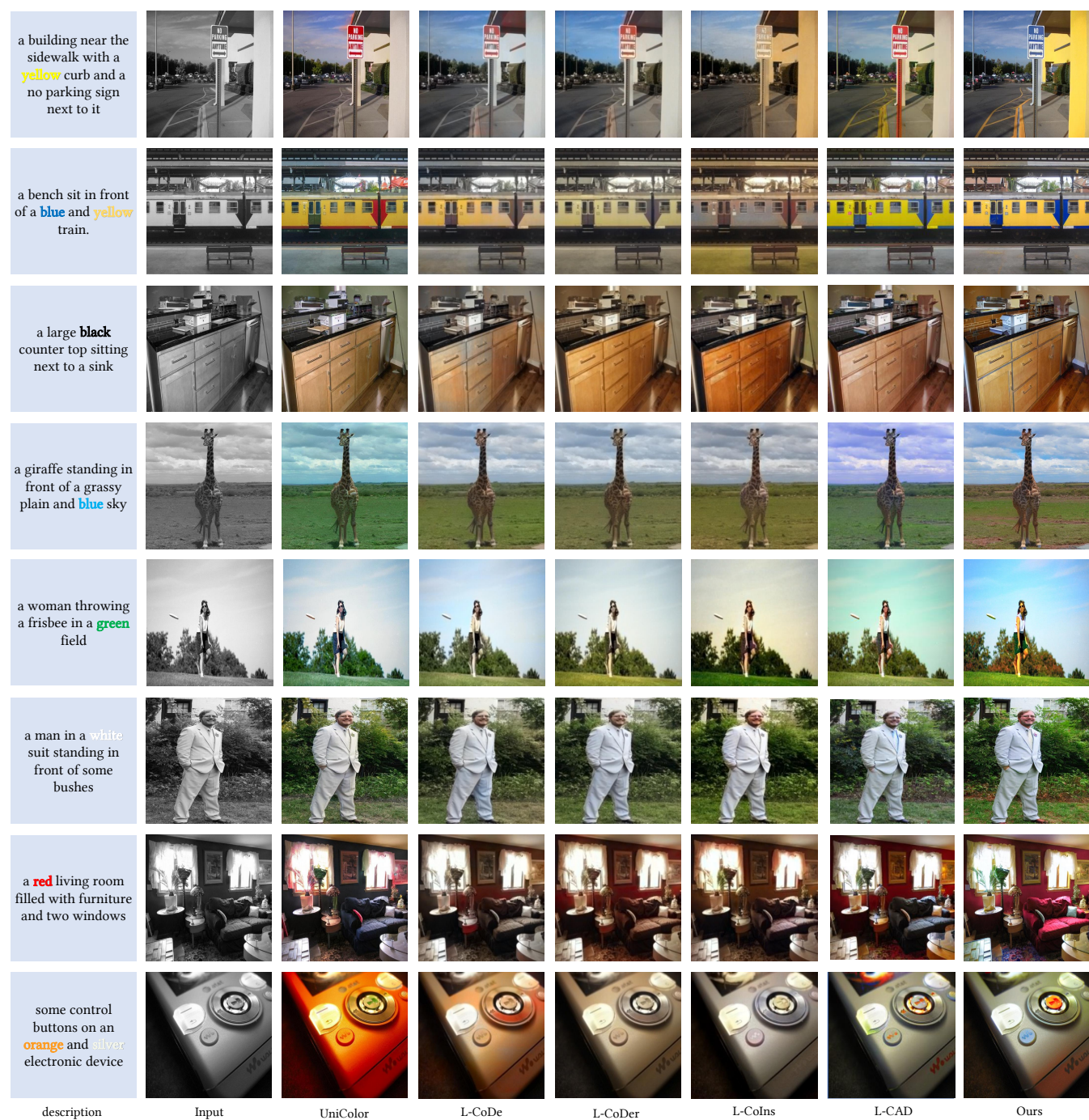


Figure 2: Qualitative results compared with previous language-based colorization methods.



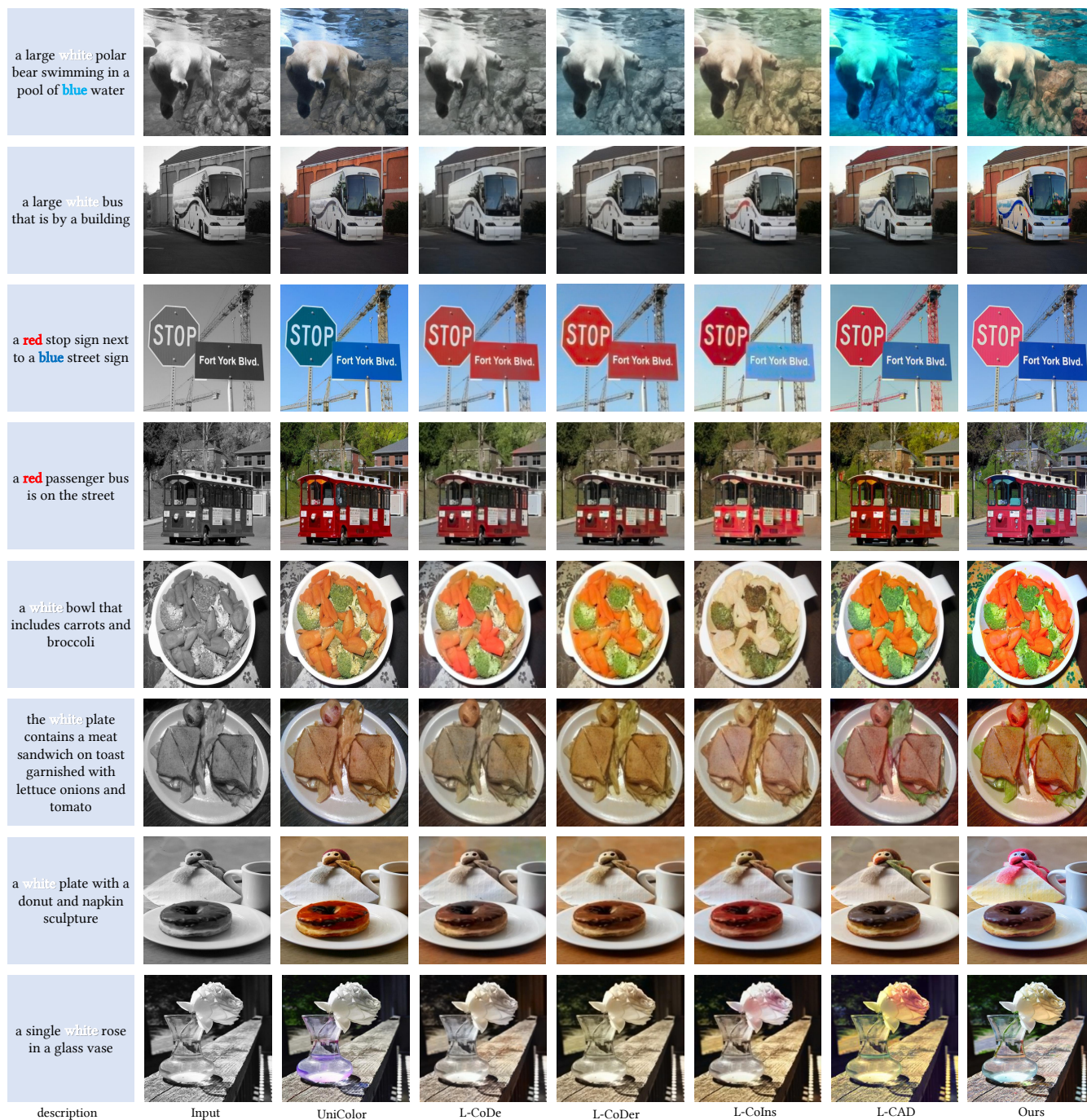


Figure 3: Qualitative results compared with previous language-based colorization methods.



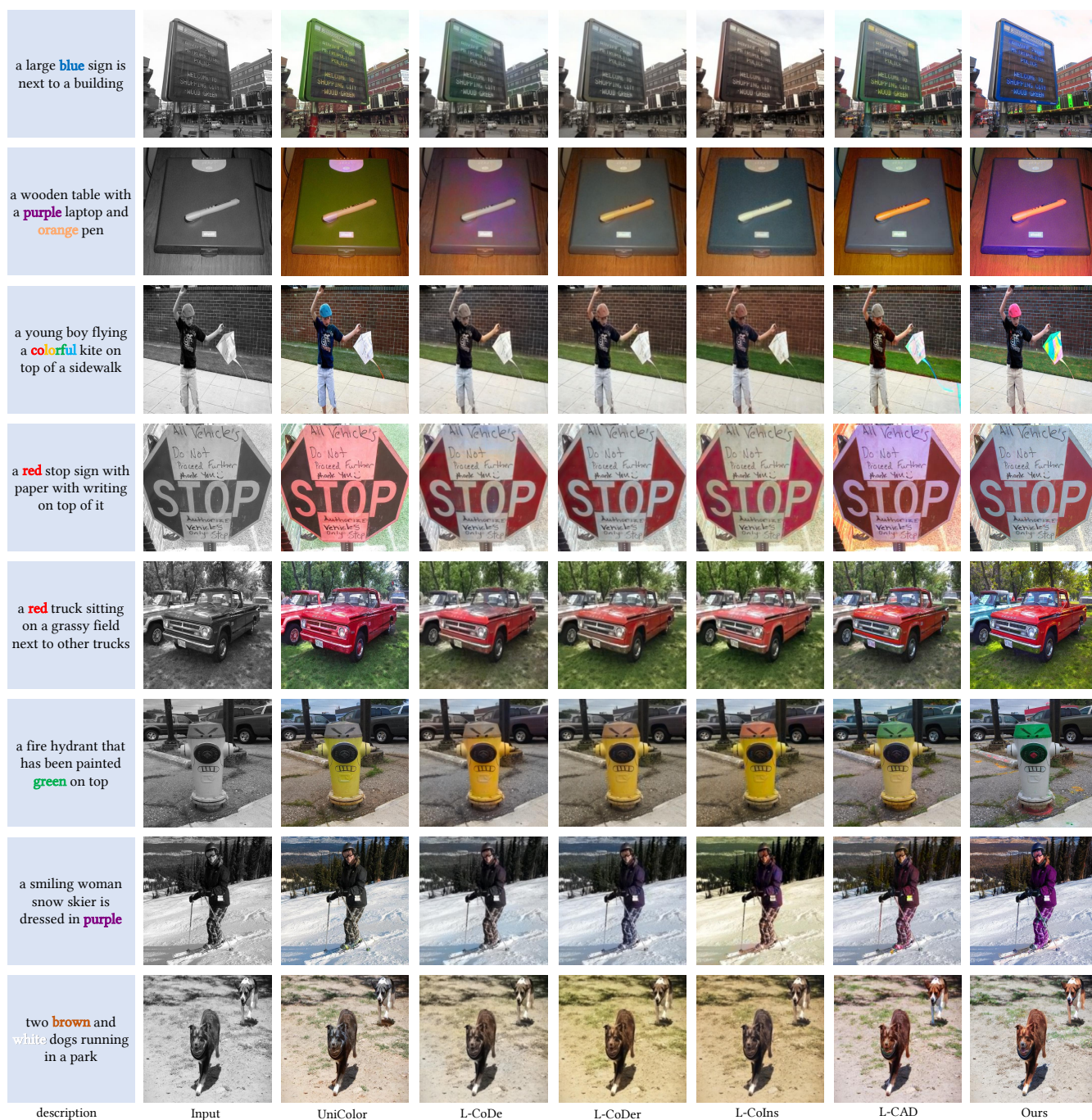


Figure 4: Qualitative results compared with previous language-based colorization methods.