Supplementary Material for Going Beyond Nouns With Vision & Language Models Using Synthetic Data

Paola Cascante-Bonilla[*†1,2] **Khaled Shehada**[*2,3] **James Seale Smith**[2,4] **Sivan Doveh**[6,7]
Donghyun Kim[2,7] **Rameswar Panda**[2,7] **Gül Varol**[5] **Aude Oliva**[2,3]
Vicente Ordonez[1] **Rogerio Feris**[2,7] **Leonid Karlinsky**[2,7]
[1]Rice University [2]MIT-IBM Watson AI Lab [3]MIT [4]Georgia Institute of Technology
[5]LIGM, École des Ponts [6] Weizmann Institute of Science [7]IBM Research

Appendix

In this supplementary material, we share our code and provide additional insights and experimental results that were not included in the main paper due to space constraints. In Section A we describe the implementation code for generating **SyViC** and the code for the proposed finetuning approach on **SyViC**. In Section B, we analyze the performance of recently open-sourced VL models on VL-Checklist [18] and show they have low performance, demonstrating the need for our improvements. In Section C, we provide additional results on Winoground [13] using CyCLIP [7] (excluded from the main text for space purposes). Section D demonstrates how we can improve BLIP [8] using **SyViC**. In Section E, we combine our contributions with those of *concurrent* work of [6] and demonstrate that our approach for improving VLC using synthetic data is *orthogonal / complementary* to the text-augmentation based methods. Section F provides more dataset details, describing how the metadata from each synthesized scene is used to generate a caption for each image in **SyViC**. In Section G, we explore combinations of our metadata-driven grammar-based caption generation with paraphrasing using openly available large language models. Section H provides "**SyViC** - number of models and number of samples" ablation excluded from the main paper due to lack of space. Finally, in Section I we provide some randomly sampled examples from **SyViC**.

A. Code

Our code for both **SyViC** data synthesis and the proposed finetuning approach is included in our project page: https://synthetic-vic.github.io/

[*]Equal contribution. Project page: https://synthetic-vic.github.io/
[†]Work partially done while interning at the MIT-IBM Watson AI Lab.

B. Expanding VL-Checklist [18] analysis to most recent VL models

As promised in the footnote in the introduction we have evaluated the very recently released open-source VL models, namely: METER (CVPR 22) [5], X-VLM (ICML 22) [17], and VLMO (NeurIPS 22) [1] on the most extensive VLC understanding benchmark of VL-Checklist [18] observing average performance of 56.8%, 58.9%, and 54.6% respectively. As noted in the introduction, this relatively low VLC understanding performance (below CLIP [11]) of the newest (open) VL models illustrates once again the very much needed improvement in this aspect. Consequently, it also underlines the importance of **SyViC** and the proposed finetuning approach for administering some of this improvement and highlighting the future potential of our approach and synthetic data in general for the VL modeling. We additionally explore the very recent BLIP [8] model and how it could be improved using **SyViC** and our approach in Section D.

C. Winoground Results for CyCLIP [7]

As promised in the main paper (lines 555-556), we include the Winoground [13] results of CyCLIP [7] not included in the main paper for lack of space. The results are included in Table A.1 and, compared to the CyCLIP baseline, demonstrate stable improvements of up to 1.17% group score for syn-CyCLIP finetuned on **SyViC** using our proposed approach.

D. Improving BLIP [8] using SyViC

BLIP is a recently released VL model achieving better out-of-the-box performance on VL-Checklist [18] and Winoground [13] compared to CLIP [11]. In Table A.2 we show how using our proposed **SyViC** dataset and the finetuning approach applied to BLIP, significant additional performance gains (1.48% on VL-Checklist and up to 4.67% on Winoground group score) can be achieved. BLIP is de-

	Winoground			Winoground†		
	Text	Image	**Group**	Text	Image	**Group**
CyCLIP	28.50	9.50	7.25	32.16	11.11	8.19
syn-CyCLIP	30.00 (1.5)	10.75 (1.25)	8.25 (1.0)	30.99 (-1.17)	12.87 (1.76)	9.36 (1.17)

Table A.1. Winoground [13] performance of syn-CyCLIP – finetuned on **SyViC**. † 'clean' (no-tag) subset of valid Winoground samples from [4]

signed and optimized for VL understanding and generation [8], and has a relatively low zero-shot out-of-the-box performance compared to CLIP (e.g., we observed an over 7% drop in zero-shot comparing baseline BLIP to CLIP and similar drop for syn-BLIP compared to CLIP). In more detail, we employ the retrieval flow of BLIP starting from ViT/B and CapFilt-L base model and use it as the BLIP baseline in Table A.2. We finetune BLIP on **SyViC** following the complete proposed recipe detailed in Sec. 3.2 of the main paper. We add rank-16 LoRa adapters to both BLIP encoders and the decoder (cross-attention layers in the text encoder). We fine-tune for two epochs with a learning rate of 5e-6 using an Adam optimizer with a weight decay factor of 0.05 [9].

E. Exploring a combination with text augmentation methods

As discussed in lines 108-112 in the Introduction of the main paper, *concurrent* works [6, 16] propose an *orthogonal* approach of improving VLC understanding performance via using text augmentation while training on additional real VL paired image+text data. These works use language tools to teach a model the importance of non-noun words by manipulating them (replacing words with incorrect alternatives in the text captions of real image+text pairs) and adding the resulting texts to the same batch. In order to show that our proposed approach of improving the VLC understanding performance of VL models using targeted demonstration on *both text and image side* via generating synthetic data (our **SyViC** dataset) is truly orthogonal and complementary to the text augmentation methods, we have conducted the following experiment whose results are summarized in Table A.3. Specifically, we used [6] code, kindly shared to us by the authors, to combine our **SyViC** finetuning (as described in Section 3.2 of the main paper) with the LAION [12] experiment of [6] using their text augmentation method both for the real data captions as well as for our **SyViC** synthetic data captions. More specifically, we finetune (using the method described in Section 3.2 in the main paper, also including [6] negative text augmentations and their additional losses for the negatives as detailed in [6]) on combined batches containing both LAION [12] text+image pairs and **SyViC** text+image pairs. The base model is CLIP [11] both for [6] and syn-[6]. As we can see in A.3, syn-[6] significantly (up to 12.26% on Relations and 8.46% on average) improves the base [6] performance on VL-Checklist (trained on the same LAION data without **SyViC**) and is roughly matching [6] performance on the ARO and zero-shot evaluations.

F. Metadata-driven caption text synthesis, more details

This section describes how the metadata from each synthesized scene is used to generate a caption for each image in **SyViC**. We outline a rule-based mechanism to deterministically generate dense captions given:

1. List of the objects present in the scene, each with its corresponding name and world coordinates.

2. List of humanoids present in the scene, each with its world coordinates, clothing identifier, and a textual description of the action it performs.

3. Segmented image that has a label for each pixel corresponding to the object or humanoid it belongs to.

4. Scene identifier that maps to a textual description of the scene.

We annotate a list of 115 clothing textures from the Multi-Garment [2] and SURREAL [14]. Clothing annotations include a list of textual descriptions such as the colors of the clothes, the hair/beard style, as well as any features that stand out such as logos, tattoos, and accessories. Additionally, we use the original scene descriptions provided by ThreeDWorld's scene library.

To generate the description of the objects in the image, we use the (3D) world positions of the objects to create positional relations between them. For objects that are horizontally aligned, we generate a description of which object is to the left or right of the other by comparing their corresponding pixels in the segmentation image. Furthermore, we generate a description of which object is in front of the other by translating the world coordinates into camera coordinates and comparing their z-coordinates. Unique identifiers for names are used to as placeholders to obtain those relationships, and object names are filled in once all positional relations are established, using indefinite articles when necessary. This process is applied to every pair of objects present in the scene.

	VL Checklist			Winoground			Winoground†		
	Relation	Attribute	**Average**	Text	Image	**Group**	Text	Image	
BLIP	68.45	73.11	70.78	38.00	18.25	14.50	43.86	25.15	21.06
syn-BLIP	70.18 (+1.73)	75.34 (+2.23)	72.76 (+1.48)	43.25 (+5.25)	19.75 (+1.5)	16.75 (+2.25)	52.63 (+8.77)	29.82 (+4.67)	25.73 (+4.67)

Table A.2. Performance of syn-BLIP – finetuned on **SyViC** and evaluated on VL-Checklist and Winoground. † 'clean' (no-tag) subset of valid Winoground samples from [4]. Gains and losses are highlighted in green and red respectively.

	VL Checklist			ARO					Zero-Short
	Relation	Attribute	**Average**	VG-Rel.	VG-Att.	Flickr30k	COCO	**Average**	(21 tasks)
CLIP	63.57	67.51	65.54	58.84	63.19	47.20	59.46	57.17	56.07
[6]	66.05	69.64	67.85	80.64	72.81	92.82	87.67	83.48	56.71
syn-CLIP	69.39 (+5.82)	70.37 (+2.86)	69.88 (+4.34)	71.40 (+12.56)	66.94 (+3.75)	59.06 (+11.86)	70.96 (+11.5)	67.09 (+9.9)	55.27 (-0.8)
syn-[6]	78.31 (+12.26)	74.31 (+7.67)	76.31 (+8.46)	80.79 (+0.15)	72.37 (-0.44)	92.44 (-0.38)	87.19 (-0.48)	83.20 (-0.28)	54.57 (-2.14)

Table A.3. Demonstrating that text augmentation on real paired VL data training is orthogonal/complementary to our approach. Comparing [6] performance finetuned on LAION with syn-[6] performance finetuned on LAION + **SyViC** using a combination of our approach in Section 3.2 of the main paper with [6]'s negatives text augmentations and additional negative losses (added from [6] original code kindly shared with us by the authors of [6]). The base model is CLIP [11] both for [6] and syn-[6]. Results are evaluated on VL-Checklist [18] and ARO [16]. Gains and losses are highlighted in green and red respectively. syn-[6] is comparable on ARO (with 0.28% difference) and significantly improving on VL-Checklist (with 8.46% average improvement), while having only a small decrease on the zero-shot evaluation, only 2.14% w.r.t. [6] and even smaller 1.5% compared to base CLIP model.

Furthermore, we generate descriptions of humans while referring to them using ordinal numbers. In particular, for each human present in the scene, we retrieve its action description and place it in a sentence (e.g. "The {first} person {walks forward}". Additionally, we retrieve the list of textual descriptions associated with the human's clothing, if exist. We consider each text as a separate sentence.

Finally, we compile a list of sentences containing a caption prefix, an enumeration of the objects, the pairwise positional relations between objects, a scene description, and action and clothing descriptions for each human. We concatenate the sentences together to get a full dense caption of the image. A simplified pseudo-code for generating a caption is shown below:

Listing 1. General Code for Caption Generation

```
def sample_prompt(objects, seg_image, scene_name):
  statements = ["This scene contains "]

  # Objects
  objects_statement = ""
  for obj in objects:
    article = get_article(obj.obj_name[0])
    objects_statement += article + f" {obj.obj_name}, "
  objects_statement += f"and {len(humans)} humans."
  statements.append(objects_statement)

  # Scene statement
  scene_statement = "They are in "
  scene_article = get_article(scene_name)
  scene_description = get_description(scene_description)
  scene_statement += scene_article + scene_description
  statements.append(scene_statement)

  # Positional relations
  relations = []
  n = len(len(objects))
  for i in range(n):
    for j in range(i + 1, n):
      left, right = get_left_right(seg_image, i, j)
      front, back = get_front_back(i, j)
      relations.extend([
        left + " is to the left of " + right,
        right + " is to the right of " + left,
      ])
      relations.extend([
        front + " is in front of " + back,
        back + " is behind " + front,
      ])
  shuffle(relations)
  statements.extend(relations)

  # Clothing and action
  for h in humans:
    # Action
    s_action = f"The {h.name} {h.action}."
    statements.append(s_action)

    # Clothing
    for s in h.clothe:
      s_clothe = f"The {h.name} {s.strip()}."
      statements.append(s_clothe)

  return " ".join(statements).strip()
```

Evidently, dense captions tend to be way too descriptive and hence noisy to be used fully in VL training. Therefore, we add a sampling option where statements are sampled with certain probabilities following their weights. For example, instead of mentioning all pairwise positional relations, this option allows sampling a number of sentences from the positional relations category.

G. LLM-Based Caption Paraphrasing

We additionally experiment with using the rule-based system to guide the use of large language models for caption generation / paraphrasing. Specifically, we adapt the deterministically-generated caption (as detailed in Section F) into a prompt for instruction-based text completion by replacing the prefix "This scene contains" (in the synthesized captions) to "Please describe a scene containing" and adding a suffix for text completion: "In this scene, we can see". We use the adapted texts as prompts for language models and generate text completions. We limit the generated texts to 150 tokens and use caption split averaging, as described in Section 3.2 in the main paper. We experiment with the Bloomz 7.1B [10] and Flan-T5 XXL [3] through Huggingface [15].

Table A.4 shows the performance of syn-CLIP trained using different caption generation mechanisms. We do not observe any significant performance gains when using the captions generated by openly avaialble language models that we tried over the rule-based system. This is indeed expected, as the captions generated by current open language models tend to repeat much of the content in the prompt, often correcting verb tenses or adding appropriate punctuation marks, which don't contribute to the semantic richness of the caption.

	VL-Checklist		
	Relation	Attribute	**Average**
CLIP	63.57	67.51	65.54
syn-CLIP with Rule-Based	69.39	70.37	**69.88**
syn-CLIP with Bloomz	**69.66**	69.69	69.68
syn-CLIP with Flan-T5	68.48	**71.14**	69.81

Table A.4. VL-Checklist [18] performance on variants of syn-CLIP fine-tuned on **SyViC** with captions generated using the rule-based system described in section 2 or using language models as described in section 3.

However, we remark that additional work on using LLMs for caption generation could investigate more powerful language models, or the use of visual grounding for caption generation as an additional information source, to yield better paraphrasing / captions.

H. Exploration into Synthetic Data Diversity

As promised in Ablations Section 4.4 of the main paper (lines 740-741 in "**SyViC** - number of models and number of samples") we include the effect of the number of synthetic samples and the number of object models used for **SyViC** generation analysis in Figure A.1. These ablations were not included in the main paper due to lack of space. As we can see the performance is improving consistently, both with adding more synthetic images (Fig. A.1a) and with adding more 3D models used for synthesis (Fig. A.1b).

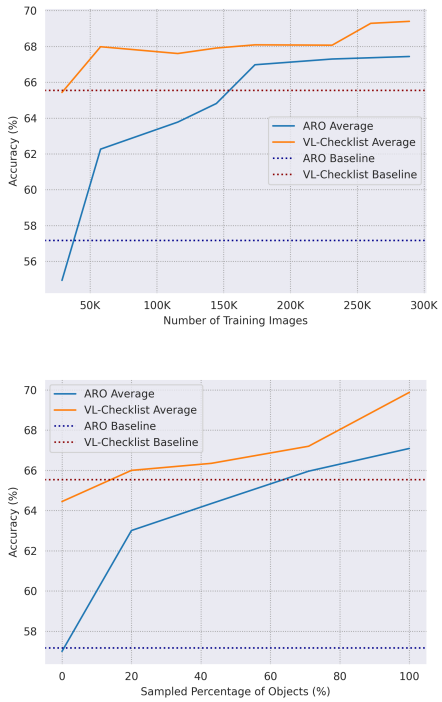


Figure A.1. Exploration into Synthetic Data Diversity. (a) effect of adding more synthetic samples to **SyViC**; (b) effect of adding more 3D object models to **SyViC**. Comparing base CLIP and syn-CLIP performances on VL-Checklist and ARO benchmarks.

I. Some Qualitative Examples with Synthetic Humans

In this section, we first showcase qualitative improvement in the compositional capabilities of CLIP after finetuning on SyViC using our proposed approach via GradCAM in Figure A.2. Next, we show textured SMPL samples in Figure A.3.

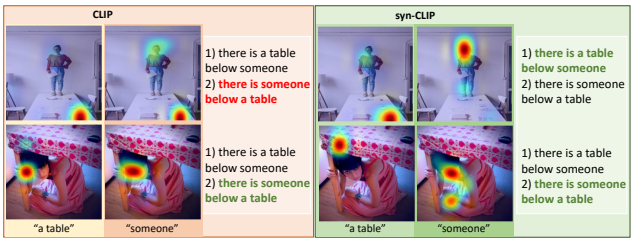


Figure A.2. GradCAM on a Winoground sample. *left* - a CLIP model attending incorrectly to table regions for both `a table` and `someone` text queries, making a mistake in prediction (red text). *right* - our syn-CLIP model correctly attends to the same making no mistakes in prediction with respect to the given inputs. Best viewed in color.

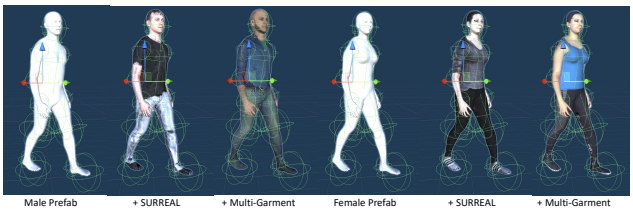
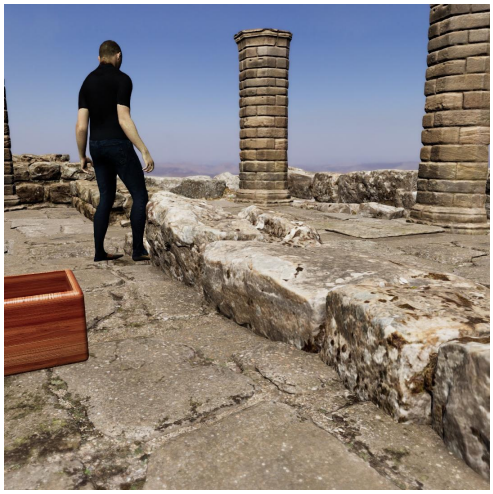


Figure A.3. Digital humans. We show male and female samples of Unity Prefabs containing SMPL templates, a set of 514 reusable 3D object assets available in **SyViC**. We add colliders to each model to allow interactions with other objects.

Finally, we showcase some visual examples from **SyViC** along with the dense captions we generate describing human actions and detailed human-object interactions and relative position descriptions in the following pages.



This scene contains a box, and one human. They are in a castle ruin with old stones. The box is to the left of the human. The box is in front of the human. The human rotate jump. The human is male. The human wears a black t-shirt and dark blue jeans.

This scene contains a cat, a car tire, and one human. They are in an abandoned shell of a factory with a grey floor that is partially tiled with yellow tiles. The cat is to the right of the car tire. The human is behind the cat. The human is to the right of the car tire. The car tire is in front of the cat. The human stand. The human has black pants on. The human is wearing a black long-sleeve shirt with an orange logo. The human is male. The human has curly blue and black hair.

This scene contains a basket, a suitcase, and one human. They are in a room with a red tiled floor and white walls. The suitcase is in front of the human. The suitcase is to the right of the human. The basket is to the left of the suitcase. The suitcase is in front of the basket. The basket is to the left of the human. The human hand up high jump. The human is wearing an olivine beanie on the head. The human is male. The human has a long beard. The human wears a purple jacket, jeans pants, and brown shoes.

This scene contains a window, a table, a car tire, and two humans. They are in a street with grey floor, green plants on the side, and houses around. The car tire is in front of the third human. The first human is behind the table. The car tire is to the left of the table. The car tire is behind the table. The window is to the left of the first human. The car tire is to the left of the first human. The window is behind the first human. The second human is to the right of the table. The table is in front of the second human. The second human is to the right of the car tire. The car tire is behind the first human. The window is behind the table. The window is to the left of the second human. The window is to the left of the table. The window is behind the second human. The second human is in front of the car tire. The second human body actions. The second human wears a white t-shirt and blue jeans with a black belt. The second human has short hair and a short beard. The second human is male. The first human dance. The first human is male. The first human has black hair. The first human wears a white t-shirt and blue jeans.

This scene contains a shirt, a stool, a fire hydrant, and two humans. They are in a room with blue floor and white walls. The shirt is behind the stool. The shirt is behind the second human. The shirt is to the right of the second human. The stool is to the left of the shirt. The second human is to the left of the first human. The first human is to the right of the fire hydrant. The second human is in front of the fire hydrant. The first human is in front of the shirt. The first human is in front of the fire hydrant. The stool is in front of the first human. The second human is to the left of the fire hydrant. The fire hydrant is behind the stool. The first human is to the right of the stool. The second human is in front of the stool. The stool is to the left of the fire hydrant. The first human is behind the second human. The second human dribble basketball. The second human wears a green football jersey and blue jeans pants. The second human has dark brown hair. The second human is male. The first human move backwards. The first human is dressed in a white shirt and dark blue jeans pants. The first human is male. The first human has a shaved beard and is wearing a brown beret.

This scene contains two children and two humans. They are in a room with a red tiled floor and white walls. The second human is in front of the first human. The first human is behind the second human. The first human one side leg walk and search. The second human has a light brown sweater on. The second human is male. The second human is standing. The second human is male. The second human is wearing a grey sweater and dark blue jeans. The second human has black hair. The second human has dark grey shoes. The first human ballet dance. The first human has white boots. The first human has black hair. The first human wears a black t-shirt and grey pants. The first human is male.

This scene contains a gift wrapping, a water faucet, and one human. They are in a room with a red tiled floor and white walls. The gift wrapping is behind the water faucet. The human is in front of the gift wrapping. The human is to the right of the gift wrapping. The water faucet is behind the human. The water faucet is to the right of the gift wrapping. The human sleepwalk. The human is wearing a dark grey t-shirt and black sports shorts. The human has short blonde hair. The human is male.

This scene contains a loudspeaker, a hat, a dog, a gas cooker, a gas cooker, a table, and one human. They are in a room with a red tiled floor and white walls. The hat is to the right of the dog. The gas cooker is to the right of the gas cooker. The human is to the left of the gas cooker. The loudspeaker is to the left of the hat. The loudspeaker is behind the gas cooker. The dog is behind the gas cooker. The gas cooker is in front of the loudspeaker. The gas cooker is in front of the gas cooker. The hat is to the right of the human. The hat is in front of the gas cooker. The gas cooker is to the left of the human. The dog is behind the hat. The loudspeaker is behind the table. The table is in front of the gas cooker. The gas cooker is to the right of the dog. The human is in front of the dog. The gas cooker is behind the human. The table is to the right of the loudspeaker. The dog is to the right of the loudspeaker. The gas cooker is to the left of the dog. The table is in front of the dog. The table is to the left of the gas cooker. The table is to the left of the hat. The hat is in front of the gas cooker. The human is in front of the loudspeaker. The human is in front of the gas cooker. The gas cooker is to the left of the table. The hat is in front of the loudspeaker. The loudspeaker is to the left of the human. The loudspeaker is behind the dog. The gas cooker is to the right of the loudspeaker. The gas cooker is behind the table. The human is behind the table. The hat is to the right of the gas cooker. The table is in front of the hat. The human hand up high jump. The human has a pair of green and white shoes. The human is wearing a black hoodie and dark blue jeans pants. The human is female. The human has her head covered by a beige head scarf.

References

[1] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. 2022.

[2] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[4] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[5] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, et al. Teaching structured vision&language concepts to vision&language models. *arXiv preprint arXiv:2211.11733*, 2022.

[7] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

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

[10] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2022.

[11] 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.

[12] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

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

[14] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.

[16] 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*, 2023.

[17] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[18] 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.