Supplementary Material for "Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles"

Anonymous CVPR submission

Paper ID 8531

This appendix is organized as follows:

- In Section 1, we further illustrate the commonsense lacking issue by providing additional comparison of fundamental VL datasets with commonly used NLP data.
- In Section 2, we provide more visualizations of success examples of our method on the proposed diagnostic benchmark for both text-image and image-text retrieval.
- In Section 3, we provide more visualizations of success examples of our method on the OK-VQA benchmark.
- In Section 4, we summarize the statistics of the proposed diagnostic test data.
- In Section 5, we study the failure case of our DANCE augmented model.
- In Section 6, we report additional details of human study and implementation of DANCE.

1. Commonsense in Fundamental VL Data vs NLP Data

1.1. Distribution Comparison Between Current VL Data and More Natural Language Data

We further explore the commonsense lacking issue in the current fundamental VL data by comparing them with common natural language processing (NLP) data. Here we compare the distributions of the syntactic categories and words of the most popular VL datasets (COCO [8] and CC 12M [1]) with three commonly used NLP datasets: ConceptNet [15] the knowledge base dataset, Wikipedia [4] the popular [2, 6, 10, 13, 18] cleaned English-language articles with the size of 16GB, C4 [12] the popular used [3, 5, 7, 11, 14, 16, 17] English-language text sourced from the Common Crawl web scrape with the size of 745GB. The syntactic categories and word distributions comparison is shown in Fig. 1.

The upper part of Fig. 1 shows the distribution of the most frequent part-of-speech (POS) tags with punctuation marks excluded, and the lower part shows the most frequent word tokens. There is a significant difference between top POS tag/word token distributions of VL datasets compared with those of the regular texts. Similar to our observation in the main paper, the most frequent words in the text in existing VL datasets are nouns (NOUN) for **individual entities**, like "*street*", "*table*", "*train*". In contrast, all the NLP datasets have apparently more verbs (VERB), like "*have*", "*used*", "*find*", "*want*", "*happen*" that contains richer information about the **relationship between entities**. Besides, the NLP datasets include more particles (PRT), like "*to*", and pronouns (PRON) like "*your*", which are associated with **interconnection** information. This further illustrates the lacking commonsense issue in the fundamental VL datasets.

While the implicit information about the **interconnections between entities** is in high demand for developing commonsense and reasoning ability, the fundamental VL datasets are lacking it. This motivates us to use commonsense knowledge to improve VL data. In addition, the distribution of ours training data is also included for comparison. We can see that our data is similar to NLP data in terms of the interconnection between entities.

1.2. ChatGPT-Based Commonsense Measurement of VL Data and Natural Language Data

We further design to utilize the ChatGPT's powerful in-context learning to measure the amount of commonsense information for each dataset. We have the score ranges 0-10, and scores descriptive and general language equally, by providing

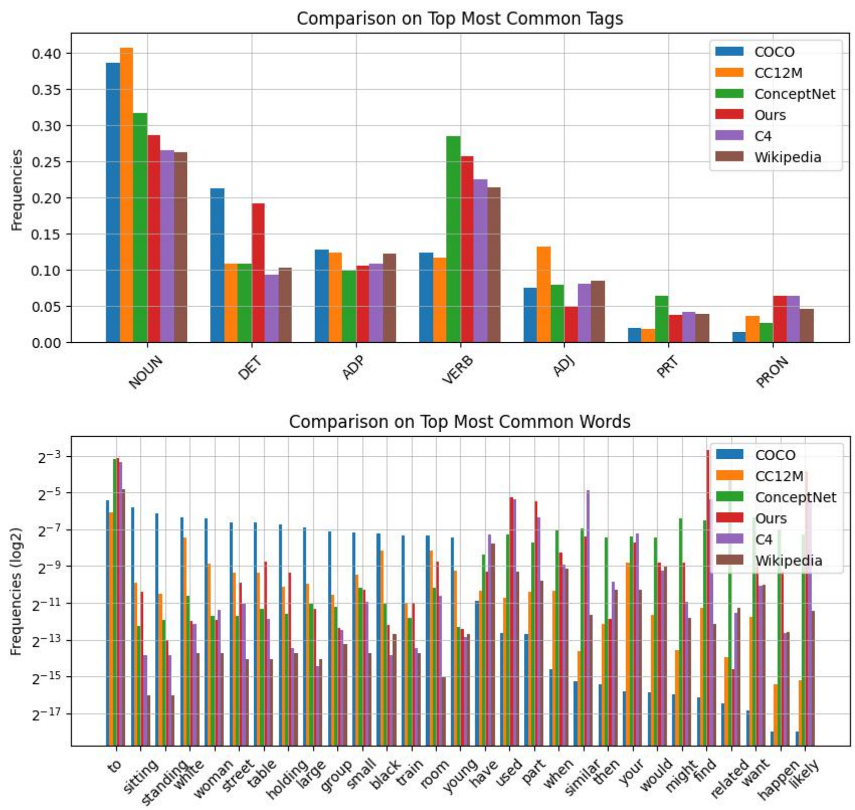

Figure 1. Comparison of the syntactic categories and words distributions of fundamental VL data (COCO [8] and CC12M [1]), ours training data generated by DANCE, and commonly used NLP data (ConceptNet [15], Wikipedia [4] and C4 [12]). Commonsense is lacking in VL data compared with NLP data, and is improved by DANCE strategy.

well-chosen text-score exemplars [9]. Then, 1K sentences (as no official API is available yet and unofficial approaches are limited) from each dataset are merged, shuffled and fed into it. Our entity names are put back. VL data COCO and CC12M still scored much lower than language data and ours even after minimizing the impact of language style, which further confirms the VL data's lack of commonsense.

Data	Wikipedia	C4	Ours	ConceptNet	CC12M	COCO
ChatGPT Score	7.53	6.82	8.03	8.26	4.91	5.25

Table 1. Comparison of the commonsense information amount of VL dataset and natural language dataset by ChatGPT.

2. Additional Qualitative Results on Our Diagnostic Benchmark

In Fig. 2 and Fig. 3, we show additional qualitative comparison with the state-of-the-art VL-models on our diagnostic test set for text-image and image-text retrieval respectively. In Fig. 2, from left to right is the input text, the input images including a correct one (in blue) and two incorrect ones (in red), the scores by each individual model, and the commonsense knowledge from the knowledge graph [15] that required for retrieval. In Fig. 3, from left to right is the input image, the input texts including a correct one (in blue) and two incorrect ones (in red), the scores, and the related commonsense knowledge from the knowledge graph. We can see that all the baselines fail to identify the correct answers, which further illustrates the lacking of commonsense ability in the popular VL-models. In contrast, our DANCE pre-trained model successfully retrieves the correct ones. We note that all these images and the knowledge are held out from the training set. This further demonstrates the reasoning ability enhanced by our DANCE strategy.

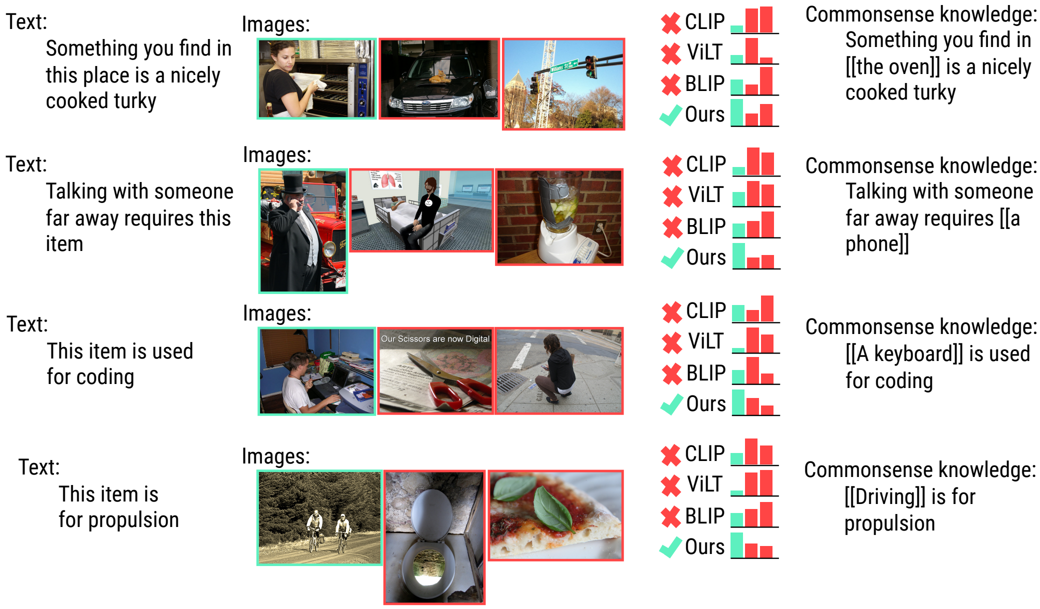

Figure 2. Qualitative examples from our diagnostic test set for text-image retrieval.

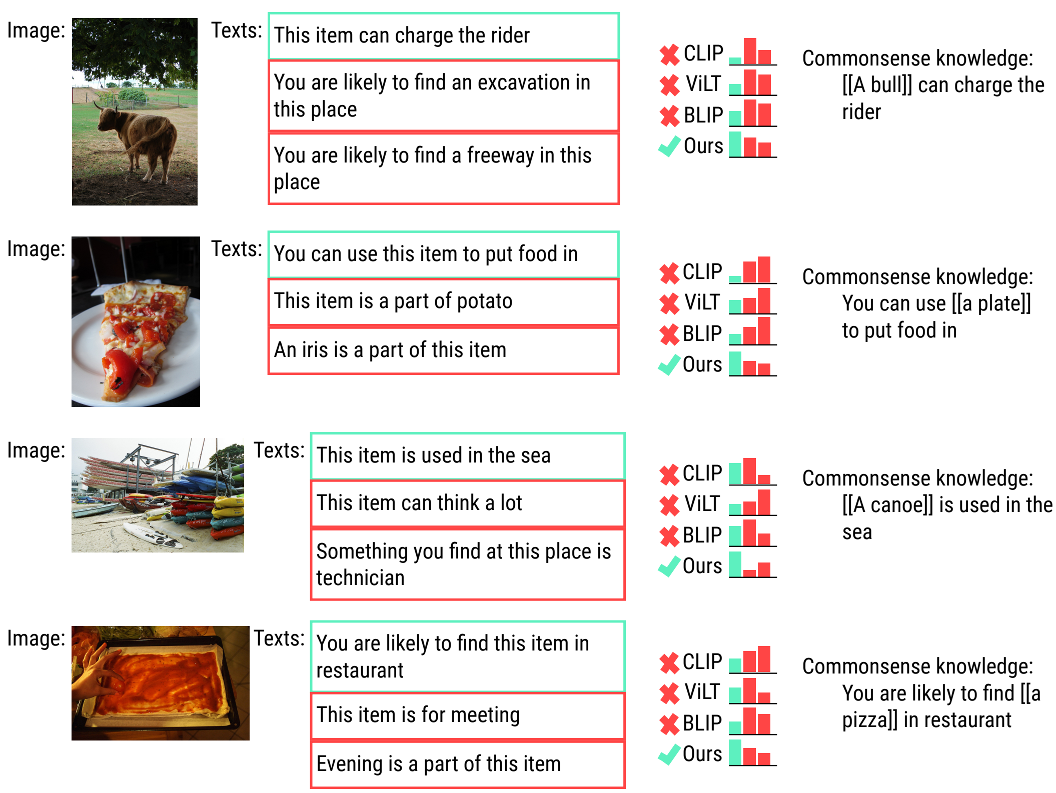

Figure 3. Qualitative examples from our diagnostic test set for image-text retrieval.

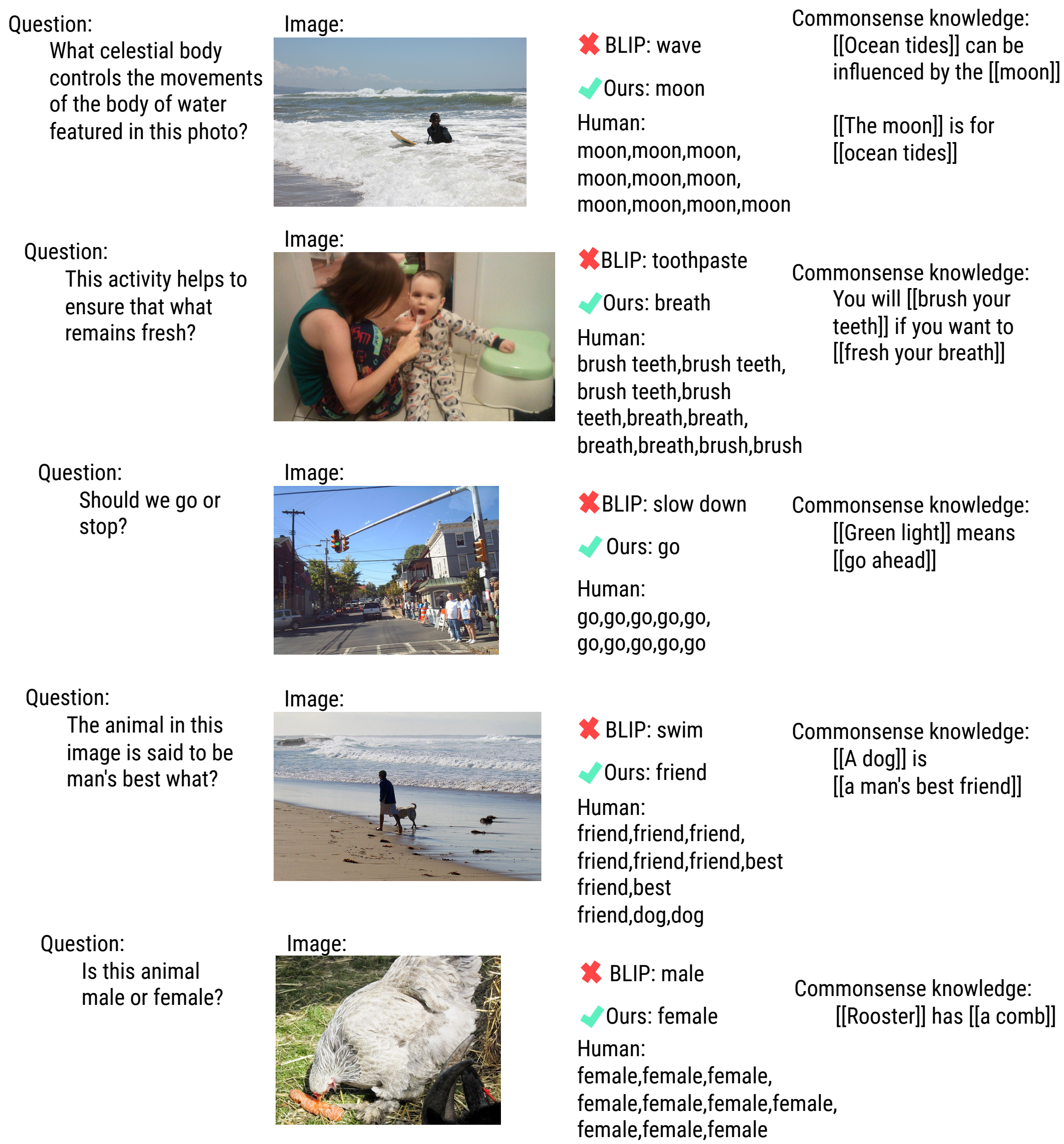

Figure 4. Qualitative examples from the commonsense-aware benchmark OK-VQA.

3. Additional Qualitative Results on OK-VQA Benchmark

In Fig. 4, we show additional qualitative comparison with the state-of-the-art VL-models on the official validation split of the popular commonsense-aware OK-VQA dataset. We note that the validation split is not included during fine-tuning. From left to right is the input question, the input image, the answers by the baseline model BLIP, the DANCE pre-trained model and human, and the related commonsense knowledge from the knowledge graph. The baseline model struggles with these questions and predicts some relevant but wrong answers, which further demonstrates the lack of commonsense ability in the current VL-models. DANCE improves the VL-model's commonsense ability in numerous aspects, including the commonsense knowledge of physics as shown in the first row, the commonsense of human behavior and motivation in the second and third rows, and the knowledge about animals in the fourth and fifth rows. This further demonstrates the commonsense ability enhanced by our DANCE strategy.

4. Statistics of Our Diagnostic Benchmark

	Text-Image seen	Text-Image unseen	Image-Text seen	Image-Text unseen
# Images	4949	4974	500	500
# Texts	500	500	13930	14889
# Seen Images	0	0	0	0
# Seen Texts	500	0	13930	0

Table 2. Statistics of different splits of our diagnostic benchmark.

Figure 5. Case study of failure on the OK-VQA benchmark.

In Table 2, we show the statistics of the four different splits of our diagnostic retrieval test set. Each row respectively represents the number of different images, the number of different text or riddles in each split, and the number of different images and texts that also appear in the training data. All these images for our test set does not appear in the training set. The knowledge in both Text-Image unseen split and Image-Text unseen split is held out from the training set.

5. Failure Case on OK-VQA Benchmark

In the main paper, we mainly focus on enhancing the VL-model's ability to general commonsense via combining the VL data lacking commonsense with commonsense knowledge graphs. However, our model learned from this commonsense-augmented data still suffers in some special real-life scenarios. Here we visualize the failure case of the model with DANCE pre-training in Fig. 5. The model fails to answer a question about counting or quantity. This indicates that the sense of numbers or the mathematical reasoning ability is still weak in existing VL-models, which is also not included in existing commonsense knowledge bases.

6. Additional Details of Human Study and Implementation

For human evaluation in our main paper, take text-image retrieval for example, annotators are given 15 text samples per hour and chose n (Line 466) matching images for each. English proficiency is required. The payment is 6.5 USD/hour.

In our implementation of data generation strategy, to extract entities from captions, we use spaCy to extract noun phrases, remove determiners and adjectives, then double-check POS tag with NLTK. Our manual check of 50 captions found that 88% (126 out of 143) of extraction were successful. 68% of entities are matched with the knowledge base in subsequent alignment. Though polysemy may cause some noise, human ACC (83%) on our dataset in Table 1 indirectly demonstrates the low noise rate of generated data pairs. Moreover, even if the data has some noise, the pre-training quality is not affected as suggested by Table 3 in our main paper.

References

[1] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 1, 2

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[3] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2021. 1

[4] Wikimedia Foundation. Wikimedia downloads. 1, 2

[5] Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. In *International conference on machine learning*, pages 5506–5518. PMLR, 2021. 1

[6] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. 1

[7] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. Fnet: Mixing tokens with fourier transforms. *CoRR*, abs/2105.03824, 2021. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[9] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. 2

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[11] Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Févry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5758–5773, 2021. 1

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019. 1, 2

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 1

[14] Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Bideman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*, 2022. 1

[15] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 1, 2

[16] Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021. 1

[17] Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations*, 2021. 1

[18] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019. 1