CLIPtone: Unsupervised Learning for Text-based Image Tone Adjustment

Supplementary Material

S.1. User Study Details

For the user study, we recruited 20 participants from our institution and each participant was asked to complete a questionnaire as shown in Fig. 1. The questionnaire consists of an answer sheet and examples for randomly selected 30 images and 30 text descriptions from our test set. In each example, an input image and randomly arranged result images of different methods are displayed. The compared methods include CLIPstyler [2], IP2P [1], and CLIPtone, but do not include T2ONet [3] as it is limited to predefined text descriptions. Each participant was asked to rate each result on a scale from 1 (poor) to 5 (excellent) based on the following questions:

- (Q1) How well do you think the structure of the input image has been preserved in the resulting image?
- (Q2) How appropriately do you think the modifications were made considering the text description?
- (Q3) How would you rate the aesthetic quality of the resulting image?

S.2. Detailed Test Set Descriptions

For quantitative comparison, we manually collected 50 text descriptions that represent tonal properties such as brightness, contrast, color balance, and the overall mood of an image. The descriptions are as follows:
[“warm”, “bright”, “dark”, “gloomy”, “bold”, “dull”, “dramatic”, “low contrast”, “colorful”, “vibrant”, “HDR”, “vintage”, “saturated”, “neon light”, “daylight”, “sunlight”, “vivid”, “lively”, “high contrast”, “faded”, “beautiful”, “captivating”, “cozy”, “cold”, “monochrome”, “grayscale”, “old”, “modern”, “cyberpunk”, “pastel tone”, “moonlight”, “nighttime”, “desaturated”, “twilight”, “pop art”, “gothic”, “polaroid”, “analog”, “silhouette”, “cinematic”, “dreamy”, “romantic”, “mysterious”, “relaxed”, “dynamic”, “melancholic”, “urban”, “ethereal”, “matte”, “edgy”]
where the 21 underlined words are those usable in T2ONet [3].

S.3. Additional Validation of Hypothesis with Different CLIP Models

Fig. 2 shows additional hypothesis validation experiments in Sec. 3 in the main paper with different CLIP models: ViT-B/32 and ViT-L/14. Consistent with the result presented in the main paper, both results demonstrate consistent increases in similarities in filtered images, supporting our hypothesis that CLIP can serve as a perceptual criterion for tone adjustments.

S.4. Trade-off between Content Loss and CLIP Loss

Fig. 3 presents a qualitative example to show the trade-off between the content-preserving loss $\mathcal{L}_{\text{content}}$ and the CLIP loss $\mathcal{L}_{\text{CLIP}}$. As shown in the figure, a larger value for the weight λ_{content} leads to tone adjustment results that are closer to the input image, while a larger value for the weight λ_{CLIP} leads to more significant alterations in the tonal properties of the input image.

S.5. Color Biases of CLIPtone

As mentioned in the main paper, CLIPtone is not free from limitations. One limitation is that it exhibits unintended color biases for certain text descriptions as shown in Fig. 4. In the figure, CLIPtone adjusts the input images towards purple for the input text description “vivid photo”. Notably, this kind of color bias is not exclusive to our model but is also found in other CLIP-based models, such as CLIPstyler [2] and IP2P [1]. This result indicates that such color biases are caused by the inherent biases within the pre-trained CLIP models.

S.6. Additional Qualitative Comparisons

Fig. 5 shows additional qualitative comparisons of different methods. It highlights that only CLIPtone successfully makes the appropriate adjustments aligned with the text descriptions while preserving the structure of the image. Meanwhile, T2ONet [3] makes only subtle adjustments, such as slight brightness enhancements, failing to adjust adequately according to descriptions. Both CLIPstyler [2] and IP2P [1] make adjustments aligned with the descriptions, but struggle to preserve the input image’s structure.

S.7. Additional Quantitative Comparisons with Other Versions of the CLIP Model

We further report additional quantitative results calculated with other versions of the CLIP model in Tab. 1. In the table, RN50 is the version used to train CLIPtone, ViT-L/14 for IP2P [1], ViT-B/32 for CLIPstyler [2], and RN101 is not used to train any model. As analyzed in the main paper, T2ONet [3] generally achieves high scores in CLIP Image Similarity, yet it scores the lowest in CLIP Text-Image Direction Similarity. In contrast, CLIPstyler [2] shows an opposite trend in these metrics. Both IP2P [1] and CLIPtone record high scores in both metrics, and among them, CLIPtone records higher scores.

S.8. Additional Examples of the Zero-Shot Prediction

We provide additional examples of the zero-shot prediction in Fig. 6. Although none of the descriptions were included in our training dataset, CLIPtone successfully captures the unique tonal properties from the text descriptions and makes proper adjustments to input images.

S.9. Results of CLIPtone with Long Text Descriptions.

We provide results of CLIPtone with long text descriptions in Fig. 7. Despite being trained on relatively short descriptions, the results show that CLIPtone can effectively handle complex and long text descriptions.

S.10. Additional Results of CLIPtone with Various Text Descriptions.

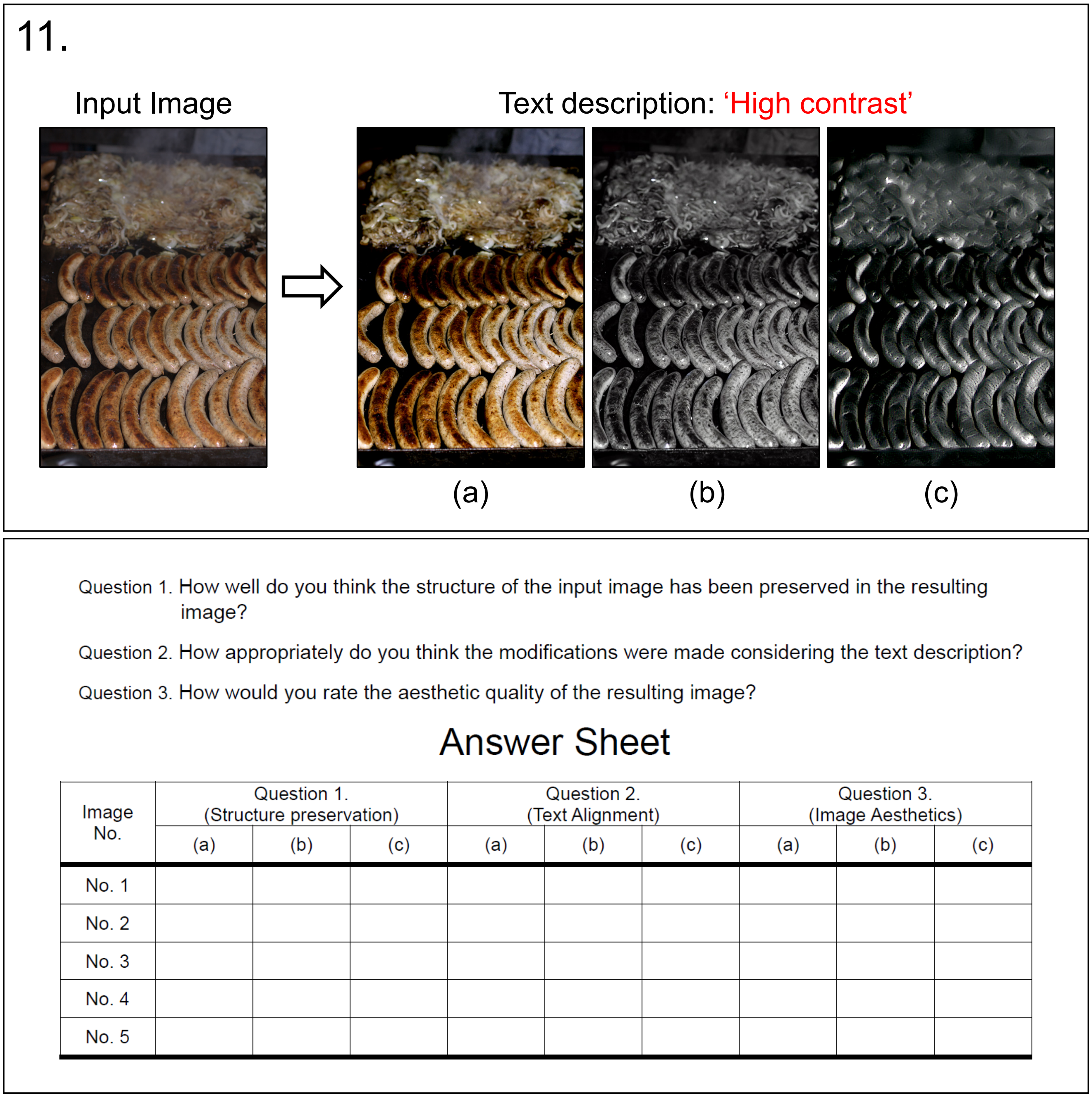
We provide additional results of CLIPtone with various text descriptions in Fig. 8, Fig. 9 and Fig. 10. With a comprehensive understanding of natural languages, CLIPtone supports a wide and diverse range of adjustments, including those previously deemed impossible.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 3

[2] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 1, 3

[3] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599, 2021. 1, 3, 5



Question 1. How well do you think the structure of the input image has been preserved in the resulting image?

Question 2. How appropriately do you think the modifications were made considering the text description?

Question 3. How would you rate the aesthetic quality of the resulting image?

Answer Sheet

Image No.	Question 1. (Structure preservation)			Question 2. (Text Alignment)			Question 3. (Image Aesthetics)		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
No. 1									
No. 2									
No. 3									
No. 4									
No. 5									

Figure 1. Our user study questionnaire. Each participant was presented upper figure consisting of an input image and results of different methods and asked to complete below answer sheet.

CLIP's version	RN50 (CLIPtone)		VIT-L/14 (IP2P)		ViT-B/32 (CLIPstyler)		RN101 (None)	
Method	CLIP Image Similarity	CLIP Text-Image Direction Similarity	CLIP Image Similarity	CLIP Text-Image Direction Similarity	CLIP Image Similarity	CLIP Text-Image Direction Similarity	CLIP Image Similarity	CLIP Text-Image Direction Similarity
T2ONet [3]	**0.994**	0.029	**0.994**	-0.001	**0.994**	-0.009	**0.999**	0.030
CLIPstyler [2]	0.604	0.082	0.657	0.059	0.656	**0.111**	0.755	**0.088**
IP2P [1]	0.915	0.065	0.917	0.060	0.928	0.078	0.958	0.074
CLIPtone	0.964	**0.089**	0.975	**0.061**	0.966	0.069	0.988	0.084

Table 1. Quantitative comparisons with other versions of CLIP model. The parentheses alongside each CLIP version denote the methods using the version in their training stage.

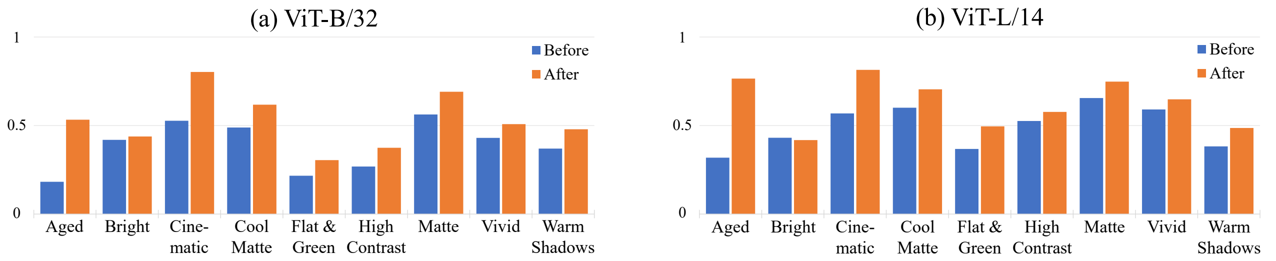
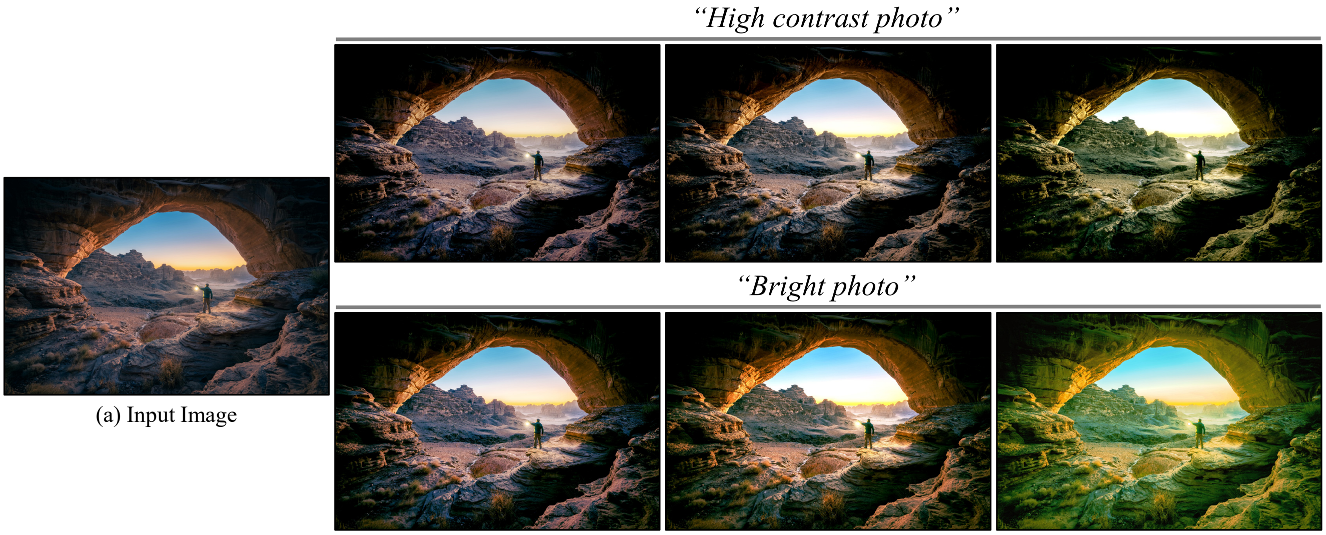


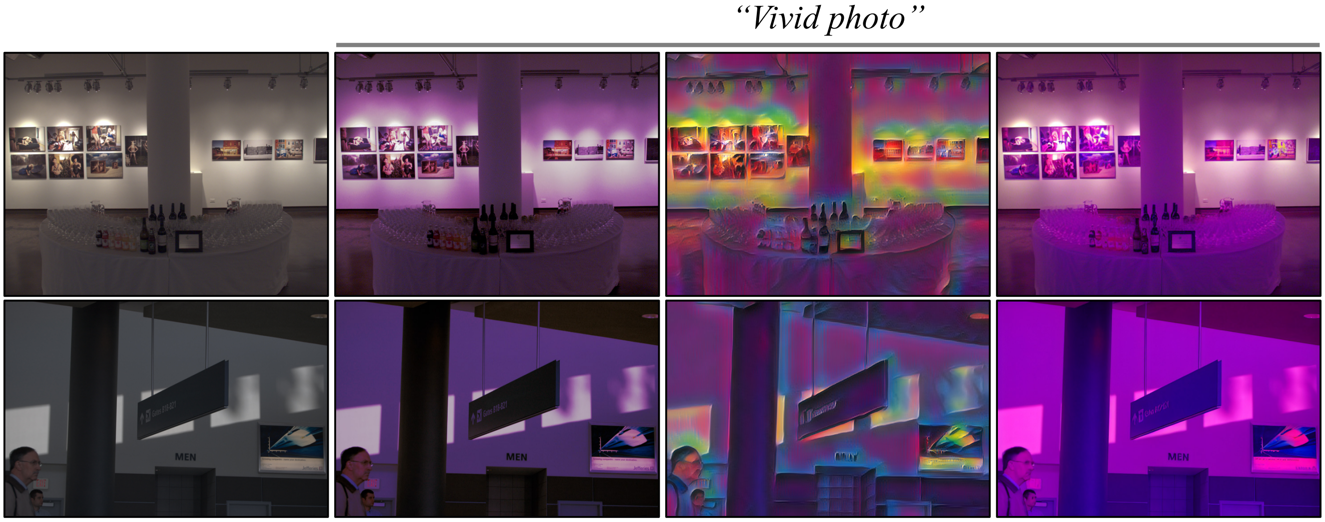
Figure 2. Additional hypothesis validation experiments using different CLIP models. Both results consistently show increases in similarities in filtered images, which again implies that CLIP is capable of assessing tonal properties of images, aligning with human perception.



(a) Input Image

(b) $\lambda_{content} = 1.5,\ \lambda_{CLIP} = 0.5$ (c) $\lambda_{content} = 1.0,\ \lambda_{CLIP} = 1.0$ (d) $\lambda_{content} = 0.5,\ \lambda_{CLIP} = 1.5$

Figure 3. Effects of the content loss $\mathcal{L}_{\text{content}}$ and CLIP loss $\mathcal{L}_{\text{CLIP}}$. Larger weight λ_{content} for the content loss leads to tone adjusted results that are closer to the input image, while larger weight λ_{content} for the CLIP loss leads to more significant alterations.



Input Image CLIPtone CLIPstyler [2] IP2P [1]

Figure 4. Examples of color bias of CLIP-based methods. CLIPtone and other CLIP-based methods show similar color-biased results that adjust the input image toward purple tone.

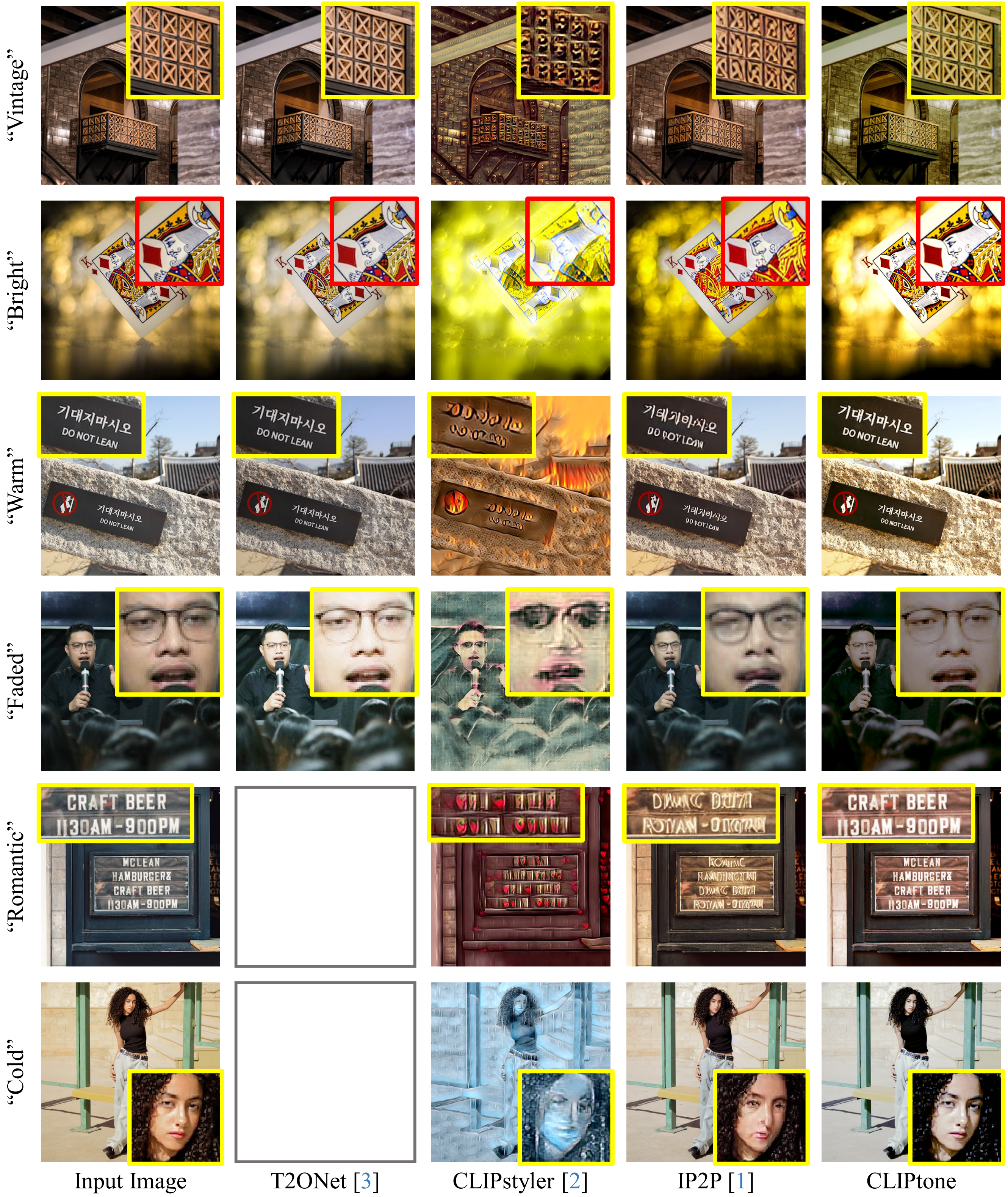


Figure 5. Additional qualitative comparison of different methods. T2ONet [3] does not support "Romantic" and "Cold".

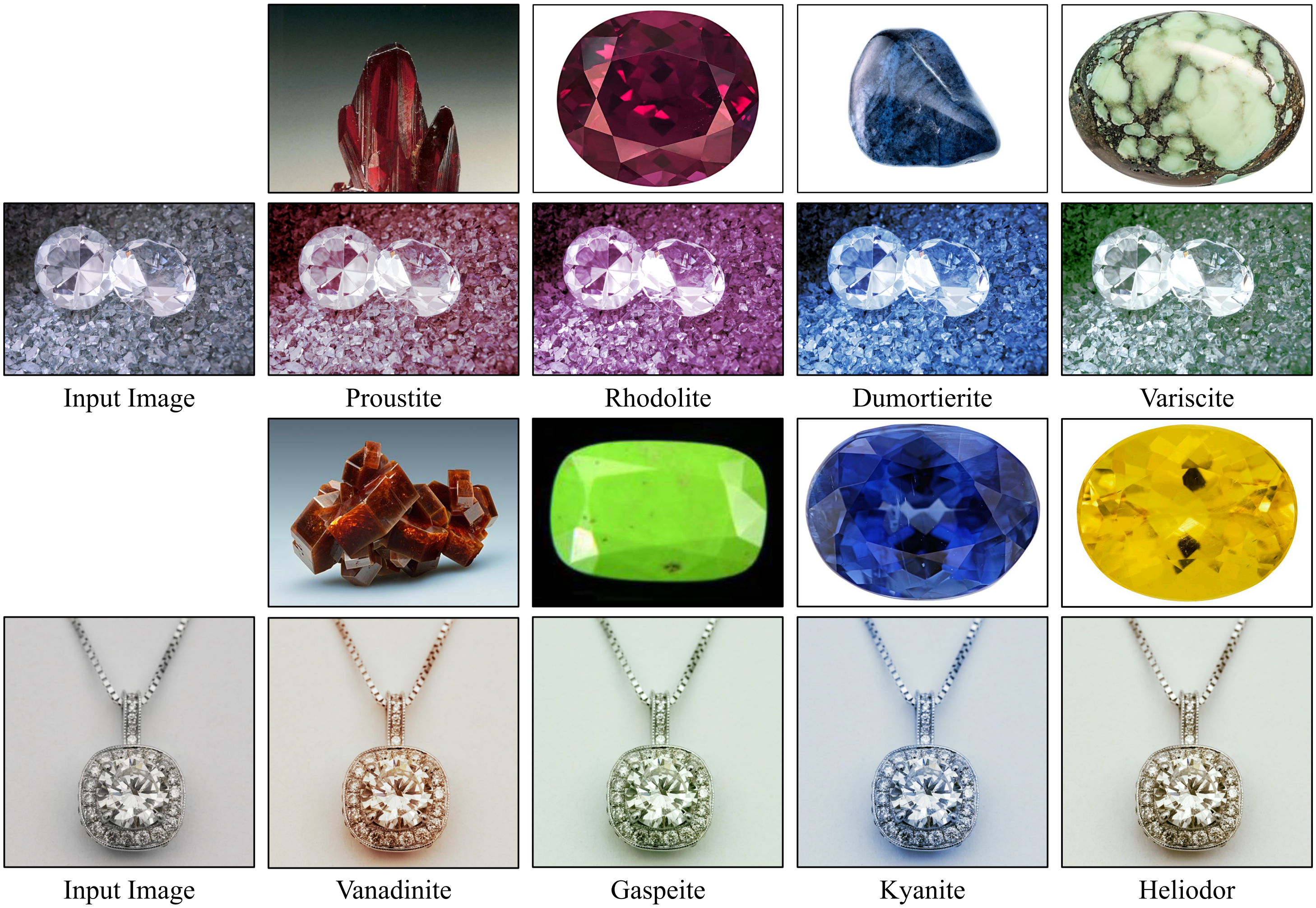
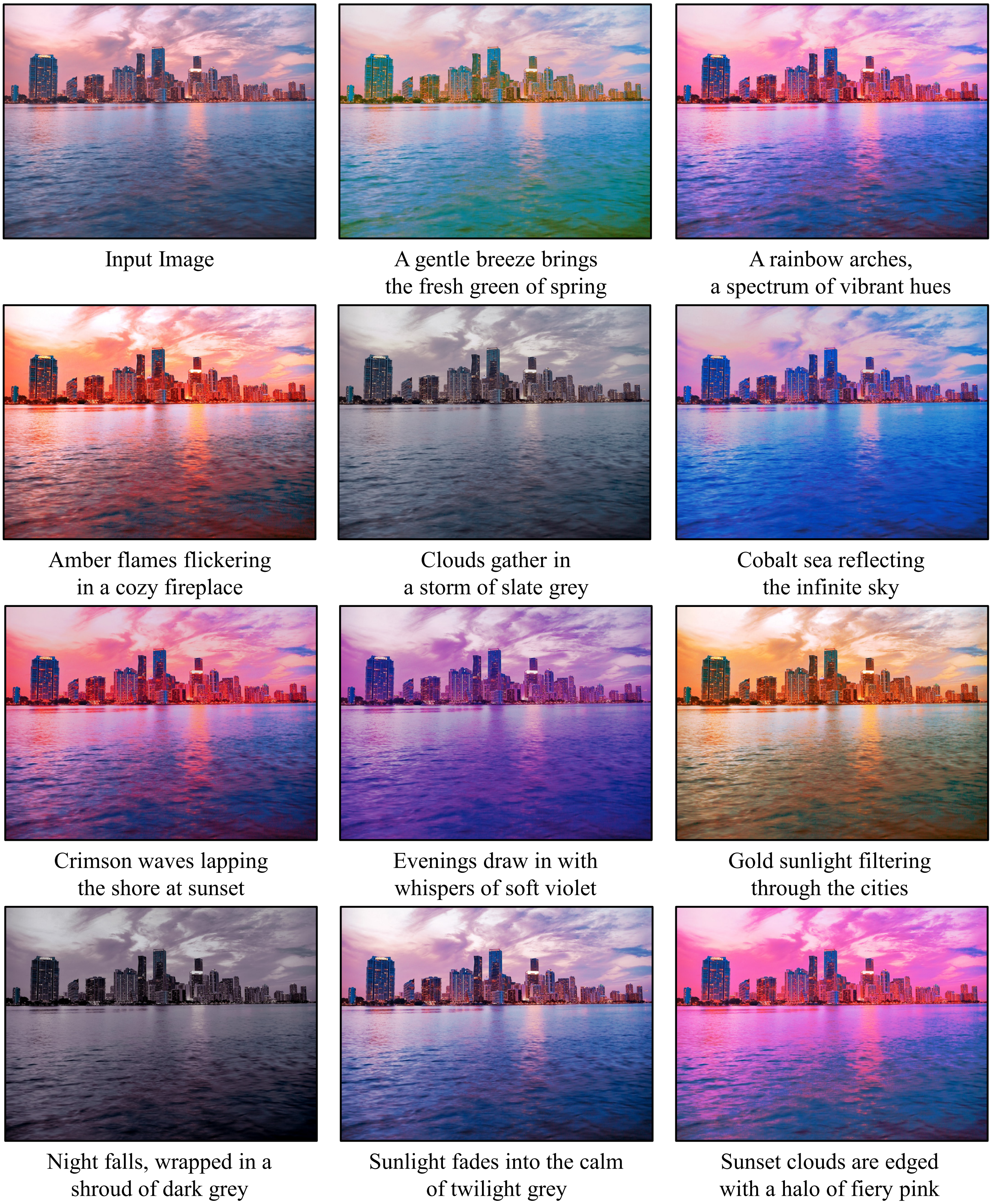


Figure 6. Additional examples of CLIPtone to demonstrate its zero-shot prediction capability. CLIPtone successfully executes proper adjustments for words related to minerals, even those not encountered during training. In each example, the top row displays instances of each word to facilitate comprehension.



Input Image

A gentle breeze brings the fresh green of spring

A rainbow arches, a spectrum of vibrant hues

Amber flames flickering in a cozy fireplace

Clouds gather in a storm of slate grey

Cobalt sea reflecting the infinite sky

Crimson waves lapping the shore at sunset

Evenings draw in with whispers of soft violet

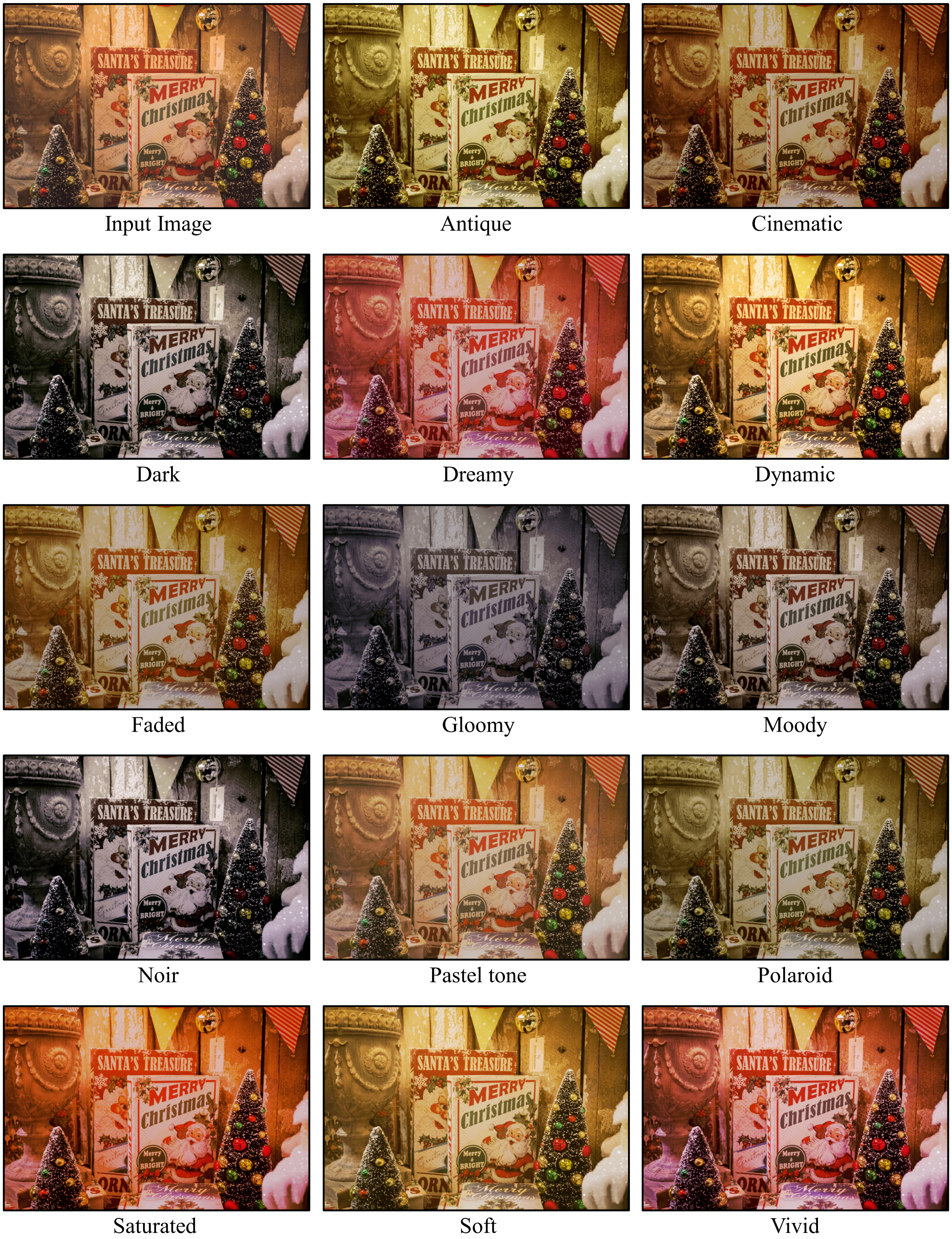
Gold sunlight filtering through the cities

Night falls, wrapped in a shroud of dark grey

Sunlight fades into the calm of twilight grey

Sunset clouds are edged with a halo of fiery pink

Figure 7. Additional results of CLIPtone with long and complex text descriptions.



Input Image

Antique

Cinematic

Dark

Dreamy

Dynamic

Faded

Gloomy

Moody

Noir

Pastel tone

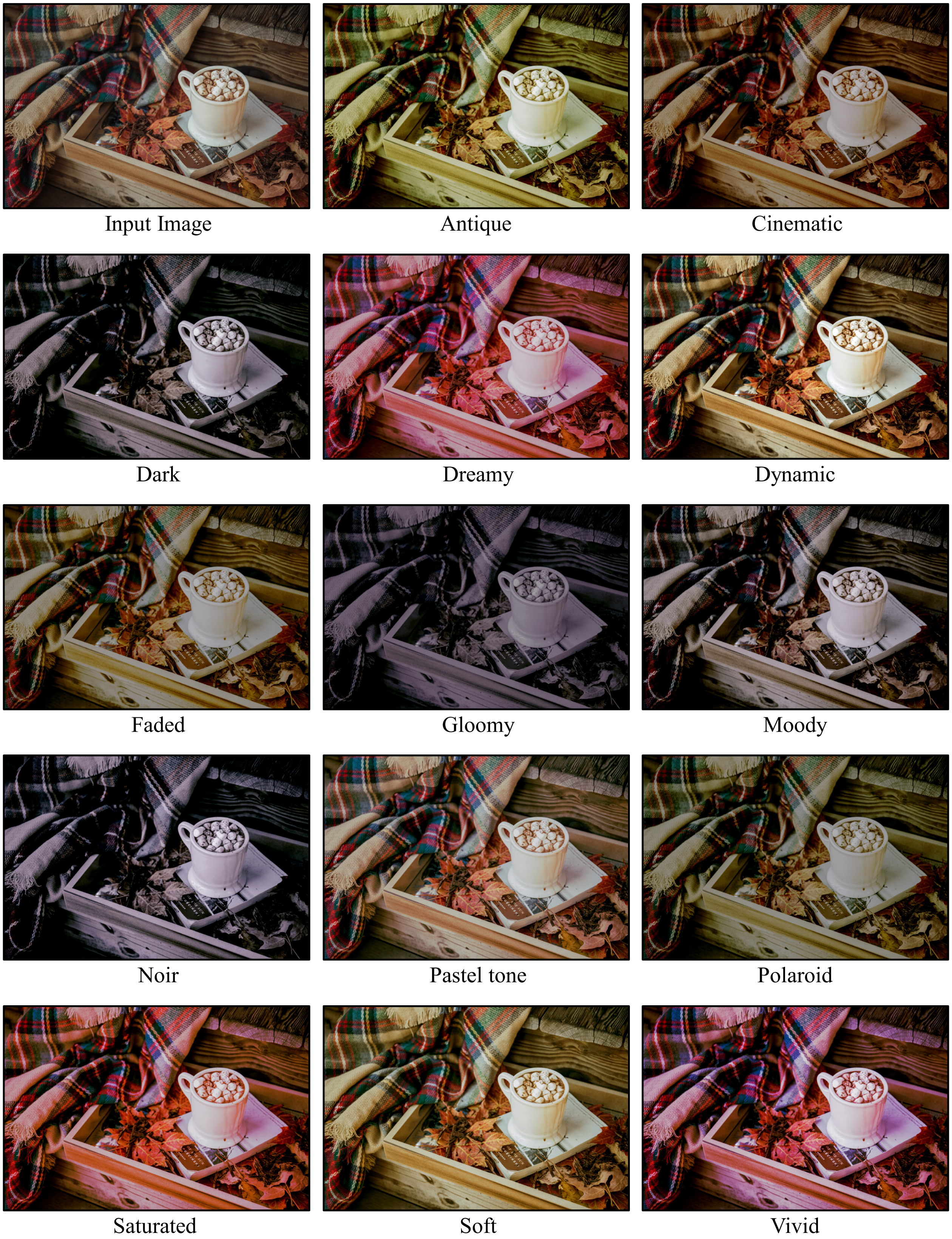
Polaroid

Saturated

Soft

Vivid

Figure 8. Additional results of CLIPtone with various text descriptions.



Input Image Antique Cinematic

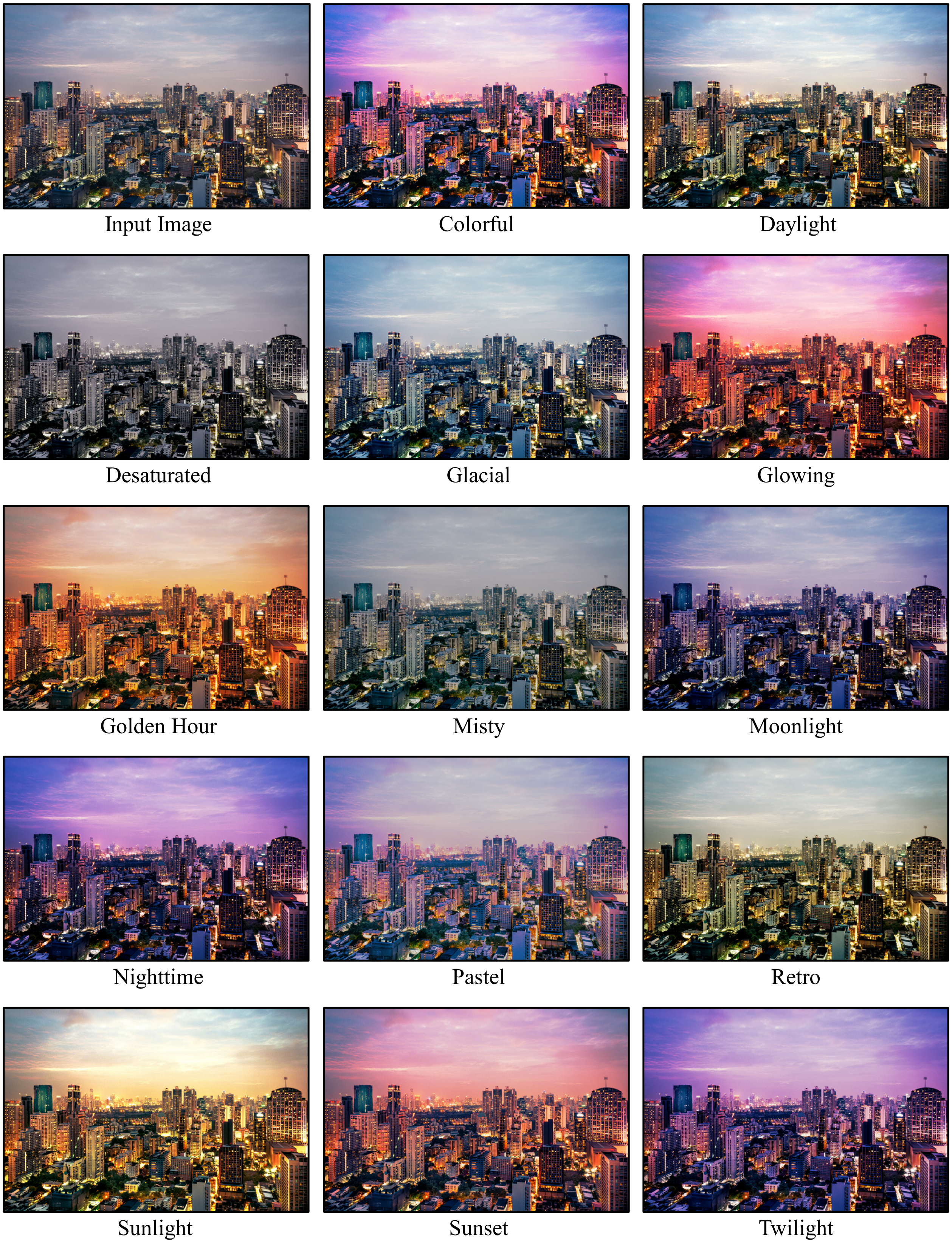
Dark Dreamy Dynamic

Faded Gloomy Moody

Noir Pastel tone Polaroid

Saturated Soft Vivid

Figure 9. Additional results of CLIPtone with various text descriptions.



Input Image

Colorful

Daylight

Desaturated

Glacial

Glowing

Golden Hour

Misty

Moonlight

Nighttime

Pastel

Retro

Sunlight

Sunset

Twilight

Figure 10. Additional results of CLIPtone with various text descriptions.