

# Supplementary Material: Overcoming the Pitfalls of Vision-Language Model for Image-Text Retrieval

Paper ID 1163

In this supplementary material, we will further detail the following aspects omitted in the main paper.

- **Section A:** More quantitative comparison results, including integrating the proposed model with SAF [1], SGR [1], SGRAF [1], and BLIP [2].
- **Section B:** More ablation results, including the influence of various loss styles and different fusion strategies.
- **Section C:** More qualitative results, including the visualization of patch-wise similarity maps and the modality gap, which further demonstrate the effectiveness of the proposed approach.

## A MORE COMPARISON RESULTS

To validate the effectiveness of our proposed Local and Generative-driven Modality Gap Correction (LG-MGC), we further integrate it with four image-text retrieval baselines, namely SAF [1], SGR [1], SGRAF [1], and BLIP [2]. Among them, the first three baselines (i.e., SAF, SGR, and SGRAF) are the standard image-text retrieval models without pre-training, whereas BLIP [2] is a pre-trained cross-modal alignment model. The results on the Flickr30K and MS-COCO are summarized in Table 1, and our results are annotated with purple background. To be specific, for each baseline method, we present the original result reported in the paper (if given), our re-implemented result, and the result obtained by adding our module (i.e., +LG-MGC). From Table 1, we can observe that our re-implemented results are very close or even higher to the original performances, which ensures the credibility of our experimental results. Moreover, it is clear that the proposed method could enhance the performance of all baselines on both the image-to-text retrieval and text-to-image retrieval tasks. Specifically, on the Flickr30K dataset, models augmented with our +LG-MGC exceed the baseline models with satisfactory improvements from 2.9% to 4.1% in terms of RSUM. On the MS-COCO dataset, our approach can also achieve 1.3% to 4.3% performance gains. These improvements indicate that our approach is helpful for cross-modal alignment, benefiting from

the local and generative-driven semantic learning. A closer look of the results displayed at the bottom of Table 1 reveals that our approach remains effective even for the more complex and large model BLIP [2], which achieves a performance improvement of 2.9% on the Flickr30K dataset and 4.3% on the MS-COCO dataset in terms of RSUM. It is worth noting that integrating our *LG-MGC* module with the baselines does not increase the number of trainable parameters, underscoring the efficiency of our proposed method. These results show the advantage of our novel designing again.

## B MORE ABLATION RESULTS

**Influence of Different Loss Styles.** In Table 2, we investigate three different distance measurement methods (i.e., *KL*, *L2*, and *InfoNCE*) within our Generative-driven Semantic Translation (GST) module, as detailed in Eq. (8) of our main paper. Through the comparison among the *Baseline* (i.e., CLIP<sub>Vit-B/16</sub> [4]), *Baseline+GST<sub>KL</sub>*, *Baseline+GST<sub>L2</sub>*, and *Baseline+GST<sub>InfoNCE</sub>*, we can observe that the models with the *GST* module consistently exceed the *Baseline* with clear improvements. This demonstrates that the generative-driven semantic translation can indeed provide an effective signal to supervise the cross-modal alignment. Moreover, the similar performances of *Baseline+GST<sub>KL</sub>*, *Baseline+GST<sub>L2</sub>*, and *Baseline+GST<sub>InfoNCE</sub>* indicate that our approach is robust across different metric learning strategies. Among them, the *InfoNCE* loss slightly improves the performance over the *KL* and *L2* by 1.9% and 1.7%, respectively, in terms of RSUM. Consequently, we opt for the *InfoNCE* loss in our generative-driven semantic translation module.

Table 2: Performance variation with different losses in GST.

Eval Task → Loss Type ↓	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
Baseline	88.4	98.7	99.5	76.1	94.6	97.2	554.5
+GST <sub>KL</sub>	91.8	99.1	99.6	78.9	95.4	97.8	562.6
+GST <sub>L2</sub>	91.5	98.9	99.7	79.0	95.7	98.0	562.8
+GST <sub>InfoNCE</sub>	92.6	99.5	99.7	78.9	95.5	98.2	564.5

Table 1: Comparisons with state-of-the-art methods on Flickr30k and MSCOCO. † denotes the improved results by the authors compared to the original paper.

Data Split →	Flickr30K (1K)						MS-COCO (5K)							
Eval Task →	Image-to-Text			Text-to-Image			RSUM	Image-to-Text			Text-to-Image			RSUM
Method ↓	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
(Faster-RCNN, ResNet-101, BiGRU, without pre-training)														
SAF <sub>(CVPR'21)</sub> [1]	73.7	93.3	96.3	56.1	81.5	88.0	488.9	53.3	-	90.1	39.8	-	80.2	-
SAF <sup>†</sup>	74.0	92.9	97.4	56.2	81.8	88.8	491.1	55.7	83.3	91.0	40.1	69.6	80.3	420.0
+LG-MGC (Ours)	74.5	93.7	97.6	57.4	82.6	89.1	<b>494.9</b>	56.0	83.3	91.1	40.3	70.1	80.5	<b>421.3</b>
SGR <sub>(CVPR'21)</sub> [1]	75.2	93.3	96.6	56.2	81.0	86.5	488.8	56.9	-	90.5	40.2	-	79.8	-
SGR <sup>†</sup>	74.9	93.2	96.7	56.0	81.6	87.6	490.0	57.5	83.9	91.2	40.3	69.6	80.4	422.9
+LG-MGC (Ours)	77.0	93.0	96.8	56.4	82.4	87.6	<b>493.2</b>	58.1	83.8	91.3	40.6	69.9	80.8	<b>424.5</b>
SGRAF <sub>(CVPR'21)</sub> [1]	77.8	94.1	97.4	58.5	83.0	88.8	499.6	57.8	-	91.6	41.9	-	81.3	-
SGRAF <sup>†</sup>	77.7	94.3	97.4	58.5	82.9	89.3	500.1	59.3	85.1	91.8	41.8	71.2	81.7	430.9
+LG-MGC (Ours)	79.2	94.6	97.4	59.0	84.2	89.8	<b>504.2</b>	59.8	85.5	92.1	42.5	71.7	81.9	<b>433.5</b>
(Dual-Encoder, pre-training)														
BLIP <sub>(ICML'22)</sub> [1]	96.6	99.8	100.0	87.2	97.5	98.8	579.9	80.6	95.2	97.6	63.1	85.3	91.1	512.9
BLIP <sup>†</sup>	96.2	99.7	100.0	86.2	97.4	98.9	578.4	80.4	95.0	97.5	62.8	85.3	91.2	512.2
+LG-MGC (Ours)	97.1	99.9	100.0	87.8	97.4	99.1	<b>581.3</b>	81.3	95.7	98.2	63.8	86.0	91.5	<b>516.5</b>

**Influence of Different Fusion Strategies.** As described in Eq. (4) and Eq. (5) of Section 3.2 in our main paper, the proposed Local-driven Semantic Completion (LSC) module concatenates the global visual and textual representations with selected local features. Therefore, we leverage different fusion strategies to fuse the global and local information, including max pooling (i.e.,  $+LSC_{Max}$ ), mean pooling (i.e.,  $+LSC_{Mean}$ ), and concatenation (i.e.,  $+LSC_{Concat}$ ). The results are presented in Table 3. From the table, it is apparent that all fusion strategies could achieve improved performances, with increases ranging from 6.7% to 10% in terms of the RSUM. These findings indicate that our local-driven semantic completion module enhances the model’s ability to learn specific fine-grained visual and textual concepts, thereby improving the cross-modal alignment. Besides, among these variants, we observe that the *Baseline*+ $LSC_{Concat}$  achieves the highest performance. Consequently, we select the concatenation fusion strategy for our LSC module.

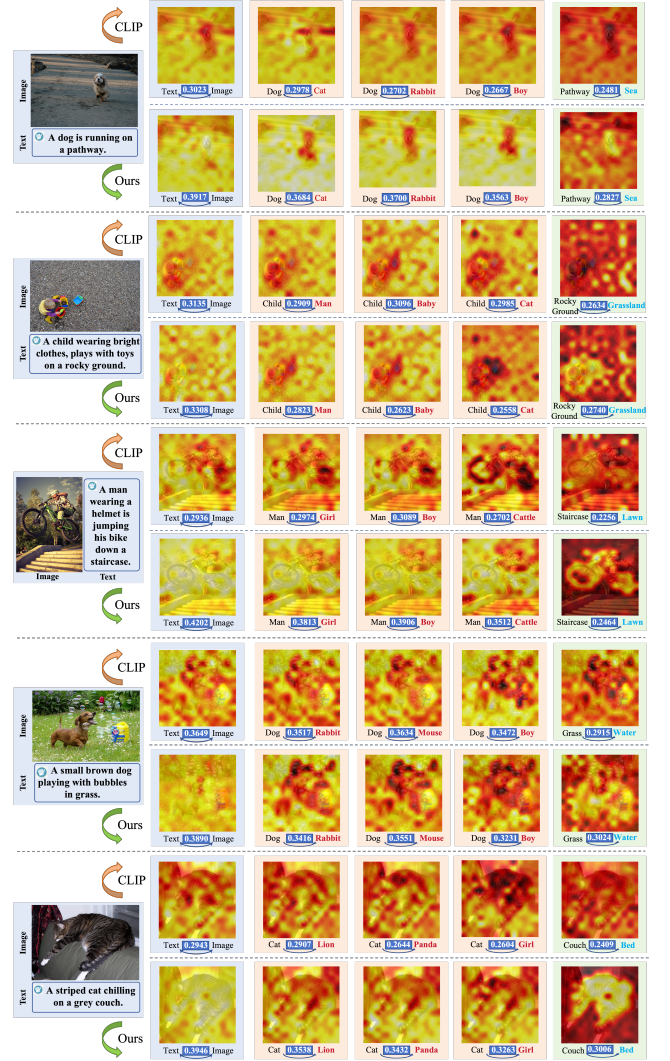
**Table 3: Performance variation with different fusion strategies in GST.**

Eval Task →	Image-to-Text			Text-to-Image			RSUM
Loss Type ↓	R@1	R@5	R@10	R@1	R@5	R@10	
Baseline	88.4	98.7	99.5	76.1	94.6	97.2	554.5
+ $LSC_{Max}$	91.4	99.3	99.8	77.9	95.0	97.8	561.2
+ $LSC_{Mean}$	92.1	99.4	99.9	77.4	95.4	97.9	562.1
+ $LSC_{Concat}$	92.6	99.5	99.7	78.9	95.5	98.2	564.5

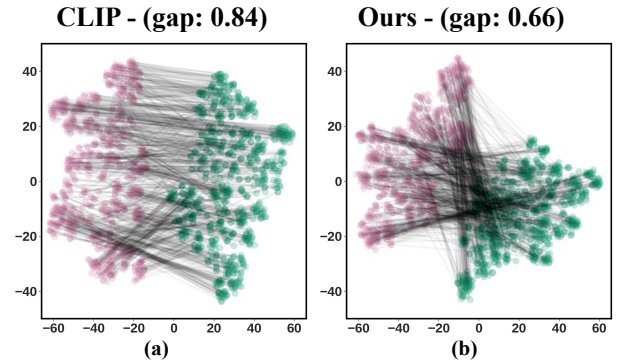
## C MORE VISUALIZATION RESULTS

**Visualization of the Patch-Wise Similarity Maps.** To qualitatively verify the effectiveness of the proposed method in integrating fine-grained local features into the global embeddings, we present more visualization results for the Flickr30K and MS-COCO datasets. To be specific, following Yi et al. [5], we calculate the similarity between each image patch and the global textual representation, which could be interpreted as the contribution of each patch feature to the overall global feature. For each sample, we visualize the similarity maps based on both the vanilla CLIP model [4] and our proposed approach, and the results are displayed in Fig. 1. From left to right, the results represent the similarity map between each image patch and the original textual feature, with the text replaced fine-grained local visual regions, and with the text replaced dominate background descriptions, respectively. Note that, for better visibility, each pixel in the patched similarity map has been smoothed using its surrounding 16 pixels. Moreover, in the figure, the brighter color indicates higher similarity, whereas the darker color denotes lower similarity.

According to the results, one could have the following observations and conclusions. First, from the qualitative results displayed in the first column of Fig. 1, we can observe that the similarity maps generated by our proposed method are generally brighter than those computed using the *CLIP* model. The results indicate that the representation learned from our method is capable of capturing comprehensive visual information, including both dominate background elements and fine-grained visual details, such as the *dog*, *child*, *man*, and so on. Compared to our method, the baseline model, CLIP, tends to focus on the dominated scenes (i.e., background). Correspondingly, compared to CLIP, our method consistently achieves a higher similarity (indicated by the blue rectangle) between the



**Figure 1: Similarity maps from the vanilla CLIP and our proposed model on both Flickr30K and MS-COCO test set. The brighter the color, the higher the value.**



**Figure 2: Visualization of the modality gap on MS-COCO for the CLIP (a) and our proposed method (b).**



image and text. Second, by comparing the results from the second to the fourth column with that shown in the first column, we can observe that the similarity maps based on the CLIP show almost no change, indicating that the baseline is insensitive to local changes. In contrast, when we replace the word *Dog* in the text with *Cat*, *Rabbit*, and *Boy* in the first example, our method clearly shifts its focus away from the image areas that contain *Dog*. Similar patterns can also be observed in the additional examples provided in Fig. 1. This phenomenon indicates that the proposed approach could indeed enable the model to more effectively focus on minute visual areas, thereby enhancing the model's fine-grained alignment capabilities. Third, by comparing the results of the fifth column with that shown in the first column, one can observe that both the baseline and our proposed method are very sensitive to the text changes concerning the background. However, unlike the baseline, which almost ceases to focus on any specific area of the image (i.e., all areas in the image become darker), the proposed method remains attentive to the visual regions about other object descriptions in the text, such as the *dog*, *child*, *man*, *bike*, and so on. Overall, the similarity maps further confirm that the vanilla CLIP model is insensitive to local changes, while through explicitly and implicitly integrating local information into the global representation, our method can enhance the model to learn more comprehensive representations.

**Visualization of the Modality Gap on MS-COCO.** To testify whether our method can alleviate the heterogeneous modality gap, we further compute the cross-modal distance on the MS-COCO dataset and present the results in Fig. 2. Specifically, following the recipe from [3], given 1000 image-text pairs from the test set of MS-COCO dataset, we first calculate the similarity between

different samples based on the euclidean distance (i.e., the distance between the blue dots and orange dots in Fig. 2). Then, the modality gap is assessed by calculating the difference between the centers of image embeddings and text embeddings by  $\Delta_{gap} = \frac{1}{n} \sum_{i=1}^n f_{cls_i}^v - \frac{1}{n} \sum_{i=1}^n f_{eos_i}^t$ , and the results are shown in the top of Fig. 2. From the comparison between Fig. 2 (a) and (b), we can observe that the proposed model could clearly reduce the gap between different modalities on the MS-COCO dataset. The results demonstrate that our local and generative-driven modality gap correction model is both superior and effective. It successfully mitigates the representation disparities between vision and language, thereby boosting the cross-modal retrieval performance.

## REFERENCES

- [1] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 1218–1226.
- [2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International conference on machine learning (ICML)*. 12888–12900.
- [3] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*. 17612–17625.
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 8748–8763.
- [5] Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, and Hongtao Lu. 2023. A simple framework for text-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7071–7080.