Enhancing Vision-Language Compositional Understanding with Multimodal Synthetic Data

Appendix

The Appendix is organized as follows:
- Section A1 presents additional details about SPARCL.
- Section A2 provides further details on the experimental setup.
- Section A3 includes additional experimental results.

A1. Details of SPARCL

The prompts used as input to the LLM for generating negative and positive captions are presented in Figure A1.

> You are an assistant assigned to help a human user edit a given sentence that describes an image. Make a minor change to the sentence by randomly altering, omitting, inserting, or replacing one word or phrase. Although the change should be minor, it must result in a significant difference in the sentence's meaning, making it unable to describe the original image. Use the provided template and respond with a single, valid sentence.
> User: {}
> Assistant: Sure! Here's my edit:

(a) Prompts used to generate negative captions.

> You are an assistant assigned to help a user edit a sentence that describes an image. Make a minor change to the sentence by randomly altering, omitting, inserting, or replacing one word or phrase. The new sentence must strictly retain the same meaning as the original sentence. Use the provided template and respond with a single, valid sentence.
> User: {}
> Assistant: Sure! Here's my edit:

(b) Prompts used to generate positive captions.

Figure A1. Prompts used to generate negative and positive captions.

A2. Experimental Setup

Data Synthesis. For caption generation, we utilize the Llama-2-Chat 13B model[1], with the temperature set to 0.9, top-k set to 100, and top-p set to 0.9 for sampling. For image generation, we use the LCM model[2] for its swift inference with few steps [61]. The pretrained CLIP ViT-L/14 [74] is used as the image feature extractor for injecting image features. We perform 8 inference steps with LCM to generate each image.

Hyperparameter Selection. First, we use only real training samples to select τ and b. The optimal values are determined by searching for the ones that minimize the training loss at the first training step, aiming to preserve the output distribution from the pretrained model. After searching, we set $\tau = 0.01$ and $b = -30.0$. Next, we select the base learning rate, weight decay, and LoRA adapter rank based on performance on the COCO-2014 validation set, in which the model is trained exclusively on real samples. According to the performance on the validation set, these hyperparameter are set to a base learning rate of 0.01, weight decay of 0.5, and LoRA adapter rank of 16. Then, we construct a validation set composed of the CIFAR-10 [47] test set and a randomly selected 5% of samples from ARO-Attribute and ARO-Relation, to balance the performance on coarse-grained and fine-grained tasks. Using this validation set, we train the model on both real and synthetic samples and use the validation performance to determine the remaining hyperparameters: m_0, α, β, γ and λ. The effects of these hyperparameters are shown in Table A5.

A3. Experimental Results

Performance on each subset of the four benchmarks. Table A1, A2 and A3 present the performance of different methods on each subset of the four benchmarks.

Comparison with other images generation methods. We compare our image generation method with StyleAligned [27]. For a fair comparison, we use an ablated version #7 of SPARCL (Sec. 4.4, main paper) without the adaptive margin loss. Both methods use synthetic captions that we generate. As shown in Table A4, StyleAligned performs about 1% worse than our method on the four compositional benchmarks, which illustrates the effectiveness of image feature injection in SPARCL. In Figure A2, we show two synthetic images from StyleAligned, where it fails to align the generated content with the synthetic captions. We hypothesize that the diffusion trajectory of the real image imposes strong constraints on the image generation model, making StyleAligned difficult to edit the image to match the synthetic caption. This issue is similar to the zero-shot image editing methods [6, 22, 63], which provide incorrect guidance during model training and lead to limited improvements on compositional understanding tasks. More-

[1] https://huggingface.co/meta-llama/Llama-2-13b-chat
[2] https://huggingface.co/SimianLuo/LCM_Dreamshaper_v7

Table A1. Comparison of accuracy (%) between SPARCL and baselines on ARO and VL-CheckList. "img" represents images, "cap" represents captions, "syn" represents synthetic data.

Method	Training Data					ARO			VL-CheckList			
	Source	# real img	# real cap	# syn img	# syn cap	Relation	Attribute	Average	Attribute	Relation	Object	Average
CLIP-ZeroShot[74]	-	-	-	-	-	59.22	62.86	61.03	67.05	66.71	85.72	73.16
CLIP-Finetune[74]	COCO	82K	410K	0	0	63.02	65.16	64.09	66.74	64.43	86.86	72.78
SDS-CLIP [3]	COCO	82K	410K	0	0	53.0	62.0	57.5	-	-	-	-
[79]	COCO	0	0	82K	82K	-	-	-	70.7	53.8	85.1	69.87
AMR-NegCLIP [83]	COCO	100K	100K	0	500K	83.2	75.6	79.4	-	-	-	-
NegCLIP [106]	COCO	100K	100K	0	500K	81.0	71.0	76.0	70.9	68.9	84.1	74.6
MosaiCLIP [85]	COCO	109K	109K	0	981K	82.6	78.0	80.3	70.1	71.3	89.0	76.8
FSC-CLIP [68]	COCO	100K	100K	0	1.5M	-	-	-	-	-	-	77.20
CE-CLIP [109]	COCO	82K	410K	0	2M	83.00	76.40	79.70	72.62	71.75	84.65	76.34
COMO [49]	COCO	113K	567K	567K	567K	-	-	-	73.44	71.16	86.20	76.93
SPARCL	COCO	82K	410K	820K	820K	80.10	74.19	77.15	73.72	72.99	90.76	79.16
SPEC [70]	LAION	20K	20K	20K	20K	73.7	66.4	70.1	-	-	-	-
[18]	CC3M	3M	3M	0	9M	-	-	-	71.97	68.95	85.00	75.31
CE-CLIP+ [109]	COCO+CC3M	3M	3M	0	15M	83.6	77.1	80.35	76.76	74.70	86.30	79.25
CLOVE [11]	LAION-COCO	>1B	>1B	0	>1B	69.0	77.4	73.2	-	-	-	-
syn-CLIP [9]	SyViC	0	0	>1M	>1M	71.40	66.94	69.17	70.37	69.39	84.75	74.84
FiGCLIP [45]	VidSitu	20K videos		0	0	68.01	65.99	67.00	-	-	-	-

Table A2. Comparison of accuracy (%) between SPARCL and baselines on SugarCrepe. "img" represents images, "cap" represents captions, "syn" represents synthetic data.

Method	Training Data					Add		Replace			Swap		Average
	Source	# real img	# real cap	# syn img	# syn cap	Attribute	Object	Attribute	Object	Relation	Attribute	Object	
CLIP-ZeroShot[74]	-	-	-	-	-	69.22	77.40	80.33	90.98	69.49	64.71	61.63	73.39
CLIP [74] (Finetune)	COCO	82K	410K	0	0	78.03	88.12	85.79	93.58	73.83	71.77	68.29	79.92
AMR-NegCLIP [83]	COCO	100K	100K	0	500K	-	-	-	-	-	-	-	79.92
NegCLIP [106]	COCO	100K	100K	0	500K	82.80	88.80	85.91	92.68	76.46	75.38	75.20	82.46
FSC-CLIP [68]	COCO	100K	100K	0	1.5M	-	-	-	-	-	-	-	85.10
CE-CLIP [109]	COCO	82K	410K	0	2M	93.4	92.4	88.8	93.1	79.0	77.0	72.8	85.2
SPARCL	COCO	82K	410K	820K	820K	93.49	92.43	88.95	95.82	78.94	81.38	78.77	87.11
CounterCurate [108]	Flickr	30K	30K	150K	150K	86.71	90.35	87.94	95.94	76.24	73.57	68.57	82.76
CE-CLIP+ [109]	COCO+CC3M	3M	3M	0	15M	94.9	93.8	90.8	93.8	83.2	79.3	76.8	87.5
CLOVE [11]	LAION-COCO	>1B	>1B	0	>1B	-	-	-	-	-	-	-	79.92
IL-CLIP [114]	CC12M	12M	12M	0	0	-	-	-	-	-	-	-	70.34
SF-CLIP [80]	YFCC15M	15M	15M	0	0	-	-	-	-	-	-	-	71.20
FiGCLIP [45]	VidSitu	20K videos		0	0	72.5	77.4	81.1	91.8	69.4	66.1	63.8	74.6

over, StyleAligned requires DDIM inversion to obtain the inverted diffusion trajectory from the real image, making it computationally expensive and impractical for large-scale image generation.

Effects of image feature injection. In Figure A3 and A4, we present examples of synthetic images to illustrate how image feature injection helps mitigate unintended changes. In Figure A3, we observe that feature injection helps to generate images with similar object size and viewing angle to the real image. For example, in (a), the real image depicts a wide shot of a girl, while the synthetic image without feature injection produces a close-up shot despite aligning with the caption. With feature injection, the synthetic image maintains a wide shot, resembling the real image. Similar effects are seen in (b) and (c). In (d), the synthetic image with feature injection preserves the viewing angle of the real image, whereas the one without feature injection devi-

Table A3. Comparison of accuracy (%) between SPARCL and baselines on SugarCrepe++. “img” represents images, “cap” represents captions, “syn” represents synthetic data.

Method	Training Data					Replace			Swap		Average
	Source	# real img	# real cap	# syn img	# syn cap	Attribute	Object	Relation	Attribute	Object	
CLIP-ZeroShot[74]	-	-	-	-	-	65.61	86.80	56.26	45.21	45.18	59.81
CLIP-Finetune[74]	COCO	82K	410K	0	0	69.03	90.61	56.33	49.24	46.21	62.27
NegCLIP[106]	COCO	100K	100K	0	500K	69.41	89.53	52.27	57.99	55.25	64.89
SPARCL	COCO	82K	410K	820K	820K	68.90	89.76	52.34	57.95	61.63	66.12
[18]	CC3M	3M	3M	0	9M	56.98	80.93	47.30	48.4	42.98	55.32

Table A4. Performance comparison (%) between SPARCL and StyleAligned.

Variant	ARO	VL-CheckList	SugarCrepe	SugarCrepe++	Average
StyleAligned [27]	72.60	75.03	85.70	65.25	74.65
SPARCL (#7)	74.12	76.35	85.40	66.44	75.58

(a) **Real caption**: A pizza sitting on top of a pan next to a paper cut out.
Synthetic caption: A paper cut out sitting on top of a pan next to a pizza.

(b) **Real caption**: A person on skis skiing down a mountain slope.
Synthetic caption: A person on rollerblades rolling down a grassy hill.

Figure A2. Examples of synthetic samples from StyleAligned. The algorithm did not alter the image content according to the caption.

ates from it. In Figure A4, we observe that feature injection helps generate backgrounds that resemble the real image. For example, in (a), the real image and the synthetic image without feature injection depicts an outdoor street scene, creating a noticeable difference. With feature injection, the single-colored background makes the synthetic image more similar to the real one. In (b), the sky occupies much of background in the real image as well as the image generated with feature injection, whereas the one without feature injection shows little sky. Also, the basketball is present in both the real and the synthetic image with feature injection but not in the middle image. Similar effects are observed in (c) and (d). These examples show that image feature injection reduces unintended variations not captured by the caption, enhancing the usefulness of synthetic samples for training VLMs.

Effects of hyperparameters. Table A5 presents the performance of SPARCL with different hyperparameter settings. For λ, we observe that $\lambda = 0.01$ achieves the highest average validation accuracy, leading us to select it for subsequent experiments. Similarly, for α, the best performance is obtained with $\alpha = 10.0$, which is used in other experiments. When evaluating different values of m_0, we find that $m_0 = 0.005$ yields the best results. Finally, we examine various combinations of β and γ and observe that $\beta = -0.02$ and $\gamma = 1.0$ provide the best validation performance. Thus, this combination is selected as the optimal hyperparameter setting.

Table A5. Performance of SPARCL with different hyperparameters. “ARO-Rel” refers to the ARO-Relation validation subset, and “ARO-Att” refers to the ARO-Attribute validation subset, both consisting of a randomly selected 5% of the full set, as described in Sec. A2.

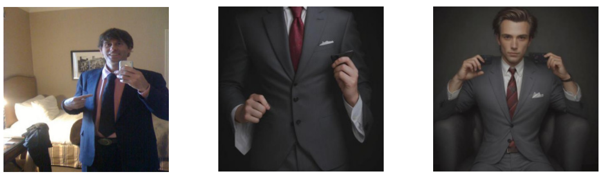
λ	α	m_0	β	γ	Validation				Test Average
					CIFAR-10	ARO-Rel	ARO-Att	Average	
0.0	0.0	-	-	-	86.56	78.79	76.52	80.62	75.58
0.001	0.0	0.01	0.0	0.0	85.02	78.21	72.72	78.65	75.95
0.01	0.0	0.01	0.0	0.0	83.66	81.77	76.46	80.63	76.78
0.1	0.0	0.01	0.0	0.0	85.02	81.29	78.94	80.47	76.94
0.01	1.0	0.01	0.0	0.0	83.18	81.10	76.52	80.27	77.21
0.01	10.0	0.01	0.0	0.0	86.46	81.89	75.05	81.13	77.27
0.01	100.0	0.01	0.0	0.0	87.64	74.93	75.19	79.25	75.28
0.01	10.0	0.005	0.0	0.0	85.91	79.87	78.76	81.51	77.08
0.01	10.0	0.01	0.0	0.0	86.46	81.89	75.05	81.13	77.27
0.01	10.0	0.02	0.0	0.0	84.85	79.41	75.28	79.85	76.79
0.01	10.0	0.005	-0.02	1.0	86.46	80.08	78.90	81.81	77.38
0.01	10.0	0.005	-0.03	1.0	86.75	81.08	76.43	81.42	77.25
0.01	10.0	0.005	-0.02	3.0	87.31	80.79	76.52	81.54	77.23

Real image | Synthetic image w/o image feature injection | Synthetic image with image feature injection

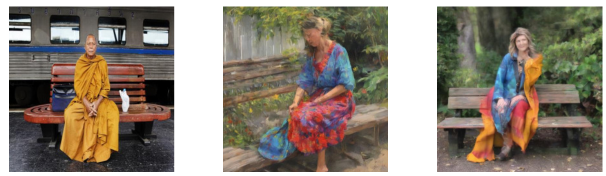
Real image | Synthetic image w/o image feature injection | Synthetic image with image feature injection

(a) Real caption: A girl in a pink snowsuit skiing on a ski slope.
Synthetic caption: A pink girl skiing on a slope.

(b) Real caption: A zebra standing next to a stone wall.
Synthetic caption: A zebra standing next to a wavy stone wall.



(c) Real caption: A man in suit and tie holding a cell phone posing for a picture.
Synthetic caption: A suit-clad man posing with a tie-adorned cell phone for a picture.

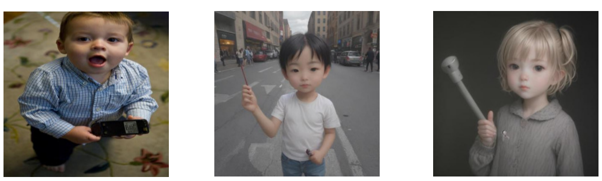


(d) Real caption: Woman with bright smock sitting on a wooden bench.
Synthetic caption: Woman with vibrant smock sitting on a weathered bench.

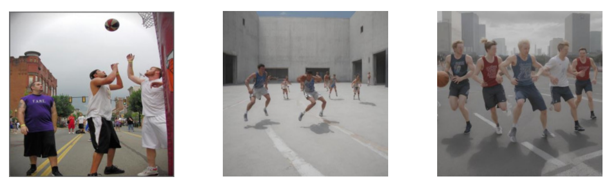
Figure A3. Examples of synthetic samples without and with image feature injection. In these examples, the image feature injection technique achieves alignment of the subject size and the viewing angle with those in real images.

Real image | Synthetic image w/o image feature injection | Synthetic image with image feature injection

Real image | Synthetic image w/o image feature injection | Synthetic image with image feature injection

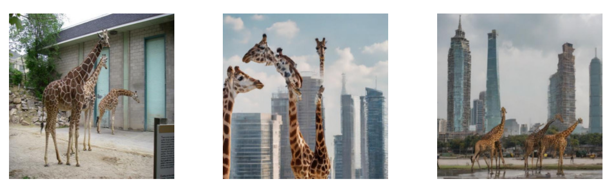


(a) Real caption: A little kid that is holding a phone.
Synthetic caption: A little kid who is holding a selfie stick.



(b) Real caption: A group of guys playing basketball on a city street.
Synthetic caption: A group of hoopsters dribbling down a concrete floor.

(c) Real caption: A red fire hydrant next to wall made of stone.
Synthetic caption: A fiery red fire hydrant stands next to a sturdy stone wall.



(d) Real caption: A group of giraffes standing next to a building.
Synthetic caption: A group of giraffes standing next to a skyscraper.

Figure A4. Examples of synthetic samples without and with image feature injection. In these examples, the image feature injection primarily helps to generate backgrounds that resemble those in real images. For example, in (d), both the first and the third images show the ground, whereas the second image does not.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 3

[2] Manel Baradad Jurjo, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. *Advances in Neural Information Processing Systems*, 34:2556–2569, 2021. 3

[3] Samyadeep Basu, Maziar Sanjabi, Daniela Massiceti, Shell Xu Hu, and Soheil Feizi. Augmenting clip with improved visio-linguistic reasoning. *arXiv preprint arXiv:2307.09233*, 2023. 2, 6

[4] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020. 3

[5] Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, et al. Improving fine-grained understanding in image-text pre-training. *arXiv preprint arXiv:2401.09865*, 2024. 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 2

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 3

[8] Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. Fff: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14172–14182, 2024. 5

[9] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165, 2023. 2, 6

[10] Paola Cascante-Bonilla, Yu Hou, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. Natural language inference improves compositionality in vision-language models. *arXiv preprint arXiv:2410.22315*, 2024. 3

[11] Santiago Castro, Amir Ziai, Avneesh Saluja, Zhuoning Yuan, and Rada Mihalcea. Clove: Encoding compositional language in contrastive vision-language models. *arXiv preprint arXiv:2402.15021*, 2024. 2, 3, 6, 7

[12] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 1, 3

[13] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International conference on machine learning*, pages 1062–1070. PMLR, 2019. 3

[14] Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*, 6(1):84, 2020. 3

[15] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[16] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 3

[17] Sivan Doveh, Assaf Arbelle, Sivan Harary, Amit Alfassy, Roei Herzig, Donghyun Kim, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *arXiv preprint arXiv:2305.19595*, 2023. 2

[18] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 2, 3, 5, 6, 7

[19] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *arXiv preprint arXiv:2406.11171*, 2024. 2, 6

[20] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1

[21] Jiahui Gao, Renjie Pi, Lin Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. Self-guided noise-free data generation for efficient zero-shot learning. In *International Conference on Learning Representations (ICLR 2023)*, 2023. 3

[22] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. *arXiv preprint arXiv:2403.14602*, 2024. 2, 1

[23] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural net-

works. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1, 3

[24] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 6

[25] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 3

[26] Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842, 2022. 3

[27] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 1, 3

[28] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*, 2023. 2

[29] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for deep learning. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[30] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017. 3

[31] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*, 2022. 3

[32] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*, 2023. 1, 2, 3, 5, 6, 8

[33] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[34] Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024. 2

[35] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 4

[36] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2417–2425, 2024. 2

[37] Zhizhong Huang, Junping Zhang, and Hongming Shan. Twin contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11661–11670, 2023. 3

[38] Sarah Ibrahimi, Arnaud Sors, Rafael Sampaio de Rezende, and Stéphane Clinchant. Learning with label noise for image retrieval by selecting interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2181–2190, 2022. 3

[39] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021. 3

[40] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[41] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018. 3

[42] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 1, 2

[43] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 3

[44] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[45] Zeeshan Khan, Makarand Tapaswi, et al. Figclip: Fine-grained clip adaptation via densely annotated videos. *arXiv preprint arXiv:2401.07669*, 2024. 2, 6

[46] Bumsoo Kim, Yeonsik Jo, Jinhyung Kim, and Seunghwan Kim. Misalign, contrast then distill: Rethinking misalignments in language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2023. 3

[47] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[48] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020. 3

[49] Chengen Lai, Shengli Song, Sitong Yan, and Guangneng Hu. Improving vision and language concepts understanding with multimodal counterfactual samples. In *European*

Conference on Computer Vision, pages 174–191. Springer, 2024. 2, 3, 6, 8

[50] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5051–5059, 2019. 3

[51] Junnan Li, Caiming Xiong, and Steven CH Hoi. Learning from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9485–9494, 2021. 3

[52] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[53] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[54] Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. Desco: Learning object recognition with rich language descriptions. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[55] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[56] Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao. Noise-resistant deep metric learning with ranking-based instance selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6811–6820, 2021. 3

[57] Hao Liu, Tom Zahavy, Volodymyr Mnih, and Satinder Singh. Palm up: Playing in the latent manifold for unsupervised pretraining. *Advances in Neural Information Processing Systems*, 35:35880–35893, 2022. 3

[58] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[59] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[60] Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Imagination-augmented natural language understanding. *arXiv preprint arXiv:2204.08535*, 2022. 3

[61] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1, 3

[62] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 1

[63] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 1

[64] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477, 2022. 3

[65] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR, 2023. 3

[66] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2311.17076*, 2023. 2

[67] Sergey I Nikolenko. *Synthetic data for deep learning*. Springer, 2021. 3

[68] Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multi-modal capabilities of pre-trained vlms for improving vision-linguistic compositionality. *arXiv preprint arXiv:2410.05210*, 2024. 2, 3, 6

[69] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *arXiv preprint arXiv:2411.02545*, 2024. 7

[70] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. *arXiv preprint arXiv:2312.00081*, 2023. 2, 6

[71] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13279–13288, 2024. 2, 3

[72] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 3

[73] Alec Radford. Improving language understanding by generative pre-training. 2018. 3

[74] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6, 1, 3

[75] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[76] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In *2019*

IEEE automatic speech recognition and understanding workshop (ASRU), pages 996–1002. IEEE, 2019. 3

[77] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073. IEEE, 2020. 3

[78] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[79] Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573, 2024. 2, 3, 6

[80] Sepehr Sameni, Kushal Kafle, Hao Tan, and Simon Jenni. Building vision-language models on solid foundations with masked distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14216–14226, 2024. 2, 6

[81] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[82] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024. 2

[83] Ziyi Shou and Fangzhen Lin. Enhancing semantic understanding in vision language models using meaning representation negative generation. In *Fourth Workshop on Knowledge-infused Learning*, 2024. 2, 3, 6

[84] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[85] Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. *arXiv preprint arXiv:2305.13812*, 2023. 2, 6

[86] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024. 2, 3

[87] Tianyi Tang, Yushuo Chen, Yifan Du, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Learning to imagine: Visually-augmented natural language generation. *arXiv preprint arXiv:2305.16944*, 2023. 3

[88] 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, 2

[89] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[90] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2

[91] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 3

[92] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018. 3

[93] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):1–13, 2020. 3

[94] Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[95] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021. 3

[96] Fei Wang, Liang Ding, Jun Rao, Ye Liu, Li Shen, and Changxing Ding. Can linguistic knowledge improve multimodal alignment in vision-language pretraining? *arXiv preprint arXiv:2308.12898*, 2023. 2

[97] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 3

[98] Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*, 2021. 3

[99] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*, 2020. 3

[100] Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. Z-lavi: Zero-shot language solver fueled by visual imagination. *arXiv preprint arXiv:2210.12261*, 2022. 3

[101] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[102] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, pages 10789–10798. PMLR, 2020. 3

[103] Nir Yellinek, Leonid Karlinsky, and Raja Giryes. 3vl: using trees to teach vision & language models compositional concepts. *arXiv preprint arXiv:2312.17345*, 2023. 2

[104] Hu Yu, Hao Luo, Fan Wang, and Feng Zhao. Uncovering the text embedding in text-to-image diffusion models. *arXiv preprint arXiv:2404.01154*, 2024. 3

[105] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2, 3

[106] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 3, 5, 6

[107] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 5

[108] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. *arXiv preprint arXiv:2402.13254*, 2024. 2, 3, 6

[109] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic fine-grained understanding. *arXiv preprint arXiv:2306.08832*, 2023. 2, 3, 5, 6, 7, 8

[110] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2, 3

[111] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 3

[112] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 2

[113] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 1, 2, 6

[114] Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13785–13795, 2024. 2, 6

[115] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

[116] Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*, 2022. 3