015 016

017

018

019

020 021

022 023

024

025 026

027 028

029

030 031

032

033

034

035

036

037

038

039

040

041

042

043

044

045

046

047

048

049 050

051

052

053

107

	054
	055
	056
	057
Supplementary Material for "Improving Commonsense in Vision-Language	058
Models via Knowledge Graph Riddles"	059
	060
	061
Anonymous CVPR submission	062
	064
Deper ID 9521	065
raper in 6551	066
	067
	068
This appendix is organized as follows:	069
• In Section 1, we further illustrate the commonsense lacking issue by providing additional comparison of fundamental	070
VL datasets with commonly used NLP data.	071
	072
• In Section 2, we provide more visualizations of success examples of our method on the proposed diagnostic benchmark	073
for both text-image and image-text retrieval.	074
• In Section 3, we provide more visualizations of success examples of our method on the OK-VQA benchmark.	075
	076
• In Section 4, we summarize the statistics of the proposed diagnostic test data.	077
• In Section 5, we study the failure case of our DANCE augmented model.	078
	080
• In Section 6, we report additional details of human study and implementation of DANCE.	081
1 Commongoings in Eurodomontal VI. Data va NI D Data	082
1. Commonsense in Fundamental VL Data vs NLF Data	083
1.1. Distribution Comparison Between Current VL Data and More Natural Language Data	084
We further explore the commonsense lacking issue in the current fundamental VL data by comparing them with common	085
natural language processing (NLP) data. Here we compare the distributions of the syntactic categories and words of the most	086
nutural language processing (1VL) and CC 12M [1]) with three commonly used NLP datasets: ConcentNet [15] the knowledge	087
base dataset. Wikipedia [4] the popular [2, 6, 10, 13, 18] cleaned English-language articles with the size of 16GB, C4 [12]	088
the popular used [3, 5, 7, 11, 14, 16, 17] English-language text sourced from the Common Crawl web scrape with the size of	089
745GB. The syntactic categories and word distributions comparison is shown in Fig. 1.	090
The upper part of Fig. 1 shows the distribution of the most frequent part-of-speech (POS) tags with punctuation marks	091
excluded, and the lower part shows the most frequent word tokens. There is a significant difference between top POS tag/word	092
token distributions of VL datasets compared with those of the regular texts. Similar to our observation in the main paper,	094
the most frequent words in the text in existing VL datasets are nouns (NOUN) for <b>individual entities</b> , like "street", "table",	095
" <i>train</i> ". In contrast, all the NLP datasets have apparently more verbs (VERB), like " <i>have</i> ", " <i>used</i> ", " <i>find</i> ", " <i>want</i> ", " <i>happen</i> "	096
that contains richer information about the <b>relationship between entities</b> . Besides, the NLP datasets include more particles	097
(PRI), like "to", and pronouns (PRON) like "your", which are associated with <b>interconnection</b> information. This further	098
Illustrates the lacking commonsense issue in the fundamental VL datasets.	099
sense and reasoning ability the fundamental VL datasets are lacking it. This motivates us to use commonsense knowledge to	100
improve VI data. In addition, the distribution of ours training data is also included for comparison. We can see that our data	101
is similar to NLP data in terms of the interconnection between entities.	102
	103
1.2. ChatGPT-Based Commonsense Measurement of VL Data and Natural Language Data	104
We further design to utilize the ChatGPT's powerful in-context learning to measure the amount of commonsense infor-	105
mation for each dataset. We have the score ranges 0-10, and scores descriptive and general language equally, by providing	100



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.

216	Tavé		Commonsense knowledge	270
217	Iext: Images:		Something you find in	271
218	this place is a nicely		<ul> <li>[[the oven]] is a nicely</li> </ul>	272
219	cooked turky		<ul> <li>cooked turky</li> </ul>	273
220			-	274
221	Text: Images:	🔜 🗶 CLIP 📕	Commonsense knowledge:	275
222	Talking with someone	🗶 🕺 Vilt 📲	Talking with someone	276
223	far away requires this	🗶 🕺 🗱 BLIP 💶	far away requires [[a	277
224		🔊 🗸 Ours 🗾	pnonejj	278
225				279
226	Text: Images:	CLIP	Commonsense knowledge:	280
227	This item is used	ViLT	[[A keyboard]] is used	281
228	for coding	K BLIP	for coding	282
229		Vurs 🚺	L	283
230	_ Images:	🗶 CLIP 🗕		284
231	Text:		Commonsense knowledge:	285
232	for propulsion		- [[Driving]] is for	286
233		Ours	- propulsion	287
234			-	288
235				289
236				290
237	Figure 2. Qualitative examples from our diagno	ostic test set for tex	t-image retrieval.	291
238		_		292
239	Image: Texts: This item can charge the rider			293
240	You are likely to find an execution in		Commonsense knowledge:	294
241	this place		[[A bull]] can charge the	295
242			lidei	296
243	You are likely to find a freeway in this	Vurs		297
244	place			298
245		-		299
246	Image: Texts: You can use this item to put food in			300
247	This item is a part of potato		Commonsense knowledge:	301
248	This item is a part of polato		to put food in	302
249	An iris is a part of this item			303
250				304
251				305
252	Image: Texts: This item is used in the sea	7		306
253			Commonsense knowledge:	307
254	This item can think a lot	🗶 ViLT 💶	[[A canoe]] is used in the	308
255	Something you find at this place is		sea	309
256	technician	Vurs		310
257				311
258	Image: Texts: You are likely to find this item in	]		312
259	restaurant	CLIP	Commonsense knowledge:	312
260	This is my in far we astim	- XVilt	You are likely to find [[a	21/
261	I his item is for meeting	BLIP	pizza]] in restaurant	215
262	Evening is a part of this item	Vours		216
263				217
264	Figure 3. Qualitative examples from our diagno	ostic test set for ima	age-text retrieval.	212
265				210
200				319

#### CVPR 2023 Submission #8531. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

324				Commonoonoo knowledge:	378
325	Question:	lmage:	😫 BLIP: wave	[[Ocean tides]] can be	379
326	controls the movements	and the second sec		influenced by the [[moon]]	380
327	of the body of water	and the second sec	Vours: moon	,	381
328	featured in this photo?		Human:	[[The moon]] is for	382
329		at a start and a start	moon,moon,moon,	[[ocean tides]]	383
330				n	384
331		Imagai	11001,11001,11001,11001		385
332	Question:	image.	<b>\$</b> BLIP: toothpaste		386
333	This activity helps to		Ourse breath	Commonsense knowledge:	387
334	ensure that what	Charles .	Vurs: breath	teeth]] if you want to	388
335	Ternains fresh?		Human:	[[fresh your breath]]	389
336			brush teeth brush	,,	390
337			teeth.breath.breath.		391
338			breath,breath,brush,bru	sh	392
339	Question:	Image:			393
340	Should we go or	+	🗱 BLIP: slow down	Commonsense knowledge:	394
341	stop?		🖌 Qurs: do	[[Green light]] means	395
342			V ouro: go	[[go ahead]]	396
343			Human:		397
344			do'do'do'do'do' do'do'do'do'do'		398
345			90,90,90,90,90		399
346	Question:	Image:			400
347	The animal in this		🗱 BLIP: swim	Commonsense knowledge:	401
348	image is said to be	Construction of the second states of the second	Ours: friend	[[A dog]] is	402
349	man's best what?		Vurs. menu	[[a man's best friend]]	403
350			Human: friend friend friend		404
351			friend friend friend hest		405
352		and the second sec	friend,best		406
353			friend,dog,dog		407
354	Question:	Image:			408
355	Is this animal		样 BLIP: male	Commonsense knowledge:	409
356	male or female?		Ours: female	[[Rooster]] has [[a comb]]	410
357		Che and the	Human:		411
358			female,female,female,		412
359			female,female,female,fe	emale,	413
360			female,female,female		414
361					415
362	Figure 4. Q	pualitative examples from the co	ommonsense-aware benc	chmark OK-VQA.	416
363					417

# 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] Luczu Lucz Therry Lashra Airelia, Iluz Exlection and Services Orthum Free Self-supervised for a 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
  - [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.
- [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.
- 558 [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] Pohyn Spear, Joshua Chin, and Catherine Havasi. Concentrat 5.5: An open multilingual graph of general knowledge. In *Thirty first*
  - [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.