ViLEM: Visual-Language Error Modeling for Image-Text Retrieval Supplementary Materials

In this supplementary materials, we elaborate on (1) details about pre-training datasets, downstream datasets, and evaluation metrics of downstream tasks; (2) Experimental setting for inference time measurement on MSCOCO dataset; (3) Ablation study for probability of each word being edited; (4) More visualizations about knowledge-based text edition and image-text retrieval.

Table 1. Statics of the pre-training datasets.

	COCO (Karpathy-train)	VG	CC3M	SBU	CC12M
image	113K	100K	2.81M	825K	8.78M
text	567K	769K	2.81M	825K	8.78M

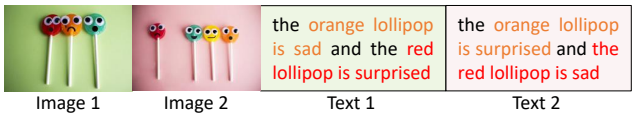

Figure 1. A sample from Winoground dataset, which contains two image-text pairs with minor differences.

A. Datasets Details

Pre-training datasets. We show the statistics of the images and texts of pre-training datasets in the Table 1

MSCOCO. MSCOCO [5] is a large image-text dataset of 123K images, where each image has 5 human-annotated captions. Following [3, 4, 6], we adopt the Karpathy split of MSCOCO, where 5K/5K/113K images are used for testing, validation and training respectively.

Flickr30K. Flickr30K contains 31K images and 159K captions. Each image is usually annotated with 5 captions. Following [2], we 1K/1K/29K images for testing, validation and training respectively.

Winoground. Winoground consists of 400 test cases and each case has two image-text pairs. As shown in Figure 1, the two image-text pairs of each case only have minor object or/and relation differences between them, which requires the model to be sensitive to local compositional semantics in the images and texts,

Table 2. Performance and inference time comparisons with our method, VinVL-base and ALBEF.

Method	image→text			text→image			R@S	Time/s
	R@1	R@5	R@10	R@1	R@5	R@10		
VinVL-base [8]	74.6	92.6	96.3	58.1	83.2	90.1	494.9	1.05×10^6
ALBEF [4]	73.1	91.4	96.0	56.8	81.5	89.2	488.0	9360
Ours	73.2	91.8	95.9	54.5	80.6	88.2	484.2	145

B. Evaluation Metrics

Retrieval. We report the widely-used R@k (k=1,5,10) for cross-modal retrieval, which is the proportion of matched samples found in the top-k retrieved results. We also report R@S to reveal the overall performance, which is defined as the sum of R@k metrics at k=$\{1,5,10\}$ of both image-to-text and text-to-image retrieval tasks. Following [1, 6], we also report the Median Rank (MedR) for video-text retrieval.

Vision-linguistic Stress Testing. We report the text score for experiments on Winoground dataset following [7], which measures whether a model can select the correct caption given an image. Given images I_0 and I_1 and captions T_0 and T_1, the text score for an example (T_0, I_0, T_1, I_1) is computed according to:

$$f(T_0, I_0, T_1, I_1) = \begin{cases} 1, & \text{if } s(T_0, I_0) > s(T_1, I_0) \\ & \text{and } s(T_1, I_1) > s(T_0, I_1) \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $s(\cdot)$ is the model's score for the image-text pair. This metric tests whether the ground truth text for a given image in Winoground dataset is scored higher than the alternative text *and* whether this hold for the other image-text pair in the example too.

C. Inference Time Measurement

Following [6], we evaluate all methods on the MSCOCO (5K) dataset with a single Tesla V100 GPU and the test batch size is set to 64. As shown in Table 2, the inference time of our method, ALBEF [4] and VinVL-base [8] are 145s, 9360s and 1.05x10^6s respectively. Our method achieves comparable performance with these "joint-encoder" methods but has much faster inference speed.

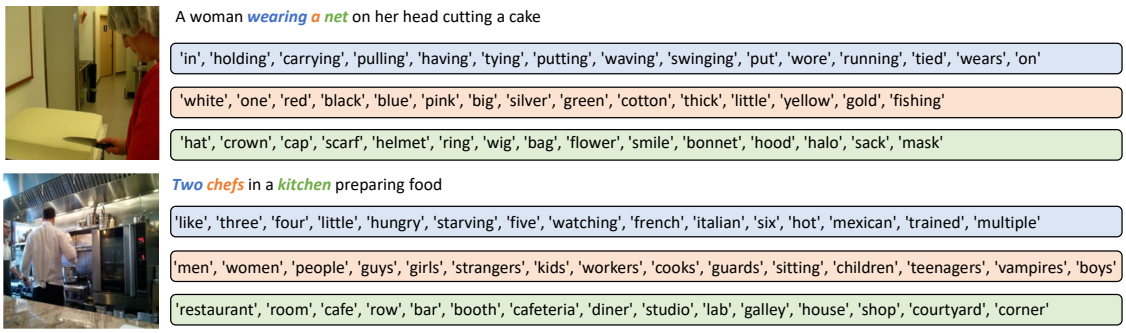

Figure 2. Examples of knowledge-based word edition. Different colored words in the texts are the words to be edited and the top-15 candidate words generated by BERT are shown in the corresponding colored boxes.

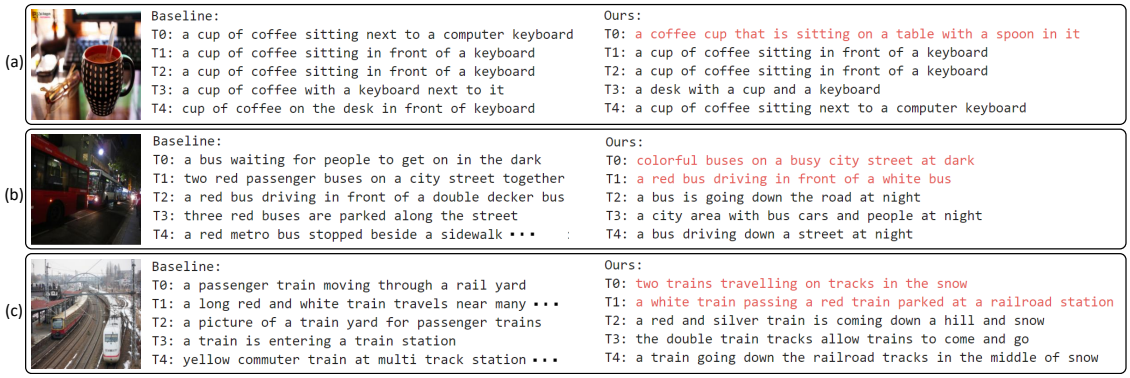

Figure 3. Illustration of image-to-text retrieval of our model and baseline model. Ground-truth captions for each image are in red color.

D. More Visualizations

Knowledge-based Word Edition. We show the candidate words generated by BERT in the Figure 2. It can be seen that we can generate diverse candidate words that are plausible but visual-incorrect, which provides useful text errors for our novel proxy task ViLEM. We can also observe that some candidate words are visual-correct, *e.g.*, "wearing"→"in" in the first image and "chefs"→"men" in the second image. But the number of these correct candidate words is relatively small, ensuring the effectiveness of our method.

Image-Text Retrieval. We show image-to-text retrieval results on the MSCOCO test set in the Figure 3. We can observe that (1) our model can capture local visual information of "spoon" in (a) while the baseline model ignores it; (2) Our model correctly recognizes the number and color differences of buses in (b), as well as the number of trains in (c), indicating that our model has a more comprehensive perception of images than the baseline model.

The text-to-image results are shown in Figure 4. It can be seen that (1) our model perceives local semantics more accurately, *e.g.*, "a metal set of bars" in (a); (2) Our model considers detailed local text semantics and find the image that contains both "tv" *and* "a bunch of chairs" in (b), but the baseline model only finds the images with only "tv" *or* "a bunch of chairs"; (3) Our model has better understanding on the complex relation, *e.g.*, "ipod cases *on* their computer screen", while the baseline model finds images with "ipod" and "screen".

E. Ablation Study

Word editing probability. We conduct an ablation study on the probability to edit word tokens. As shown in Table 3, we increase the word editing probability from 15% to 45% and observe that the retrieval performance drops gradually. When the word editing probability is set to 80%, we observe further performance degradation. We argue that a high word editing probability will cause drastic changes in text semantics and reduce the difficulty of discriminating the correctness of words, preventing the learning of local semantics of images and texts and association between them.

Selection of prepositions or articles. Editing some prepositions and articles may introduce noise. But there are also many prepositions describing spatial relations and editing them generates spatial relation errors (*e.g.*, "on"→"beside"). Similarly, editing articles can lead to number errors (*e.g.*, "a"→"two"). Correcting these errors help model understand spatial relation and number. The re-

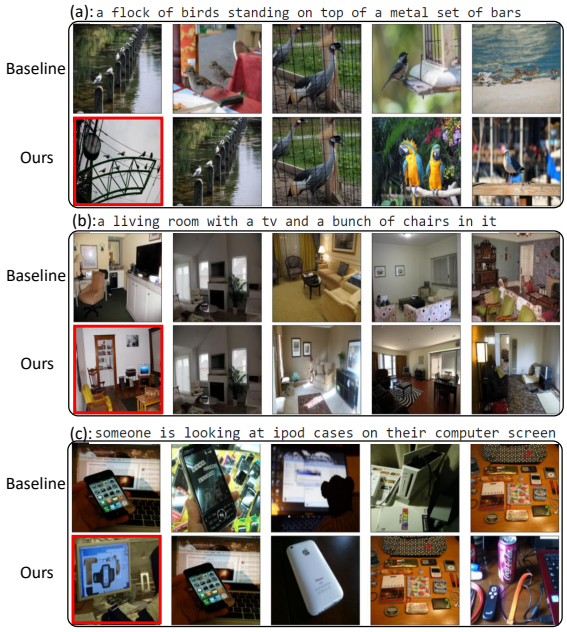

Figure 4. Illustration of text-to-image retrieval results of our model and baseline model. The ground-truth image for each text is in the red box.

Table 3. Ablation study on the word editing probability.

Prob	image→text			text→image			R@S
	R@1	R@5	R@10	R@1	R@5	R@10	
0.15	**29.1**	55.3	**68.3**	**22.0**	**45.7**	**57.7**	**278.1**
0.30	29.0	**55.4**	67.3	21.8	**45.7**	**57.7**	276.9
0.45	28.3	55.3	68.1	21.3	45.6	57.1	275.7
0.80	27.2	54.4	66.3	20.9	44.7	56.7	270.2

Table 4. Ablation study on the word selection.

Method	image→text			text→image			R@S
	R@1	R@5	R@10	R@1	R@5	R@10	
w/o pre. & art.	28.8	55.2	68.3	21.9	45.6	57.5	277.3
w/o synonyms	29.0	55.4	68.5	21.9	45.9	57.7	278.4
Ours	29.1	55.3	68.3	22.0	45.7	57.7	278.1

sult of excluding prepositions or articles are shown in Table 4 and we can observe that it achieves lightly worse results (278.1→277.3 on R@S).

Probability of selecting synonyms. We also approximately compute the probability of selecting synonyms to replace origin word with the help of WordNet. We construct synonyms set for each word via WordNet, and count that 0.3M/7.16M words (4.3%) are replaced by synonyms in CC4M dataset. Moreover, we experiment with excluding synonyms during text edition. As shown in Table 4, synonyms have little impace on performance (278.1 *vs.* 278.4 on R@S).

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1

[2] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 1

[3] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1

[4] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[6] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15692–15701, 2022. 1

[7] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1

[8] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 1