Supp: Removing Distributional Discrepancies in Captions Improves Image-Text Alignment

Yuheng Li[1], Haotian Liu[1], Mu Cai[1], Yijun Li[2], Eli Shechtman[2], Zhe Lin[2], Yong Jae Lee[1], and Krishna Kumar Singh[2]

[1] University of Wisconsin-Madison
[2] Adobe Research

In this supplement, we will provide additional implementation details, present more results, and discuss limitations.

1 More Implementation Details

Data filtering classifer As stated in the main paper, to mitigate the bias arising from distribution discrepancies between positive and negative captions, which could influence the learning of vision-language models, we devised a strategy involving a text-only classifier. Our approach involves dividing the dataset into five equally sized partitions. In each iteration, we employ four partitions as the training set to fine-tune a DistilBERT model [5], with the fifth partition reserved as the test set. This cycle is repeated until every partition has been used as the test set. The fine-tuning parameters for the DistilBERT classifier are set to a learning rate of 2e-5 and a batch size of 16, over a duration of three epochs. Upon training, the classifier is applied to the hold-out test set to identify and filter out biased data, based on the classifier's prediction confidence. The data predicted correctly with the highest 30% confidence will be excluded.

BLIP2 finetuning In Section 4.6 of the main paper, we further demonstrate the generalizability of our curated dataset by fine-tuning BLIP2 [2] on it. Specifically, we fine-tuned BLIP2's Q-former using a batch size of 128 and a learning rate of 1e-5, across 1000 steps.

Group Score for MagicBrush Like the Winoground dataset, the MagicBrush consists of quartets that include one original image I_{ori}, one edited image I_{edt}, and their respective captions C_{ori} and C_{edt}. We define text score and image score as:

$$f(C_{ori}, I_{ori}, C_{edt}, I_{edt}) = \begin{cases} 1 & \text{if } s(C_{ori}, I_{ori}) > s(C_{edt}, I_{ori}) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

and

$$g(C_{ori}, I_{ori}, C_{edt}, I_{edt}) = \begin{cases} 1 & \text{if } s(C_{edt}, I_{edt}) > s(C_{edt}, I_{ori}) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

respectively, where $s(C, I)$ represents the scoring function used to evaluate the alignment between a text-image pair. The group score is defined as follows:

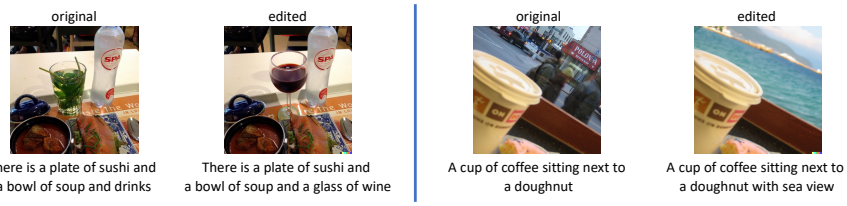

Fig. 1: Magicbrush dataset Often, the original caption remains consistent with the edited image.

Table 1: Comparison of our model against baselines across three prevalent challenges, highlighting the highest image score in bold.

	Attribute	Counting	Spatial
CLIP-ViT-L-14 [4]	70	60	56
NegCLIP [10]	76	62	62
BLIP2-ITC [2]	80	60	64
VisualGPT [1]	**98**	82	78
BLIP2-ITM [2]	90	82	58
Image-Reward [8]	94	76	58
VQ2 (BLIP-T5) [9]	90	78	60
LLaVA-1.5 [3]	90	80	80
Ours	**98**	**84**	**90**

$$h(C_{ori}, I_{ori}, C_{edt}, I_{edt}) = \begin{cases} 1 & \text{if } f(C_{ori}, I_{ori}, C_{edt}, I_{edt}) \text{ and } g(C_{ori}, I_{ori}, C_{edt}, I_{edt}) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Note that, contrary to the original definition from Winoground [7], we do not insist on $s(C_{edt}, I_{edt}) > s(C_{ori}, I_{edt})$ for the test score or $s(C_{ori}, I_{ori}) > s(C_{ori}, I_{edt})$ for the image score. This adjustment is made because, within the Magicbrush dataset, it is often observed that the original image's caption C_{ori} aligns well with the edited image I_{edt}, as demonstrated in Figure 1. Consequently, we do not penalize a model for assigning a high score to $s(C_{ori}, I_{edt})$.

2 More Results

Image Score for Attribute, Counting and Spatial Reasoning In Section 4.5 of the main paper, we specifically focus on the attribute, counting, and spatial reasoning capabilities of vision-language models by curating a custom dataset. This dataset is generated by first creating captions using GPT, which are then used to synthesize images via a Text-to-Image (T2I) model [6]. For each generated image pair, we manually select one image that aligns well with the caption (positive) and one that does not (negative). While the main paper presents our results in terms of accuracy, in this section, we introduce an image

score metric. For each text-images triplet, we assign a score of 1 if $s(C, I_{pos}) > s(C, I_{neg})$, and 0 otherwise. Table 1 shows our results. Again, our model achieves the best performance across all datasets.

Contrastive Training The results we have discussed so far are based on models trained with the original cross-entropy loss. Inspired by the CLIP [4], we have also attempted to fine-tune the LLaVA-1.5 model using our data with a contrastive loss approach. However, we empirically found that the contrastive loss was not as effective as anticipated: (67, 54.24, 46.5) (88.9, 77.3, 84.2) (95.5, 94.8, 97.6) (86.98) (compare with last row in Table1 in main paper). We hypothesize that the current cross-entropy loss might produce a similar effect to contrastive loss when dealing with two types of data (positive vs negative), and contrastive loss might be more beneficial when there is a spectrum of "negativeness," e.g., where some pairs are completely unaligned and others are only slightly incorrect. To explore this, we created three types of data: positive (aligned), slightly unaligned (the curated data in our paper), and completely unaligned (randomly shuffled image-text pairs), aiming for progressively lower scores across these categories. However, this approach proved unhelpful, as the last case was too straightforward for LLaVA to discern, providing no training signal. Thus, exploring how to create a spectrum of data alignment levels and studying if they are useful remains an interesting future research.

Qualitative Results on Winoground Figure 2 presents qualitative results for the Winoground dataset, comparing our model against the strongest baseline, LLaVA-1.5 [3], and the recently introduced VQ2 [9]. The outcomes demonstrate that our model exhibits superior compositional understanding abilities.

3 Discussion

In this paper, we introduce a novel approach for generating high-quality training data for image-text alignment models. However, our method has some limitations. Firstly, it assumes access to ground truth positive captions for images, which may not be feasible on a large scale. A promising avenue for overcoming this hurdle could involve leveraging vision-language models such as LLaVA for generating image captions, presenting an exciting direction for future research to enhance scalability. Additionally, our method may inherit certain constraints from the LLaVA image encoder, which utilizes CLIP and is tailored to process images of a fixed, relatively low resolution. Consequently, this may limit the model's ability to detect small objects or capture fine-grained details effectively.

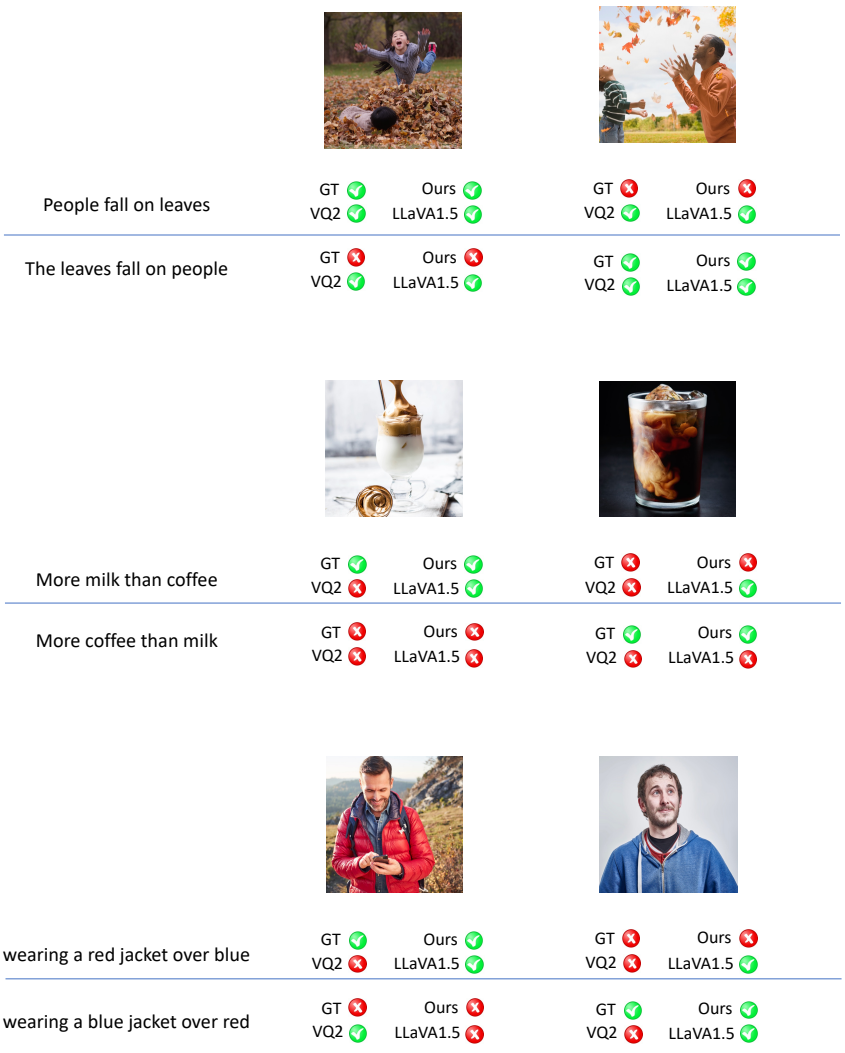

Fig. 2: Qualitative Results on Winoground

References

1. Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M.: Visualgpt: Data-efficient adaptation of pretrained language models for image captioning (June 2022)
2. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:256390509
3. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
5. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019), https://api.semanticscholar.org/CorpusID:203626972
6. stabilityai: IF repository. https://github.com/deep-floyd/IF
7. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
8. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
9. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., Szpektor, I.: What you see is what you read? improving text-image alignment evaluation. arXiv preprint arXiv:2305.10400 (2023)
10. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=KRLUvxh8uaX