SUPPLEMENTAL MATERIAL: IS PERTURBATION-BASED IMAGE PROTECTION DISRUPTIVE TO IMAGE EDITING?

1. DETAILS OF MODIFIED CAPTIONS

This section provides detailed information on the modified captions based on the Flickr8k [1] dataset. Figure 1 presents three images, each accompanied by its original caption, a closely-modified caption, and an extensively-modified caption. The original captions are sourced from the Flickr8k dataset [1]. Subsequently, two modified versions were generated using Claude AI [2]. The closely-modified captions are derived by replacing a few words in the original captions while maintaining their semantic meaning. In contrast, the extensively-modified captions deviate significantly and are semantically unrelated to the original captions.

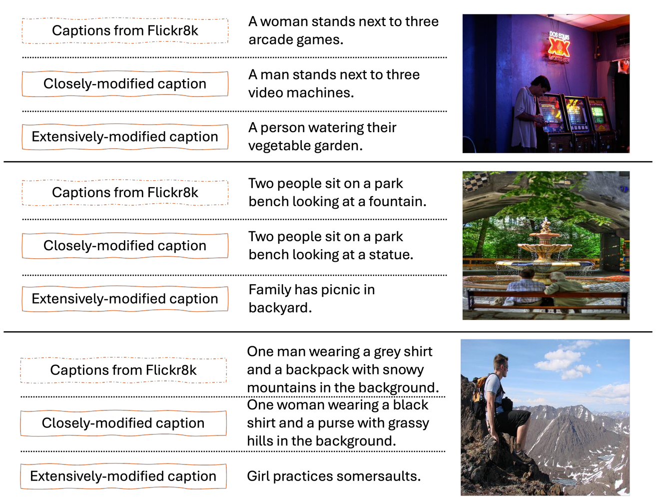

Fig. 1: We provide three examples from Flickr8k [1] dataset with their original captions, closely-modified captions, and extensively-modified captions.

To quantify the semantic similarity between the original captions and the generated versions, we utilized Google's Universal Sentence Encoder (GSE) [3]. The figures below are the GSE-evaluated distributions. A higher GSE value indicates greater similarity, whereas a lower value signifies reduced similarity.

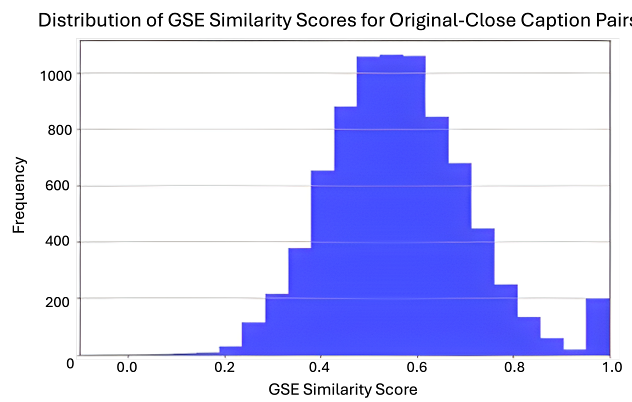

Fig. 2: The distribution of GSE similarity between the original captions and the closely-modified captions is presented. The mean value of the distribution is approximately 0.6, indicating a significant level of semantic similarity between the two sets of captions.

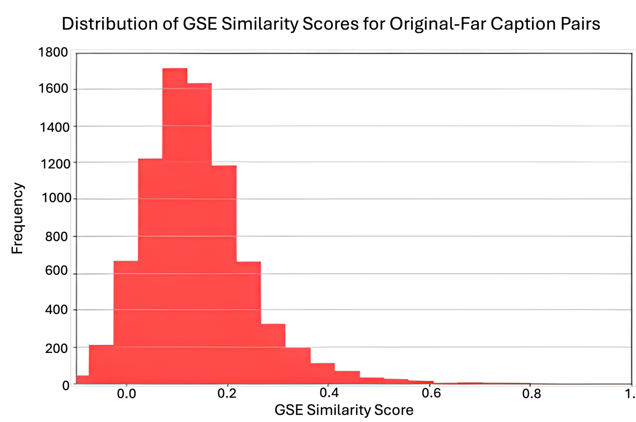

Fig. 3: The distribution of GSE similarity between the original captions and the extensively-modified captions is presented. The mean value of the distribution is approximately 0.1, indicating minimal overlap or shared semantic content between the two sets of captions.

2. ADDITIONAL RESULTS

In this section, we present the Percentage Change results when transferring the style of artwork images from WikiArt [4] to the targeted style.

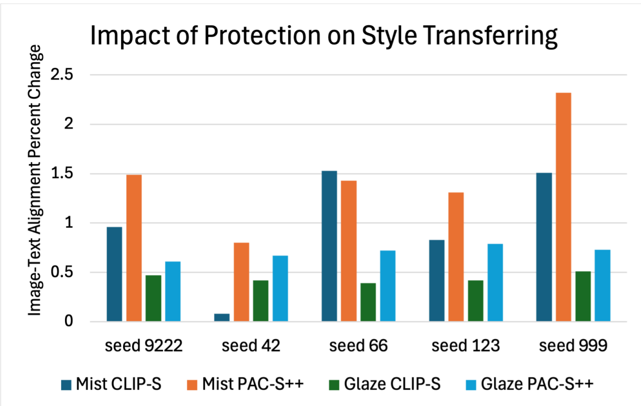

Fig. 4: Impact of protection on stylization generation on Artwork image domain. The diagram illustrates the percentage change across five generator seeds, evaluated under two protection methods (Glaze [5] and Mist [6]) and two ITA scoring metrics (CLIP-S [7] and PAC-S++ [8]).

3. ACTUAL CHANGE RESULTS FOR STYLIZATION ON NATURAL SCENE IMAGES

The Actual Change results for each style are shown in Tab 1. We analyze the frequency of cases where the Actual Change is negative or non-negative.

	ITAScore	Actual Change <0	Actual Change ≥ 0
Cubism	CLIP-S	6.87%	**93.13%**
	PAC-S++	14.87%	**85.13%**
Post-Impressionism	CLIP-S	34.00%	**66.00%**
	PAC-S++	25.25%	**74.75%**
Impressionism	CLIP-S	33.25%	**66.75%**
	PAC-S++	33.25%	**66.75%**
Surrealism	CLIP-S	27.37%	**72.63%**
	PAC-S++	42.37%	**57.63%**
Baroque	CLIP-S	21.75%	**78.25%**
	PAC-S++	31.75%	**68.25%**
Fauvism	CLIP-S	16.75%	**83.25%**
	PAC-S++	40.87%	**59.13%**
Renaissance	CLIP-S	17.25%	**82.75%**
	PAC-S++	32.62%	**67.38%**

Table 1: Actual Change results for stylization prompts that transfer to 7 styles under different ITAScore methods. Bold numbers are used to indicate the majority.

4. REFERENCES

[1] Micah Hodosh, Peter Young, and Julia Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[2] Anthropic, "Claude 3.5 sonnet," `https://www.anthropic.com/claude`, 2024.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al., "Universal sentence encoder for english," in *EMNLP*, 2018, pp. 169–174.

[4] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka, "Improved artgan for conditional synthesis of natural image and artwork," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 394–409, 2019.

[5] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," in *USENIX Security 23*, 2023, pp. 2187–2204.

[6] Chumeng Liang and Xiaoyu Wu, "Mist: Towards improved adversarial examples for diffusion models," *arXiv:2305.12683*, 2023.

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv:2104.08718*, 2021.

[8] Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara, "Positive-augmented contrastive learning for vision-and-language evaluation and training," *arXiv:2410.07336*, 2024.