Supplmentary Material:
Affective Image Filter: Reflecting Emotions from Text to Images

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

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

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

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

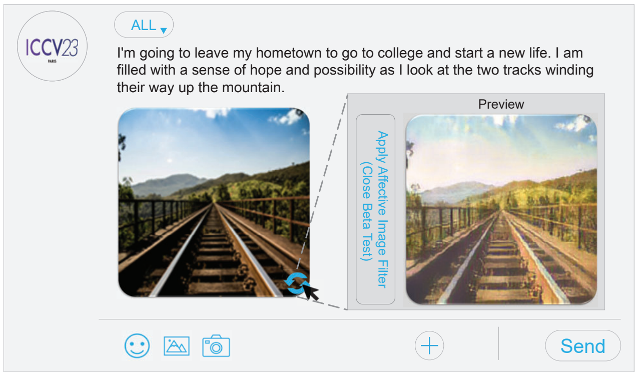
[4]Beijing Academy of Artificial Intelligence

{shuchenweng,shiboxin}@pku.edu.cn, {pxzhang,zhengchang98,lisi}@bupt.edu.cn, wangxinlong@baai.ac.cn

7. Appendix

7.1. Application

We present a potential application of the AIF model in social media in Fig. 7: When users input text and corresponding images to express their emotions and personalize their content, the platform could offer candidate images in real-time that reflect their emotions, thereby helping them stand out from the crowd and attract more followers.

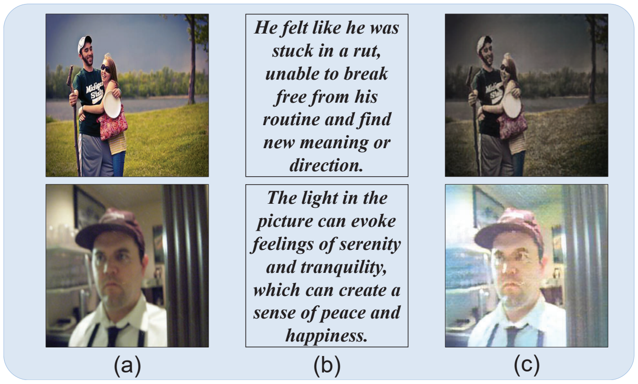


I'm going to leave my hometown to go to college and start a new life. I am filled with a sense of hope and possibility as I look at the two tracks winding their way up the mountain.

Figure 7: Application of the AIF model on social media.

7.2. Failure cases

Since the visual content of images is a crucial feature to evoke specific emotional responses from human observers, the AIF model may have difficulty in producing convincing results when the user-provided content image does not properly express the intended emotion of the text. We show two failure cases in Fig. 8, where neither does the AIF model reflect the frustrated emotion to hugging couples with happy smiles nor does it create a filter that makes the man with a disgusted expression look peaceful and happy.



He felt like he was stuck in a rut, unable to break free from his routine and find new meaning or direction.

The light in the picture can evoke feelings of serenity and tranquility, which can create a sense of peace and happiness.

Figure 8: Failure cases of the AIF model. (a) User-provided content images. (b) Texts that reflect thoughts and feelings. (c) Failure results produced by the AIF model.

7.3. Hyperparameters sensitivity experiments

In our experiments, hyperparameters of training losses are empirically set, not sensitive to variations in a certain range, except for λ_s and λ_c that make a trade-off between preserving visual contents and creating colors and textures. We show the sensitivity analysis in Tab. 3

Table 3: The sensitivity analysis of hyperparameters. Best performances are highlighted in **bold**.

λ_sm	λ_as	λ_ed	λ_GAN	λ_s	λ_id	λ_c	SSIM (%) ↑	SD ↓	SG (‰) ↓	Acc (%) ↑
30	600	140	3	0.3	2	5	56.15	**5.41**	**1.39**	**29.96**
6/150	600	140	3	0.3	2	5	54.19/55.40	5.67/5.62	1.42/1.42	29.33/28.63
30	120/3000	140	3	0.3	2	5	54.56/53.67	5.65/5.46	1.47/1.45	28.86/28.68
30	600	28/700	3	0.3	2	5	53.43/54.24	5.51/5.50	1.40/1.42	28.02/29.50
30	600	140	0.6/15	0.3	2	5	54.12/53.29	5.71/5.70	1.40/1.39	29.45/29.37
30	600	140	3	0.06/1.5	2	5	**63.14**/33.05	6.08/5.76	1.46/1.50	28.69/29.58
30	600	140	3	0.3	0.4/10	5	51.27/54.25	5.72/5.71	1.42/1.43	29.85/29.27
30	600	140	3	0.3	2	1/25	33.22/62.68	5.80/5.96	1.47/1.49	29.09/26.81

7.4. Parameter setting

We create three variants to explore three different parameter settings of the AIF transformer, including the number of transformer blocks, hidden size, the MLP size, the number of attention heads, and the parameter number, as shown in Tab. 4. We further show the quantitative results for each setting in Tab. 5, which demonstrate that the performance improves as the number of parameters increases. In the main paper, we report the results of AIF (Base).

Table 4: Different parameter settings.

Model	Blocks	Hidden size	MLP size	Heads	#Params.
AIF (Small)	3	384	1536	6	14.5M
AIF (Base)	4	768	2048	8	26.4M
AIF (Large)	6	1024	3072	12	69.3M

Table 5: Quantitative experiment results of different parameter settings. Throughout the paper, ↑ (↓) means higher (lower) is better. Best performances are highlighted in **bold**.

Model	SSIM (%) ↑	SD ↓	SG (‰) ↓	Acc (%) ↑
AIF (Small)	53.57	5.5913	1.4180	28.76
AIF (Base)	56.15	5.4147	1.3881	29.96
AIF (Large)	**57.62**	**5.0279**	**1.3380**	**30.30**

7.5. Distribution estimator (DE)

As illustrated in Eq. (8) (Sec. 4.3.2) of the main paper, we pre-train a distribution estimator to assist the AIF model to reflect emotions from text to images more accurately. Instead of building the distribution estimator according to previous works [8] that focus on the visual content (denoted as visual DE), we take colors and textures as cues to build the distribution estimator tailored for the AIF task (denoted as AIF DE). As shown in Fig. 9, our AIF DE extracts multi-level features with the VGG network [7], and then calculates means and variances at each level. As such, the AIF DE discards the features of concrete visual content while obtaining the features of colors and textures.

We use the following three metrics to comprehensively evaluate the performance of the AIF DE and visual DE: *(i)* **Top:** We select the most possible emotion in the estimated distribution, and measure whether it matches the emotional category associated with the input text. *(ii)* **Any:** We measure whether the most possible emotion is one of the candidate emotional categories of the corresponding anchor image. *(iii)* **KLD:** We calculate the Kullback-Leibler divergence [3] between the estimated emotional distribution and the ground truth. As shown in Tab. 6, our AIF DE achieves higher scores on all three scores for synthesized images and ground truth images. In the main paper, we use the **Top** as the **Acc** score to measure the accuracy.

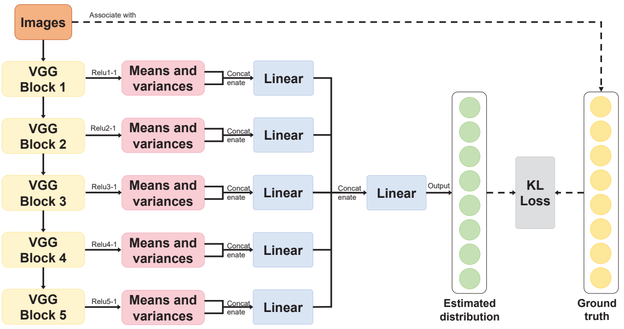


Figure 9: Architecture of the AIF DE.

Table 6: Quantitative comparisons between the AIF DE and the visual DE.

Method	AIF DE			Visual DE		
	Top (%) ↑	Any (%) ↑	KLD (‰) ↓	Top (%) ↑	Any (%) ↑	KLD (‰) ↓
ManiGAN	27.77	49.59	4.4870	24.32	44.84	5.3009
DiffusionCLIP	24.59	44.13	4.8778	23.76	43.09	5.4089
CLIPStyler	26.40	49.83	4.6767	26.30	48.67	5.1671
CLVA	25.64	51.06	4.4051	26.22	48.92	4.9922
AIF (Small)	28.76	51.01	4.2597	26.26	48.69	4.7681
AIF (Base)	29.96	52.16	4.2541	26.68	48.94	4.7651
AIF (Large)	30.30	52.70	4.1819	26.73	48.94	4.7101
Ground truth	**42.88**	**66.65**	**3.5029**	**33.61**	**56.14**	**4.2238**

7.6. Conditional-unconditional discriminator

In Eq. (9) (Sec. 4.4) of the main paper, we develop a multi-level conditional-unconditional discriminator to align synthesized images with user-provided texts and discriminate whether synthesized images are aesthetically pleasing. In this subsection, we further present the architecture of the conditional-unconditional discriminator in Fig. 10. Similar to the AIF DE in Sec. 7.5, we extract multi-level features that represent colors and textures and use them to calculate unconditional discriminator scores. We also concatenate these features with text tokens at each level to calculate conditional discriminator scores.

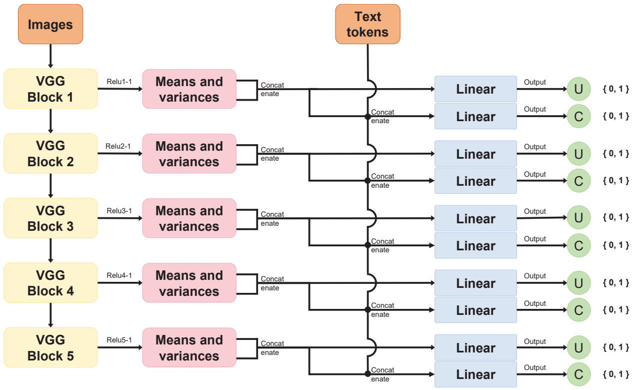
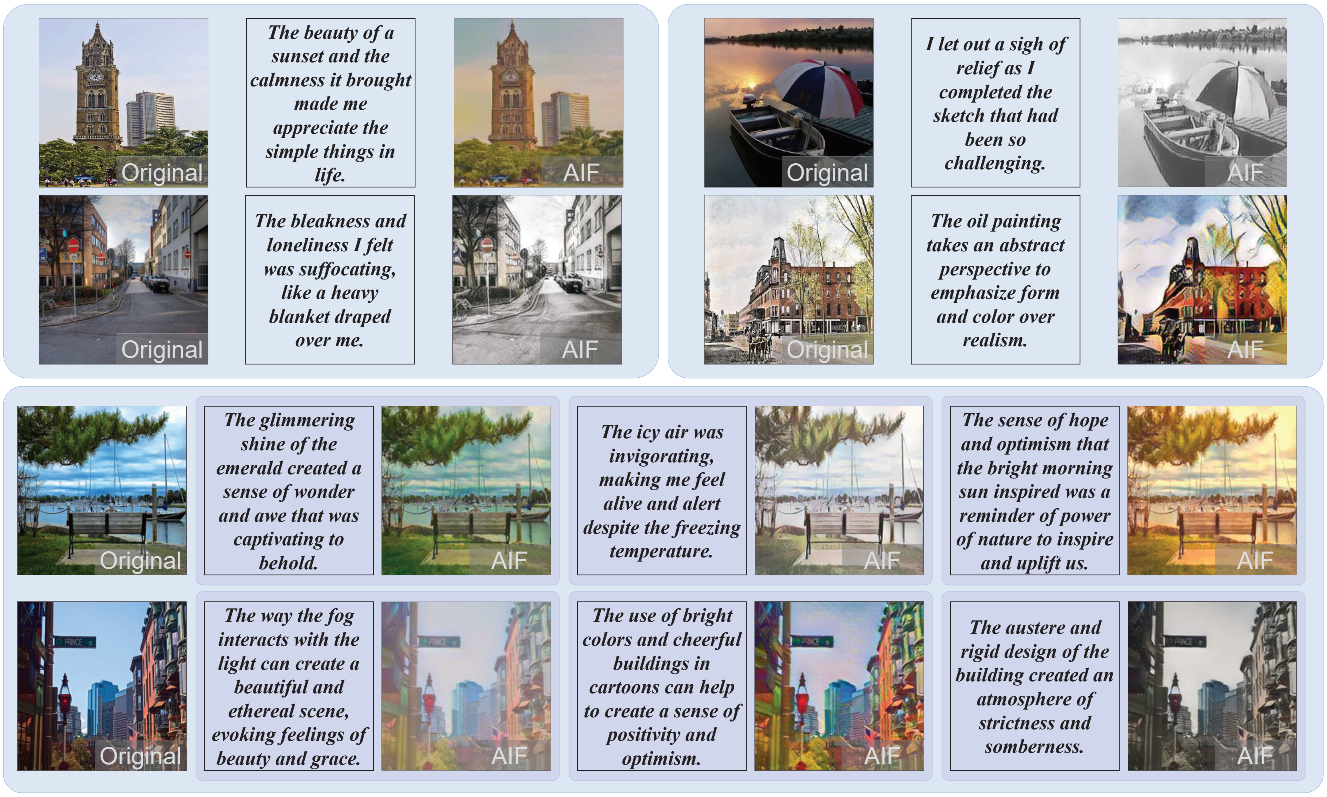


Figure 10: Architecture of the conditional-unconditional discriminator, which outputs an unconditional score (U) and a conditional score (C) at each level.


The beauty of a sunset and the calmness it brought made me appreciate the simple things in life.

I let out a sigh of relief as I completed the sketch that had been so challenging.

The bleakness and loneliness I felt was suffocating, like a heavy blanket draped over me.

The oil painting takes an abstract perspective to emphasize form and color over realism.

The glimmering shine of the emerald created a sense of wonder and awe that was captivating to behold.

The icy air was invigorating, making me feel alive and alert despite the freezing temperature.

The sense of hope and optimism that the bright morning sun inspired was a reminder of power of nature to inspire and uplift us.

The way the fog interacts with the light can create a beautiful and ethereal scene, evoking feelings of beauty and grace.

The use of bright colors and cheerful buildings in cartoons can help to create a sense of positivity and optimism.

The austere and rigid design of the building created an atmosphere of strictness and somberness.

Figure 11: More application results.

7.7. Additional qualitative results

In Fig. 11, we show the generalization of the AIF model. Users could freely reflect their personal thoughts and feelings to arbitrary images, *e.g.*, “calmness” and ”loneliness”. Using specific words, they could also customize unique styles, *e.g.*, “sketch” and “oil painting”. In addition, given an arbitrary image and different texts, the AIF model could synthesize various results.

We make additional comparisons with relevant image editing methods (*i.e.*, ManiGAN [5] and DiffusionCLIP [2]) and style transfer methods (*i.e.*, CLIPstyler [4] and CLVA [1]), and show comparison results in Fig. 12. We further show more qualitative ablation study results in Fig. 13, whose details are presented in Sec. 5.3 of the main paper.

7.8. Dataset sample

As illustrated in Sec. 3, we collect and process a large-scale dataset tailored for the AIF task, including abundant aesthetic images and corresponding text descriptions associated with the closest emotional category in the Mikel’s wheel [6]. We show more samples of the AIF dataset in Fig. 14 to inspire relevant research.

References

[1] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *ECCV*, 2022. 3, 4

[2] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 3, 4

[3] Solomon Kullback and Richard A Leibler. On information and sufficiency. *AoMS*, 1951. 2

[4] Gihyun Kwon and Jong Chul Ye. CLIPstyler: Image style transfer with a single text condition. In *CVPR*, 2022. 3, 4

[5] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. ManiGAN: Text-guided image manipulation. In *CVPR*, 2020. 3, 4

[6] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *BRM*, 2005. 3, 6

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 2

[8] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 2017. 2

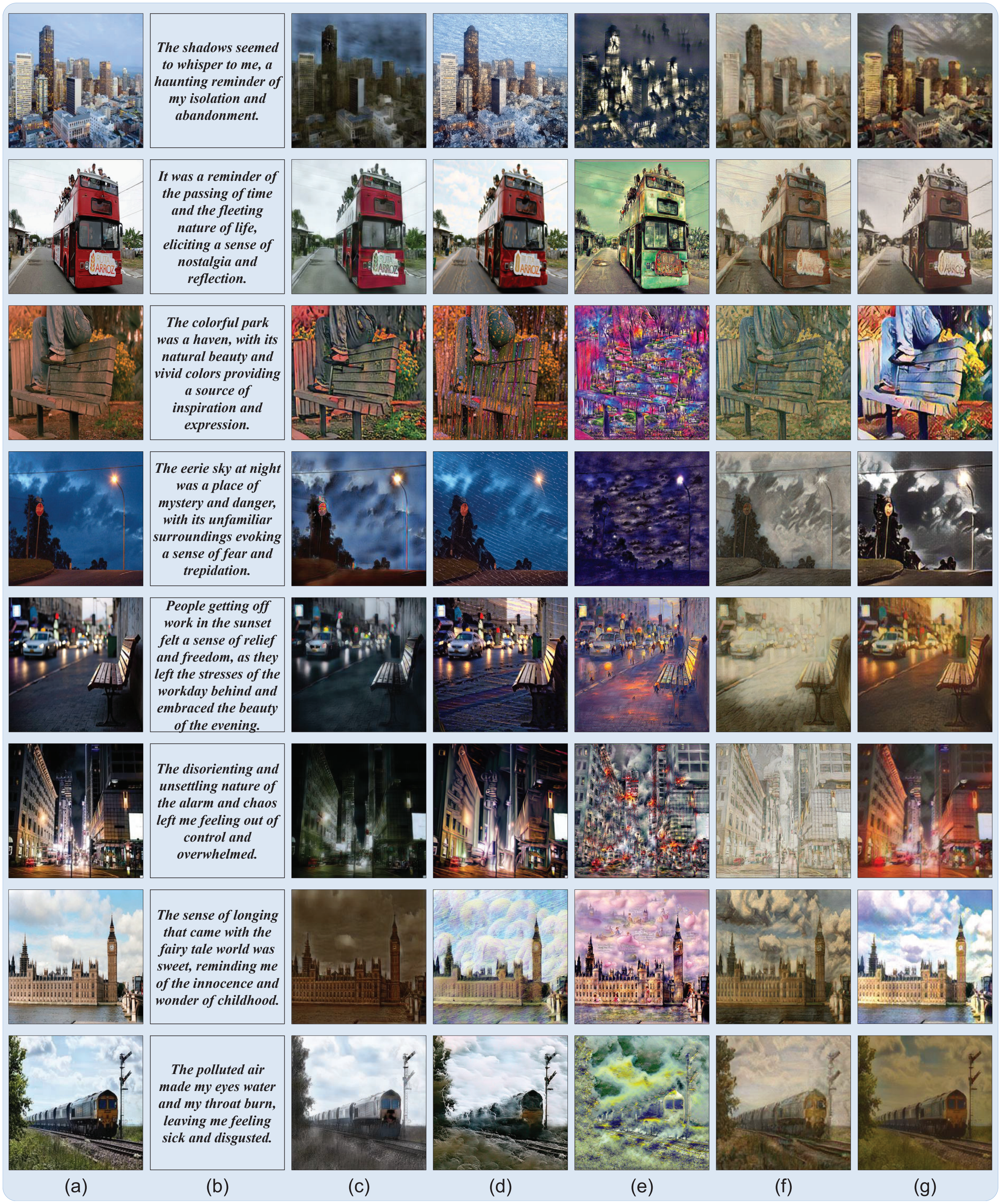
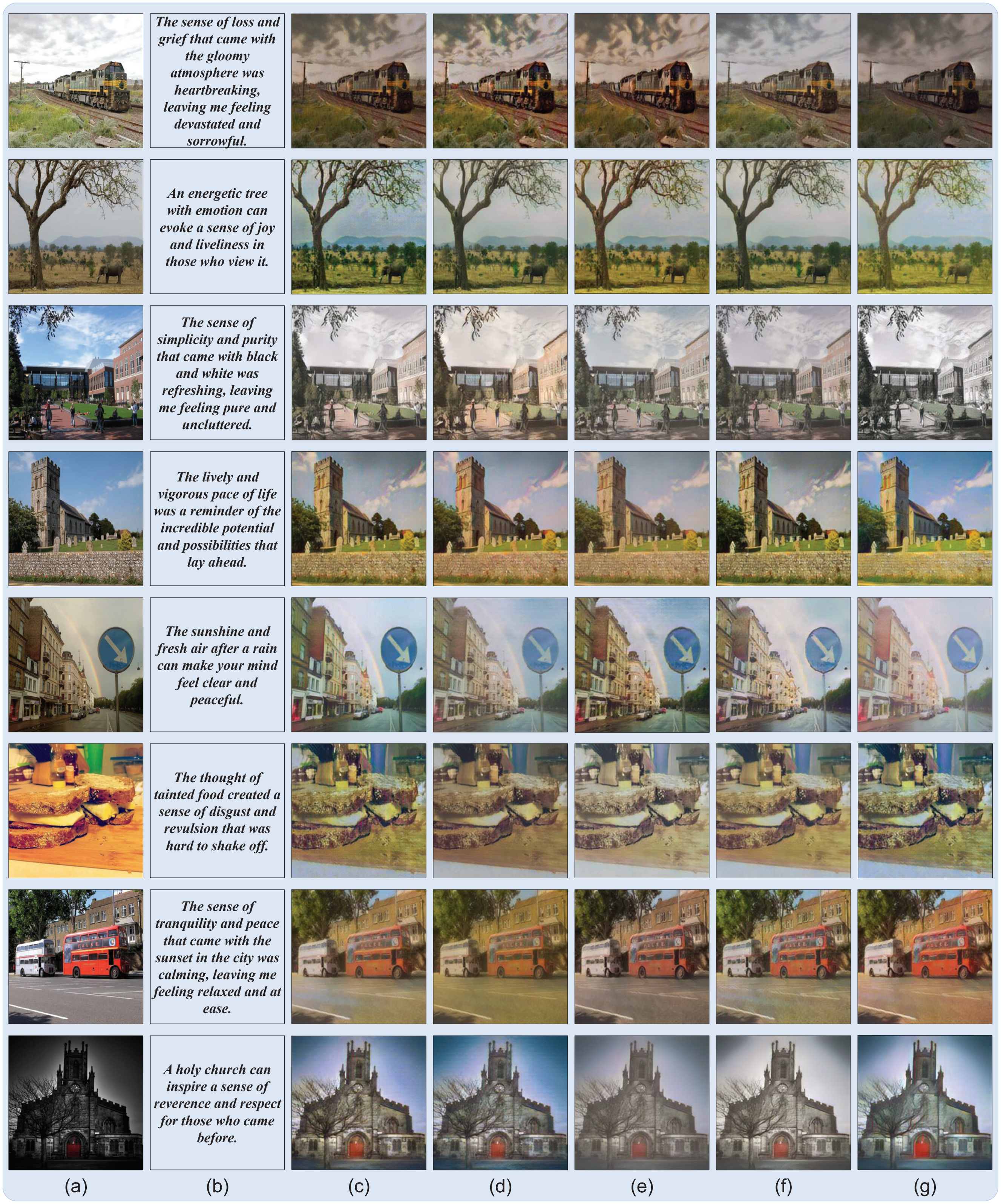

The shadows seemed to whisper to me, a haunting reminder of my isolation and abandonment.
It was a reminder of the passing of time and the fleeting nature of life, eliciting a sense of nostalgia and reflection.
The colorful park was a haven, with its natural beauty and vivid colors providing a source of inspiration and expression.
The eerie sky at night was a place of mystery and danger, with its unfamiliar surroundings evoking a sense of fear and trepidation.
People getting off work in the sunset felt a sense of relief and freedom, as they left the stresses of the workday behind and embraced the beauty of the evening.
The disorienting and unsettling nature of the alarm and chaos left me feeling out of control and overwhelmed.
The sense of longing that came with the fairy tale world was sweet, reminding me of the innocence and wonder of childhood.
The polluted air made my eyes water and my throat burn, leaving me feeling sick and disgusted.
(a)
(b)
(c)
(d)
(e)
(f)
(g)

Figure 12: Qualitative comparison results with state-of-the-art methods. (a) User-provided content images. (b) Texts that reflect thoughts and feelings. (c) ManiGAN [5]. (d) DiffusioinCLIP [2]. (e) CLIPstyler [4]. (f) CLVA [1]. (g) Our results.



The sense of loss and grief that came with the gloomy atmosphere was heartbreaking, leaving me feeling devastated and sorrowful.

An energetic tree with emotion can evoke a sense of joy and liveliness in those who view it.

The sense of simplicity and purity that came with black and white was refreshing, leaving me feeling pure and uncluttered.

The lively and vigorous pace of life was a reminder of the incredible potential and possibilities that lay ahead.

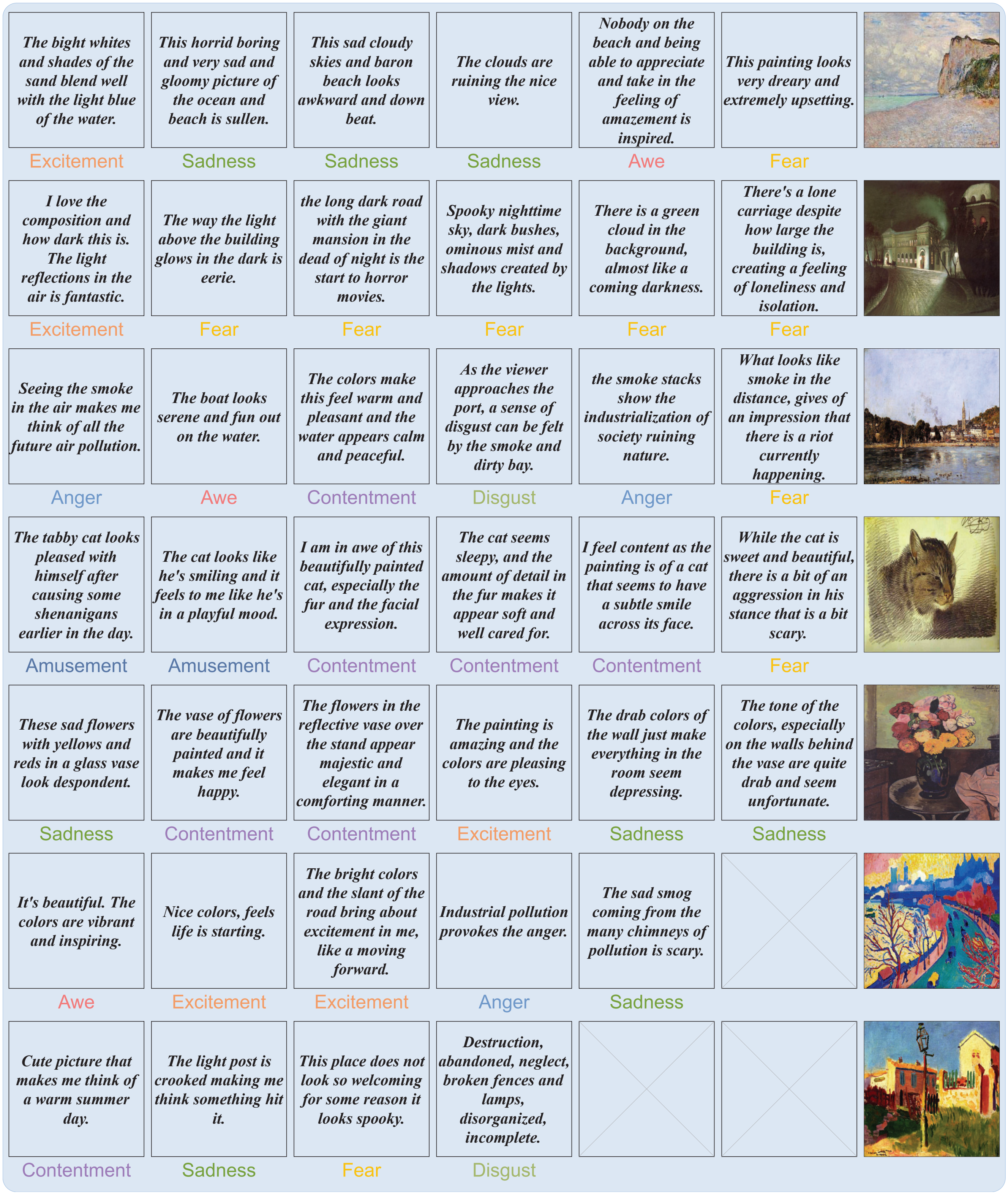
The sunshine and fresh air after a rain can make your mind feel clear and peaceful.

The thought of tainted food created a sense of disgust and revulsion that was hard to shake off.

The sense of tranquility and peace that came with the sunset in the city was calming, leaving me feeling relaxed and at ease.

A holy church can inspire a sense of reverence and respect for those who came before.

Figure 13: Ablation study results with different variants of the proposed method. (a) User-provided content images. (b) Texts that reflect thoughts and feelings. (c) W/o VAD. (d) W/o SE. (e) W/o ED. (f) W/o GAN. (g) Our results.



The bight whites and shades of the sand blend well with the light blue of the water.	*This horrid boring and very sad and gloomy picture of the ocean and beach is sullen.*	*This sad cloudy skies and baron beach looks awkward and down beat.*	*The clouds are ruining the nice view.*	*Nobody on the beach and being able to appreciate and take in the feeling of amazement is inspired.*	*This painting looks very dreary and extremely upsetting.*
Excitement	Sadness	Sadness	Sadness	Awe	Fear
I love the composition and how dark this is. The light reflections in the air is fantastic.	*The way the light above the building glows in the dark is eerie.*	*the long dark road with the giant mansion in the dead of night is the start to horror movies.*	*Spooky nighttime sky, dark bushes, ominous mist and shadows created by the lights.*	*There is a green cloud in the background, almost like a coming darkness.*	*There's a lone carriage despite how large the building is, creating a feeling of loneliness and isolation.*
Excitement	Fear	Fear	Fear	Fear	Fear
Seeing the smoke in the air makes me think of all the future air pollution.	*The boat looks serene and fun out on the water.*	*The colors make this feel warm and pleasant and the water appears calm and peaceful.*	*As the viewer approaches the port, a sense of disgust can be felt by the smoke and dirty bay.*	*the smoke stacks show the industrialization of society ruining nature.*	*What looks like smoke in the distance, gives of an impression that there is a riot currently happening.*
Anger	Awe	Contentment	Disgust	Anger	Fear
The tabby cat looks pleased with himself after causing some shenanigans earlier in the day.	*The cat looks like he's smiling and it feels to me like he's in a playful mood.*	*I am in awe of this beautifully painted cat, especially the fur and the facial expression.*	*The cat seems sleepy, and the amount of detail in the fur makes it appear soft and well cared for.*	*I feel content as the painting is of a cat that seems to have a subtle smile across its face.*	*While the cat is sweet and beautiful, there is a bit of an aggression in his stance that is a bit scary.*
Amusement	Amusement	Contentment	Contentment	Contentment	Fear
These sad flowers with yellows and reds in a glass vase look despondent.	*The vase of flowers are beautifully painted and it makes me feel happy.*	*The flowers in the reflective vase over the stand appear majestic and elegant in a comforting manner.*	*The painting is amazing and the colors are pleasing to the eyes.*	*The drab colors of the wall just make everything in the room seem depressing.*	*The tone of the colors, especially on the walls behind the vase are quite drab and seem unfortunate.*
Sadness	Contentment	Contentment	Excitement	Sadness	Sadness
It's beautiful. The colors are vibrant and inspiring.	*Nice colors, feels life is starting.*	*The bright colors and the slant of the road bring about excitement in me, like a moving forward.*	*Industrial pollution provokes the anger.*	*The sad smog coming from the many chimneys of pollution is scary.*	
Awe	Excitement	Excitement	Anger	Sadness	
Cute picture that makes me think of a warm summer day.	*The light post is crooked making me think something hit it.*	*This place does not look so welcoming for some reason it looks spooky.*	*Destruction, abandoned, neglect, broken fences and lamps, disorganized, incomplete.*		
Contentment	Sadness	Fear	Disgust		

Figure 14: More samples of the AIF dataset, where each anchor image has multiple corresponding text descriptions and each text description is categorized into the closest emotional category in the Mikel's wheel [6].