SD4Match: Learning to Prompt Stable Diffusion Model for Semantic Matching

Supplementary Material

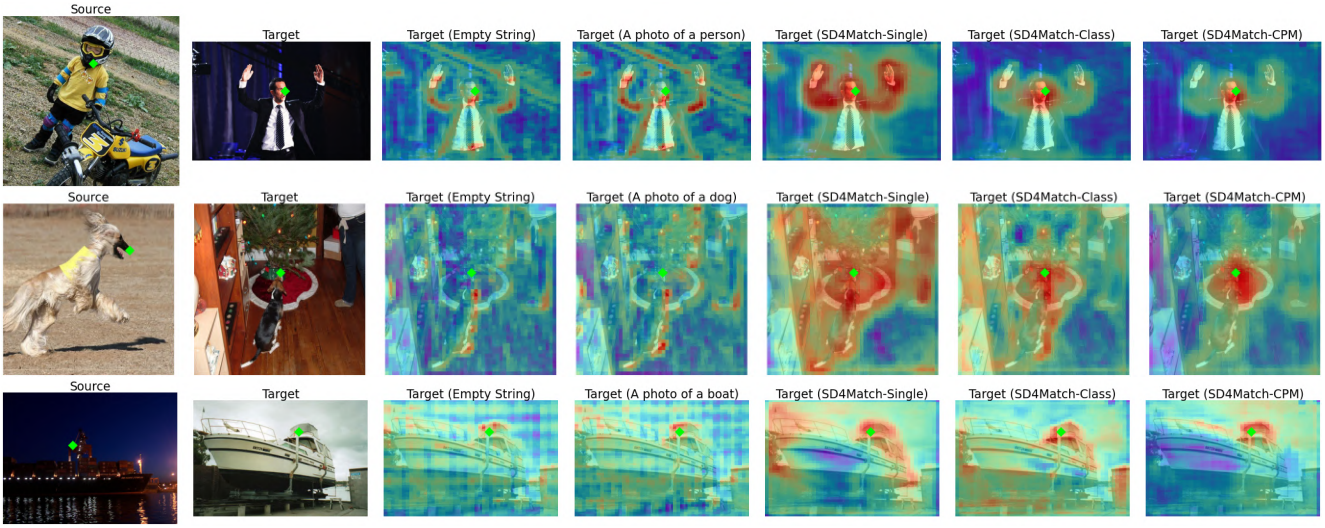

Figure 6. Correlation between the query feature and the target image with different prompts. Warmer colors indicate a higher correlation.

6. Discussion on Prompt Initialization

We discuss the impact of prompt initialization on matching accuracy in this section. We evaluate two extra initialization methods, an empty string and the text "*a photo of an object*", on SD4Match-Single as it is a generic configuration that does not require prior knowledge of the object category. We summarize their results on SPair-71k in Tab. 4. As demonstrated, the results of an empty string and "*a photo of an object*" are slightly better than the random initialization in the main manuscript. This validates the potential of our method and rooms for further improvement.

Table 4. Evaluation of different initialization methods on SPair-71k using SD4Match-Single.

Init. Method	SPair-71k @ $\alpha = 0.1$
"*a photo of an object*"	73.5
Empty String	73.4
Random Init.	72.6

7. Discussion on Image Size

We follow DIFT [50] and adopt the image size 768x768 during training and inference. We discovered that the discriminative power of SD deteriorates significantly if we deviate from such a size. For instance, on SPair-71k @ $\alpha = 0.1$, the accuracy of vanilla SD2-1 drops from 52.9 to 34.6 when the image size is halved to 384x384 and to 47.9 when doubled to 1536x1536. This is because SD2-1 is trained with the size of 768x768 so the semantic knowledge learned by SD2-1 is based on this size. Larger or smaller images will undermine the extraction of semantic information and provide a worse starting point for prompt tuning.

8. Visualization of Feature Correlation

To demonstrate the impact of different prompts on matching accuracy, we visualize the feature correlation using different prompts in Fig. 6. As shown, features extracted by learned prompts (SD4Match-Single, SD4Match-Class and SD4Match-CPM) are more discriminative in differentiating objects and backgrounds than textual prompts (Empty String and "a photo of a {*category*}"). This is attributed to the feature discriminative loss in Eq. (5) as it teaches the prompt to look for foreground objects. Moreover, SD4Match-Class and SD4Match-CPM have more localized capability than the SD4Match-Single, which shows the benefit of prior knowledge of the object category.

9. More Visualization

We provide more examples of images generated by per-class prompts in Fig. 7, images generated by conditional prompts in Fig. 8 and more qualitative comparison with baselines in Fig. 9.

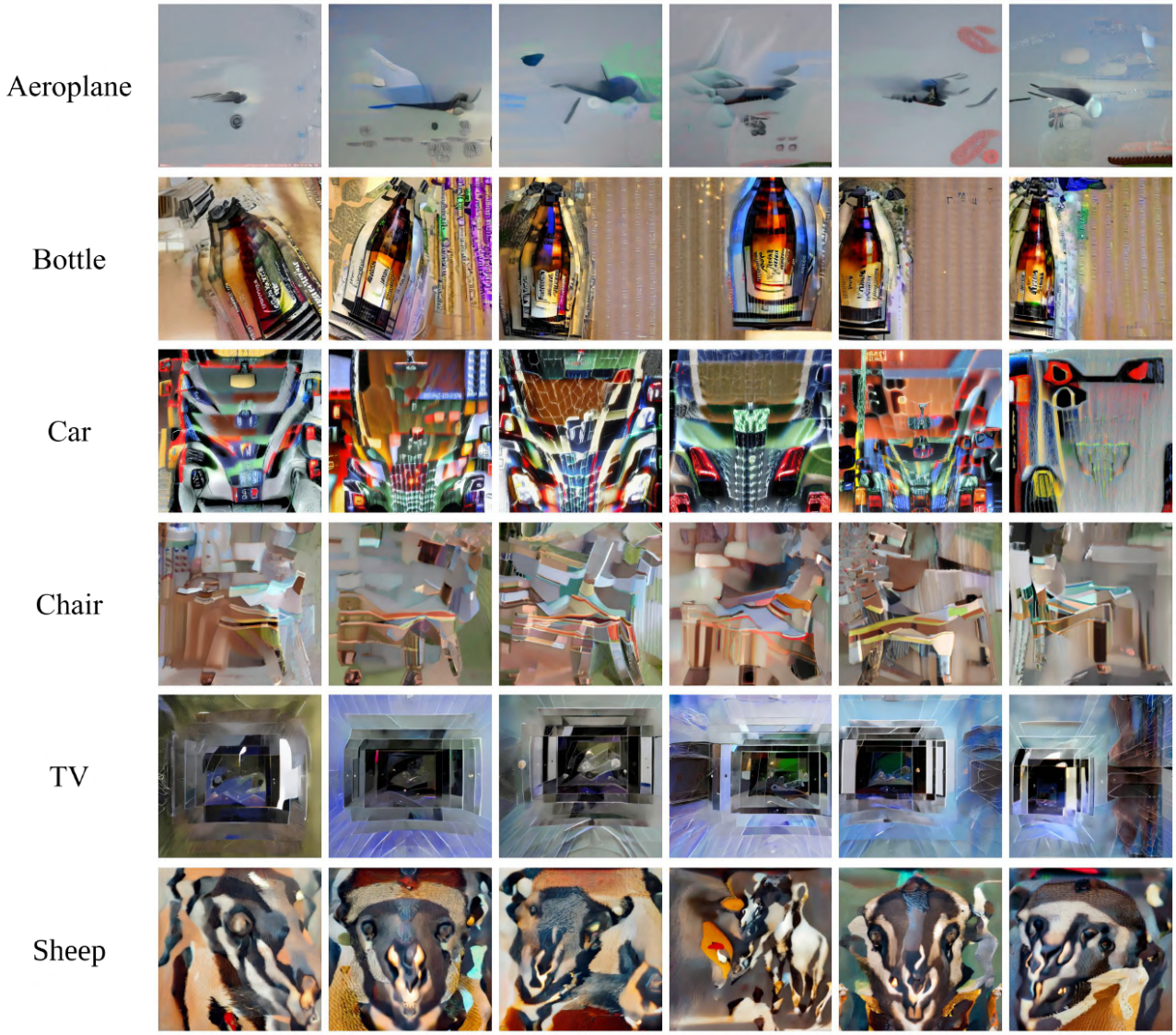

Figure 7. Visualizations of images generated by class-specific prompts learned by SD4Match-Class. For each object category, images are generated by the same prompt but with different random seeds.

Figure 8. Visualization of images generated by conditional prompts learned by SD4Match-Class. The generated image is an abstract illustration of the category of the shared object presented in Image A and Image B.

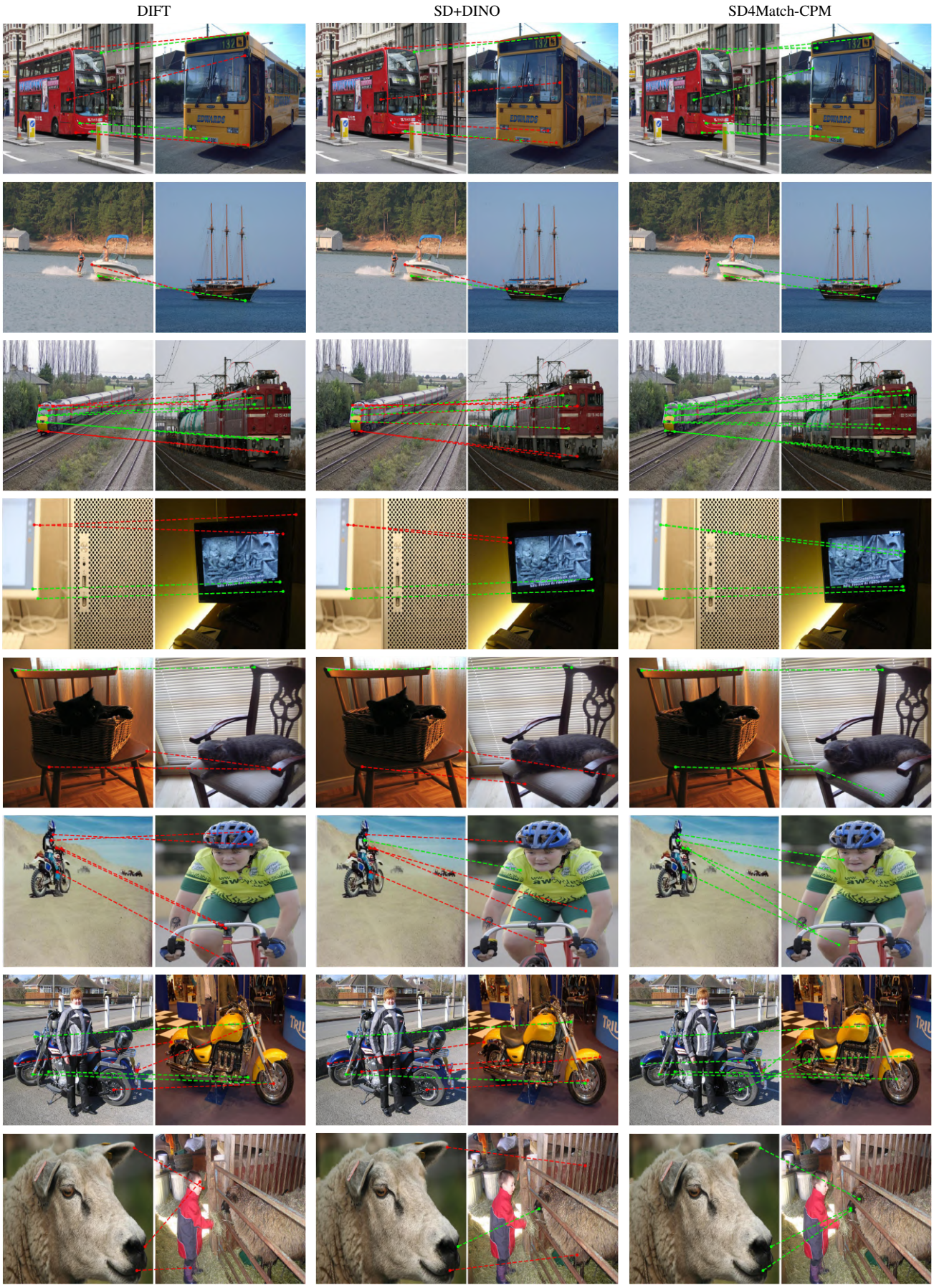

Figure 9. Qualitative comparison between DIFT, SD+DINO, and SD4Match-CPM.