

Enhancing Vision-Language Compositional Understanding with Multimodal Synthetic Data

Haoxin Li and Boyang Li
 Nanyang Technological University
 {haoxin003, boyang.li}@ntu.edu.sg

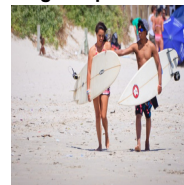
Abstract

Paired image-text data with subtle variations in-between (e.g., people holding surfboards vs. people holding shovels) hold the promise of producing Vision-Language Models with proper compositional understanding. Synthesizing such training data from generative models is a highly coveted prize due to the reduced cost of data collection. However, synthesizing training images for compositional learning presents three challenges: (1) efficiency in generating large quantities of images, (2) text alignment between the generated image and the caption in the exact place of the subtle change, and (3) image fidelity in ensuring sufficient similarity with the original real images in all other places. We propose SPARCL (*S*ynthetic *P*erturbations for *A*dvancing *R*obust *C*ompositional *L*earning), which integrates image feature injection into a fast text-to-image generative model, followed by an image style transfer step, to meet the three challenges. Further, to cope with any residual issues of text alignment, we propose an adaptive margin loss to filter out potentially incorrect synthetic samples and focus the learning on informative hard samples. Evaluation on four compositional understanding benchmarks demonstrates that SPARCL significantly improves the compositionality of CLIP, boosting the average accuracy of the CLIP base model by over 8% across all benchmarks and outperforming state-of-the-art methods by 2% on three benchmarks.

1. Introduction

Current Vision-Language Models (VLMs) still face limitation in accurately interpreting compositional relationships between objects and attributes, as demonstrated by numerous evaluations [32, 62, 88, 106, 113]. This limitation primarily stems from the absence of subtle variations in the training data [42] (e.g., the subtle variations between the two captions in Figure 1 (a)). As a result, it becomes possible to maximize empirical image-caption alignment us-

Source caption: Two people carrying surfboards on a beach.
Target caption: Two people carrying shovels on a beach.



Source image



Edited image



Generated image

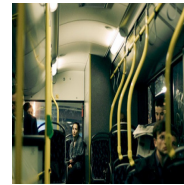
(a) Difficulty in accurately creating precise variations.



+ Real positive caption:
 A man is stirring a silver pot filled with food.
- Synthetic negative caption:
 A man is stirring a copper pot filled with food.



Hard Sample



+ Real positive caption:
 Random people sitting in a public transport bus.
- Synthetic negative caption:
 Random animals riding a purple elephant.



Easy Sample



+ Real positive caption:
 A double-decker bus parked at a station.
+ Synthetic positive caption:
 A double-decker bus parked at a bus station with a few passengers getting on.



Wrong Generation

(b) Inconsistency in cross-modal alignment quality of synthetic samples.

Figure 1. Challenges in generating and training on synthetic data: (a) When generating an image with subtle variations based on a real image and a target caption specifying the variations, an image editing model [6] struggles with text alignment (middle), while an image generation model [75] fails to maintain image fidelity (right). (b) Synthetic positive and negative image-caption pairs show different levels of alignment quality. The subtle variations in the synthetic negative caption (left) make it difficult to distinguish from the positive; the over-modified negative caption (middle) is easy to distinguish; and the hallucinated content in the synthetic positive caption (right) results in an incorrect positive.

ing shortcut features [23] rather than genuinely learning nuanced distinctions. While collecting training samples with subtle variations could enhance compositionality, this approach is time-consuming and labor-intensive, rendering it impractical at scale.

Advances in generative models [7, 12, 20, 43, 61, 75,

78, 91, 105, 110] now facilitate the synthetic generation of training samples with subtle variations between paired samples [11, 18, 49, 68, 71, 79, 83, 86, 106, 108, 109]. By starting with real image-caption pairs, generative models can create subtle edits to both captions and images, providing valuable training data with minimal manual effort. These generated variations enable VLMs to enhance their compositionality by learning from the nuanced differences.

However, generating and training on synthetic data presents two key challenges. The first is the difficulty of efficiently and accurately creating precise variations in synthetic images. To generate large-scale training data with precise variations, the image generation process must meet three criteria: *efficiency* in producing large quantities of images, *text alignment* of the generated image to the corresponding caption, which is subtly changed from the original caption, and *image fidelity* in ensuring the generated image otherwise closely matches their real counterpart. Nevertheless, current image generation methods struggle to meet all three criteria simultaneously. Per-sample optimization methods [44, 82, 94, 112] lack efficiency, while zero-shot image editing methods [6, 22, 63] and text-to-image (T2I) models [75, 78] often fail to achieve proper text alignment and image fidelity, respectively, as illustrated in Figure 1 (a). The inaccurate synthetic variations produced by these models can mislead the learning of VLMs. To alleviate this issue, we take a fast T2I model, which excels in efficiency and text alignment, and inject real image features into the text prompt features in order to enhance image fidelity. By combining this approach with AdaIN [35], we manage to substantially improve the image generation process per the three criteria.

Still, the generated data do not *perfectly* meet the text alignment criterion, and we propose to deal with the remaining problems with an innovative approach to model training. As Figure 1 (b) shows, the similarity between synthetic positive and negative pairs may vary, resulting in a mixture of hard text-image pairs, easy pairs, and incorrect pairs. Instead of treating all positive and negative samples uniformly, we propose a novel loss that differentiates between positive, hard negative, and easy negative samples. Further, we propose an adaptive margin that helps to filter out potentially incorrect generations and focus the learning on informative hard samples.

In summary, we propose SPARCL, which integrates image feature injection into a fast T2I model to improve the quality of synthetic variations in images and employs an adaptive margin loss to leverage the varying alignment quality in synthetic samples for training VLMs. Evaluations on four compositional understanding benchmarks demonstrate that SPARCL significantly enhances VLM compositionality, improving CLIP by over 5% on VL-CheckList [113] and 7% on SugarCrepe [32], while surpassing state-of-the-

art methods by 1% and 2% on the two benchmarks. The main contributions of this paper are as follows: (1) we propose image feature injection to enhance the quality of synthetic variations in images, which provide valuable training data to improve the compositionality of VLMs; (2) we introduce an adaptive margin loss to leverage varying levels of cross-modal alignment in synthetic samples to effectively differentiate positive and negative samples; (3) experimental results validate that SPARCL significantly improves the compositional understanding capabilities of CLIP models.

2. Related Work

2.1. Limitations in Compositionality of VLMs

While VLMs excel in many multi-modal tasks [40, 52, 53, 74, 84, 101, 115], they still struggle with compositional understanding—the ability to interpret novel combinations of known visual and textual components. Benchmarks like What’sUp [42] reveal difficulties in understanding spatial relationships, SPEC [70] highlights issues with object size, position, and count, and ARO [106] uncovers limitations in understanding attributes, relations, and word order. Winoground [88], SNARE [96], and VL-CheckList [113] also expose these shortcomings. SugarCrepe [32] addresses hackable biases in prior benchmarks, where text-only models achieve artificially high performance, by introducing fluent and meaningful hard negatives. Their findings suggest that previous benchmarks overestimated compositional understanding. Building on this, SugarCrepe++ [19] further introduces semantically equivalent but lexically varied captions as hard positives and shows the difficulties of VLMs in distinguishing between lexical variations.

2.2. Improving Compositionality of VLMs

Prior approaches to improving the compositionality of VLMs can be broadly classified into the following categories: (1) *Leveraging detailed image captions*: Detailed captions from dense captioning models [17, 54], simulation platforms [9], and video annotations [45] are collected to train VLMs. However, these samples often lack pairs with subtle variations, limiting their contribution to compositionality. (2) *Distilling from pretrained models*: SDS-CLIP [3], SF-CLIP [80] and IL-CLIP [114] distill knowledge from pretrained image generation models and visual-language foundation models. IL-CLIP [114] refines representations through iterative learning with pretrained vision and language agents. However, pretrained models also face limitations in compositionality [34, 90]. (3) *Incorporating structural knowledge*: MosaiCLIP [85] and StructureCLIP [36] incorporate scene graph knowledge in text features. CLIP-SGVL [28] and 3VL [103] train VLMs to predict scene graphs. [66] utilizes scene graphs as prompts to elicit compositional knowledge from VLMs without fur-

ther training. However, these methods rely on models trained with expensive dense structure annotations (e.g., scene graphs). (4) *Utilizing synthetic negative samples*: Rule-based tools [30] or large language models (LLMs) [15, 73] are used to generate negative captions by editing real captions [10, 11, 18, 68, 83, 86, 106, 109]. Image generation models [75] are also used to create or edit images for training [49, 71, 79, 108]. Despite their utility, efficiently generating large amount of images with precise variations remains challenging due to the inherent limitations of generative models. (5) *Applying fine-grained alignment constraints*: MCD [46] enforces multi-scale alignment across images with varying augmentations and the corresponding text captions. SPARC [5] learns local alignment by associating each text token with a group of local image patches. CE-CLIP [109] applies intra-modal contrastive loss and cross-modal ranking loss to improve alignment. However, uniform supervision applied to synthetic samples with varying alignment quality limits their effectiveness. In this paper, we propose SPARCL to address two key challenges in learning from synthetic data: the difficulty in accurately creating precise variations and the inconsistency in cross-modal alignment quality in synthetic data.

2.3. Training with Synthetic Data

Synthetic data for training machine learning models have been studied in various fields [14, 26, 48, 64, 67, 76, 77, 93, 99]. Synthetic data generated through simulations and graphics engines supports a wide range of tasks [16, 72, 95]. However, these synthetic datasets often diverge significantly from real-world data. Recent advances in generative models [7, 12, 43, 61, 75, 78, 91, 105, 110] have made it possible to synthesize data that more closely resembles real-world scenarios. Synthetic data are widely used in both language tasks [21, 31, 60, 64, 65, 87, 97, 98, 100, 116] and vision tasks [1, 2, 4, 25, 29, 39, 57, 81, 89, 92, 111]. Synthetic data are often noisy, necessitating noise-resistant training methods [13, 37, 41, 50, 51, 102], especially those designed for contrastive learning [38, 56]. In this paper, we generate multimodal samples with subtle variations and filter out potentially incorrect ones during training to improve the compositionality of VLMs.

3. SPARCL

To improve the compositional understanding abilities of VLMs, we propose Synthetic Perturbations for Advancing Robust Compositional Learning, or SPARCL, which generates multimodal samples with subtle variations from real samples and trains VLMs to learn nuanced differences through synthetic data. In the generation phase, SPARCL creates positive and negative captions with slight variations from real captions using an LLM, then generates images based on the real image and these modified captions through

a fast T2I model. To enhance the quality of subtle variations in synthetic images, we introduce image feature injection, which integrates real image features into the text prompt features of the T2I model to improve fidelity to the real image, as detailed in Sec. 3.2. In the training phase, SPARCL employs an adaptive margin loss that leverages varying levels of multimodal alignment in the synthetic samples to effectively learn informative nuanced distinctions, as described in Sec. 3.3. The framework of SPARCL is shown in Figure 2.

3.1. Generating Negative and Positive Captions

Captions with subtle variations are crucial for learning compositional knowledge, as shown by previous work that generates captions by randomly swapping or replacing nouns and adjectives [106, 109]. However, manually designed generation rules often introduce nonsensical or grammatically incorrect artifacts, creating shortcut features [23] that obstruct true compositional understanding [32]. To address this issue, we use an LLM to generate natural synthetic captions. To further mitigate the impact of generative artifacts, we generate both negative and positive captions, ensuring that VLMs cannot easily differentiate negative captions from positive ones based solely on artifacts. We denote the i^{th} real image-caption pair in the training set as (I_i^r, T_i^r) . Given a real caption T_i^r , we prompt the LLM to generate a synthetic negative caption T_i^{sn} and a synthetic positive caption T_i^{sp} , as specified by the prompts in Figure A1 in the Appendix.

3.2. Generating Images via Image Feature Injection

Synthetic images that exhibit subtle variations from real images while aligning with the captions are valuable but challenging to generate, as discussed in Sec. 1. Although previous works [49, 71, 79, 108] have utilized object segmentation or filtered dissimilar generations to improve fidelity to real images, they struggle with manipulating relationships or lack efficiency. We aim to enhance fidelity by injecting image features into a fast T2I model [61], which already achieves high efficiency and text alignment.

Image Feature Injection. The images generated by T2I models lacks image fidelity to real images, as no real image information is input to the models. To enhance the fidelity of synthetic images, we inject real image features into the text prompt features, enabling the model to incorporate information from the real images.

In T2I models, content and style are separated in the semantic and padding embeddings. The semantic embeddings (before the embedding of the [EOS] token) usually capture most of the image content in the text prompts, while the padding embeddings (after the [EOS] token) usually represent the image style [104]. Therefore, we can inject real image features into the padding embeddings to guide

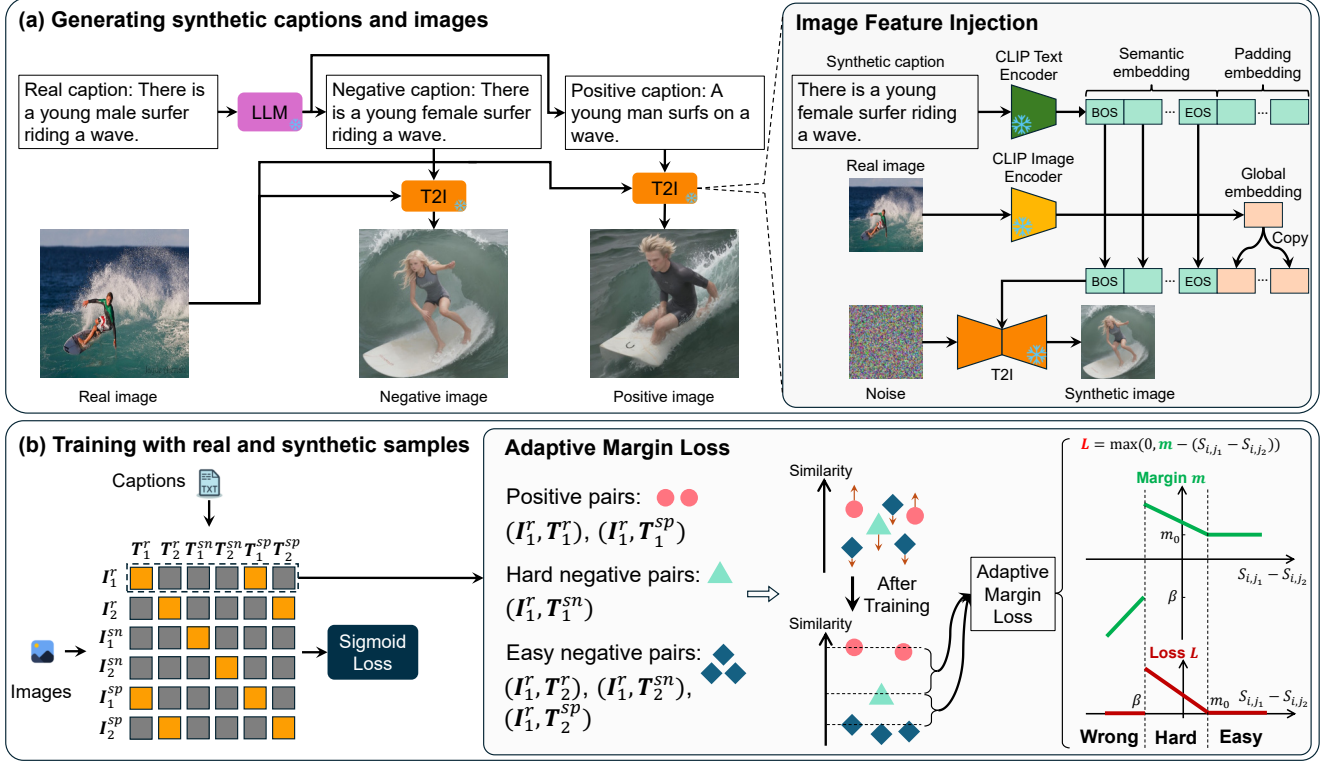


Figure 2. An overview of SPARCL. (a) Starting with a real image-caption pair, we generate synthetic positive and negative pairs with subtle variations using an LLM and a fast T2I model. To improve the quality of subtle variations in synthetic images, we introduce image feature injection to reduce unintended variations from a standard T2I model (see Sec. 3.2). (b) We train the VLM using both real and synthetic samples. In addition to a sigmoid loss for distinguishing positive and negative pairs, we apply an adaptive margin loss that leverages varying alignment levels across training samples to learn informative nuanced distinctions (see Sec. 3.3).

the model in generating images with a similar style to real images, without affecting the alignment with the captions, thanks to the decoupling of content and style in the prompt embeddings.

We extract features for a real image I_i^r and the corresponding synthetic caption T_i^s ($s \in \{sn, sp\}$) using two aligned feature encoders (e.g., CLIP image and text encoders). The image embedding f_i^r is the encoder output at the position of the [CLS] token. The text embedding is the text encoder output $e_i^s = \langle e_{i,j}^s \rangle_{j=1}^L$, where L is the maximum sequence length. Let the index of the [EOS] token in e_i^s be k_i^s , $k_i^s \leq L$. That is, the padding token [PAD] is used as input to the text encoder at positions between k_i^s and L . The key step of SPARCL is to replace the text embeddings at those positions with the image embedding, producing a new sequence of text embeddings, $\hat{e}_i^s$,

$$\hat{e}_i^s = \langle e_{i,1}^s, e_{i,2}^s, \dots, e_{i,k_i^s}^s, \underbrace{f_i^r, \dots, f_i^r}_{L-k_i^s \text{ times}} \rangle. \quad (1)$$

The embeddings $\hat{e}_i^s$ is then used as input to the T2I model to generate a synthetic image $\tilde{I}_i^s$. This process, depicted in Figure 2, reduces unintended variations in synthetic images generated by the standard T2I model and enhances image fidelity to real images, thereby improving the quality of

synthetic subtle variations.

Style Transfer. To further reduce the domain gap between synthetic and real images, we use AdaIN [35] to transfer the style of the synthetic image to that of the real image. We use a pretrained AdaIN encoder to extract content features from $\tilde{I}_i^s$ and style features from I_i^r . We normalize the content features using instance normalization and then scale and shift them with the mean and variance of the style features. The transformed features are fed into a pretrained AdaIN decoder to generate I_i^s , which is subsequently used for model training.

3.3. Training with Real and Synthetic Samples

After generating captions and images, each real image-caption pair (I_i^r, T_i^r) is extended to $(I_i^r, T_i^r, I_i^{sn}, T_i^{sn}, I_i^{sp}, T_i^{sp})$ by adding one synthetic negative and one synthetic positive pair. We then train the VLM using these extended samples. Given a batch of n sample groups, $\{(I_i^r, T_i^r, I_i^{sn}, T_i^{sn}, I_i^{sp}, T_i^{sp})\}_{i=1}^n$, we organize the images by concatenating all real, synthetic negative, and synthetic positive images into an image batch I^B of $3n$ images, and similar all captions into a caption

batch $\mathbf{T}^B$ of $3n$ captions.

We calculate the similarity between each image-caption pair in the batch. The similarity between the i^{th} image $\mathbf{I}_i^B$ and the j^{th} caption $\mathbf{T}_j^B$ is given by $S_{i,j} = c(\mathcal{E}_I(\mathbf{I}_i^B), \mathcal{E}_T(\mathbf{T}_j^B))$, where $c(\cdot, \cdot)$ denotes cosine similarity, and $\mathcal{E}_I$ and $\mathcal{E}_T$ represent the image and text encoders of the VLM, respectively. We then define the ground-truth alignment variable $M_{i,j}$, which takes value 1 if the image matches with the caption, and -1 otherwise. Clearly, the only positive image-caption pairs are $(\mathbf{I}_i^r, \mathbf{T}_i^r)$, $(\mathbf{I}_i^{sn}, \mathbf{T}_i^{sn})$, $(\mathbf{I}_i^{sp}, \mathbf{T}_i^{sp})$, $(\mathbf{I}_i^r, \mathbf{T}_i^{sp})$ and $(\mathbf{I}_i^{sp}, \mathbf{T}_i^r)$. To account for multiple positive associations of an image or a caption, we apply a sigmoid-based contrastive loss [8, 107] to encourage higher similarity for positive pairs and lower similarity for negative pairs,

$$L_{con} = -\frac{1}{3n} \sum_{i=1}^{3n} \sum_{j=1}^{3n} \log \frac{1}{1 + \exp(-M_{i,j}(S_{i,j}/\tau + b))}, \quad (2)$$

where τ is the temperature parameter and b is a bias term.

Adaptive Margin Loss. Despite best efforts toward controlling generative models, synthetic samples may still have variable quality in terms of the alignment between text and imagery. That is, purported positive (resp. negative) pairs may not be semantically similar (resp. dissimilar). To account for variation in synthetic data quality, we propose to differentiate between real data and synthetic data and between easy negatives and hard negatives. Further, we propose an adaptive margin that filters out potential incorrect samples and prioritizes learning from hard samples.

For each image in a batch, we define four sets of captions, a positive set $\mathbb{P}$, a hard negative set $\mathbb{N}_h$, an easy negative set $\mathbb{N}_e$, and a real negative set $\mathbb{N}_r$, which represent different levels of alignment within the batch. If the image is a real or synthetic positive image,

- $\mathbb{P}$ contains the real positive captions $\mathbf{T}_i^r$ and the synthetic positive captions $\mathbf{T}_i^{sp}$ of the current, i -th image.
- $\mathbb{N}_h$ contains the synthetic negative captions $\mathbf{T}_i^{sn}$ of the current image.
- $\mathbb{N}_e = \{\mathbf{T}_j^r, \mathbf{T}_j^{sp}, \mathbf{T}_j^{sn} | i \neq j\}$ contains all real and synthetic captions from all other images.
- $\mathbb{N}_r = \{\mathbf{T}_j^r | i \neq j\}$ contains all real captions belonging to the other images.

If the image is a synthetic negative image, the sets $\mathbb{N}_e$ and $\mathbb{N}_r$ are unchanged but $\mathbb{P}$ and $\mathbb{N}_h$ differ:

- $\mathbb{P}$ contains the synthetic negative captions $\mathbf{T}_i^{sn}$.
- $\mathbb{N}_h$ contains $\mathbf{T}_i^r$ and $\mathbf{T}_i^{sp}$.

We define the margin loss for the i^{th} image as:

$$\begin{aligned} L_{mar,i}^I &= \frac{1}{|\mathbb{P}| \cdot |\mathbb{N}_h|} \sum_{j_1 \in \mathbb{P}, j_2 \in \mathbb{N}_h} \max(0, m + S_{i,j_2} - S_{i,j_1}) \\ &+ \frac{1}{|\mathbb{N}_h| \cdot |\mathbb{N}_e|} \sum_{j_1 \in \mathbb{N}_h, j_2 \in \mathbb{N}_e} \max(0, m + S_{i,j_2} - S_{i,j_1}) \quad (3) \\ &+ \frac{\alpha}{|\mathbb{P}| \cdot |\mathbb{N}_r|} \sum_{j_1 \in \mathbb{P}, j_2 \in \mathbb{N}_r} \max(0, m + S_{i,j_2} - S_{i,j_1}). \end{aligned}$$

This loss encourages positive pairs to have higher similarity scores than hard negative pairs, which should have higher similarity than easy negative pairs. Additionally, a weight $\alpha > 1$ is applied to comparisons between positive pairs and real negative pairs to emphasize these comparisons, as real samples are generally correct, making comparisons involving them more reliable.

Lastly, we propose an adaptive margin in the above loss. Let $d = S_{i,j_1} - S_{i,j_2}$ denote the difference between two similarity scores in Eq. (3). The adaptive margin m is then computed as follows:

$$m = \begin{cases} d, & d < \beta \\ (\frac{m_0 - d}{m_0 - \beta} \gamma + 1)m_0, & \beta \leq d \leq m_0 \\ m_0, & d > m_0 \end{cases} \quad (4)$$

where m_0 is a base margin, $\beta < 0$ is a cutoff threshold, γ is a scaling factor. The adaptive margin is designed with the following rationale: when $d < \beta$, the purported positive pair is much less similar than the purported negative pair, suggesting the presence of incorrect or mislabeled samples, so we zero out the loss by setting the margin to d . For $\beta \leq d \leq m_0$, the margin is scaled, with smaller differences receiving larger margins to emphasize learning from harder samples. When $d > m_0$, the samples are already well-separated with the margin m_0 , so the margin is capped at m_0 , again leading to zero loss. We visualize both the adaptive margin and the corresponding loss as functions of d in Figure 2 (b).

The loss for all images is given by $L_{mar}^I = \frac{1}{3n} \sum_i L_{mar,i}^I$. Analogously, we compute L_{mar}^T for all captions. The total adaptive margin loss is $L_{mar} = L_{mar}^I + L_{mar}^T$. Finally, the overall training loss is a weighted combination of the contrastive loss and the adaptive margin loss, with weight λ :

$$L = L_{con} + \lambda L_{mar}. \quad (5)$$

4. Experiments

4.1. Datasets

Training. We use the COCO-2014 dataset [55] as the training data source. The training set consists of 82,783 images, each paired with five captions. For each real image-caption pair, we generate one positive and one negative synthetic pair. In line with previous approaches [18, 106, 109], we train the VLMs using both the original COCO-2014 training data and the synthetic samples.

Evaluation. We evaluate our model on four vision-language compositional understanding benchmarks: (1) ARO [106], which consists of 23,937 cases for relation understanding and 28,748 for attribute understanding. We exclude the subsets for order understanding, as they contain significant nonsensical and non-fluent artifacts [32].

Table 1. Comparison of accuracy (%) between SPARCL and baselines on four benchmarks. “img.” represents images, “cap.” represents captions, “syn.” represents synthetic data.

Method	Training Data					ARO	VL-CheckList	SugarCrepe	SugarCrepe++
	Source	# real img.	# real cap.	# syn. img.	# syn. cap.				
CLIP [74] (Zero-Shot)	-	-	-	-	-	61.1	73.2	73.4	59.8
CLIP [74] (Finetune)	COCO	82K	410K	0	0	64.1	72.8	79.9	62.3
SDS-CLIP [3]	COCO	82K	410K	0	0	57.5	-	-	-
[79]	COCO	0	0	82K	82K	65.0	69.9	-	-
AMR-NegCLIP [83]	COCO	100K	100K	0	500K	79.4	-	85.2	-
NegCLIP [106]	COCO	100K	100K	0	500K	76.0	74.6	82.5	64.9
MosaiCLIP [85]	COCO	109K	109K	0	981K	80.3	76.8	-	-
FSC-CLIP [68]	COCO	100K	100K	0	1.5M	-	77.2	85.1	-
CE-CLIP [109]	COCO	82K	410K	0	2M	79.7	76.3	85.2	-
COMO [49]	COCO	113K	567K	567K	567K	-	76.9	-	-
SPARCL (our method)	COCO	82K	410K	820K	820K	77.2	79.2	87.1	66.1
SPEC [70]	LAION	20K	20K	20K	20K	70.1	-	-	-
CounterCurate [108]	Flickr	30K	30K	150K	150K	-	-	82.8	-
[18]	CC3M	3M	3M	0	9M	-	75.3	-	55.3
CE-CLIP+ [109]	COCO+CC3M	3M	3M	0	15M	80.4	79.3	87.5	-
CLOVE [11]	LAION-COCO	>1B	>1B	0	>1B	73.2	-	85.1	-
IL-CLIP [114]	CC12M	12M	12M	0	0	-	-	70.3	-
SF-CLIP [80]	YFCC15M	15M	15M	0	0	-	-	71.2	-
syn-CLIP [9]	SyViC	0	0	>1M	>1M	69.2	74.8	-	-
FiGCLIP [45]	VidSitu	20K videos		0	0	67.0	-	74.6	-

(2) *VL-CheckList* [113], a large-scale benchmark with over 100,000 samples, evaluates compositionality across subsets of objects, attributes, and relationships, which are further divided into various fine-grained categories. (3) *SugarCrepe* [32] includes 7,000 test cases across seven subsets. In the above three benchmarks, each test case containing one image, one positive caption, and one negative caption. (4) *SugarCrepe++* [19] includes 4,757 test samples across five subsets, where each test case consists of one image, two positive captions, and one negative caption. All benchmarks involve classifying captions as positive or negative for the given images. We report the average accuracy across all subsets of each benchmark and include the accuracy for each subset in Appendix A3.

4.2. Implementation Details

Models. We use ViT-B/32 and ViT-L/14 architectures from OpenAI’s CLIP model [74] as our base models, initialized with pretrained checkpoints. Following syn-CLIP [9], we integrate LoRA adapters [33] into both the image and text encoders of CLIP to improve training efficiency and mitigate knowledge forgetting. Only the LoRA adapters are fine-tuned during training.

Training Setups. We use the AdamW optimizer [59] with a cosine learning rate schedule [58]. Training is conducted on two Tesla V100 GPUs, with a batch size of 128 sample groups for ViT-B/32 and 16 for ViT-L/14. The base learning rate is set to 0.01 for a total batch size of 256, and scaled

linearly [24] based on the actual batch size. Training is performed for 3,000 steps for ViT-B/32 and 15,000 steps for ViT-L/14, corresponding to fewer than 5 epochs in previous studies [18, 106, 109]. More details and hyperparameters settings are in Appendix A2.

4.3. Main Results

We compare SPARCL with several baseline methods: (1) methods that utilize synthetic samples for training [49, 68, 70, 79, 83, 106, 108, 109], (2) methods that distill knowledge from pretrained models [3, 11, 80, 114], (3) methods that incorporate knowledge from scene graphs [18, 85], and (4) methods that leverage detailed image captions [9, 45]. Both SPARCL and the baseline methods use ViT-B/32 as the base model. The results are presented in Table 1.

We observe that SPARCL achieves the best performance on VL-CheckList, SugarCrepe, and SugarCrepe++ compared to baselines trained on the same data source, namely COCO. Specifically, SPARCL surpasses the strongest baseline by 1.4% on VL-CheckList and 2.5% on SugarCrepe, underscoring its effectiveness in enhancing compositional understanding. However, SPARCL performs worse on ARO than some baselines. We hypothesize that this may be due to nonsensical or grammatically incorrect artifacts in the ARO test samples [32], which could favor methods that utilize rule-based synthetic training samples (*e.g.*, [85]) containing similar artifacts. Notably, even with only 50% of the training data, SPARCL outperforms most base-

Table 2. Ablated performance (%) of SPARCL. “SynCap” refers to synthetic captions, “SynImg” refers to synthetic images, “FeatInj” denotes image feature injection, and “CompSet” indicates the comparison sets used in Eq. (3).

Model	Variant	SynCap	SynImg	FeatInj	AdaIN	Adaptive Margin Loss		ARO	VL-CheckList	SugarCrepe	SugarCrepe++	Average
						CompSet	Margin					
ViT-B/32	#1	✗	✗	✗	✗		✗	60.49	72.61	79.36	64.85	69.33
	#2	✓	✗	✗	✗		✗	71.77	73.52	86.35	64.32	73.99
	#3	✗	✓	✗	✗		✗	62.62	71.70	79.97	64.84	69.79
	#4	✓	✓	✗	✗		✗	71.86	75.54	85.43	65.22	74.51
	#5	✓	✓	✓	✗		✗	73.40	76.56	85.54	65.78	75.32
	#6	✓	✓	✗	✓		✗	73.79	75.72	85.79	64.49	74.95
	#7	✓	✓	✓	✓		✗	74.12	76.35	85.40	66.44	75.58
	#8	✓	✓	✓	✓	All	Fixed	76.79	78.59	87.08	65.42	76.97
	#9	✓	✓	✓	✓	All	Adaptive	77.15	79.16	87.11	66.12	77.38
	#10	✓	✓	✓	✓	All	Adaptive Inversed	76.67	79.26	86.68	65.30	76.97
	#11	✓	✓	✓	✓	Only ($\mathbb{P}, \mathbb{N}_h$)	Adaptive	77.48	80.48	86.15	64.70	77.20
ViT-L/14	#1	✗	✗	✗	✗		✗	59.80	73.26	81.49	64.90	69.86
	#2	✓	✗	✗	✗		✗	72.16	75.44	87.38	64.44	74.85
	#3	✗	✓	✗	✗		✗	60.70	72.23	82.58	66.63	70.54
	#4	✓	✓	✗	✗		✗	75.20	78.29	87.29	64.61	76.35
	#5	✓	✓	✓	✗		✗	75.11	78.58	87.54	64.57	76.45
	#6	✓	✓	✗	✓		✗	74.88	79.10	87.50	64.62	76.52
	#7	✓	✓	✓	✓		✗	74.93	80.04	87.79	65.30	77.01
	#8	✓	✓	✓	✓	All	Fixed	75.21	80.75	88.02	66.41	77.60
	#9	✓	✓	✓	✓	All	Adaptive	75.83	80.81	88.23	66.83	77.93
	#10	✓	✓	✓	✓	All	Adaptive Inversed	75.15	80.71	87.93	66.61	77.60
	#11	✓	✓	✓	✓	Only ($\mathbb{P}, \mathbb{N}_h$)	Adaptive	76.26	80.57	87.42	66.28	77.63

line methods (see Table 3 for SPARCL performance with reduced training data). Compared to baselines using additional data sources, SPARCL surpasses or matches their performance. For example, SPARCL outperforms methods in [11, 18] despite their use of more training data. SPARCL performs comparably to CE-CLIP+ [109] on VL-CheckList and SugarCrepe, despite CE-CLIP+ leveraging CC3M as an additional data source and utilizing significantly more samples.

4.4. Ablation Study and Analysis

We perform ablation studies to assess the impact of each component and design choice in SPARCL. The results, presented in Table 2, show the overall performance.

Synthetic captions vs. synthetic images. To analyze the impact of synthetic captions and images, we compare Variant #1 (using only real samples) with Variant #2 (using real samples and synthetic captions) and #3 (using real samples and synthetic images). Variant #2 shows a significant improvement over Variant #1, with the average accuracy improving from 69.33% to 73.99% for ViT-B/32. However, Variant #3 obtains only a marginal improvements of about 0.5%. These results suggest that synthetic captions substantially enhance compositional understanding, while synthetic images alone provide limited benefit, which aligns with findings in [69]. We hypothesize that generative artifacts in synthetic images are more pronounced than in captions, which negatively affects the effective learning of nu-

anced distinctions by VLMs. When both synthetic captions and images are used (Variant #4), performance improves further, indicating a synergistic effect.

Image feature injection vs. AdaIN. To analyze the impact of image feature injection and AdaIN, we compare Variant #4 (without image feature injection or AdaIN) with Variant #5 (with image feature injection only) and #6 (with AdaIN only). Using ViT-B/32, we observe that Variant #5 outperforms Variant #4 across all four benchmarks, with an average gap of 0.8%. While Variant #6 shows a smaller improvement of 0.4%, it outperforms Variant #5 on ARO and SugarCrepe. These results suggest that image feature injection is more effective for improving compositionality than AdaIN, although the two methods are somewhat complementary. Through the improvements obtained using ViT-L/14 are a little bit different, both techniques improve performance over Variant #4. Combining both methods in Variant #7 leads to further improvements, with a more than 1% increase in average accuracy for ViT-B/32 and a 0.65% increase for ViT-L/14 over Variant #4. This highlights the importance of reducing unintended changes in synthetic images, as such changes could cause the model to rely on them for distinguishing positive and negative samples rather than the semantic differences specified in the corresponding captions. Examples of synthetic images in Figure A3 and A4 in the Appendix show how image feature injection helps mitigate these unintended changes.

Effects of the margin loss and adaptive margin strate-

Table 3. Performance (%) of SPARCL using different subsets of training samples. “Neg.” represents synthetic negative samples, “Pos.” represents synthetic positive samples, and “Prop.” represents the proportion of training samples.

Variant	Neg.	Pos.	Prop.	ARO	VL-CheckList	SugarCrepe	SugarCrepe++	Average
#7	✓	✗	100%	72.12	74.65	86.41	57.99	72.79
#7	✓	✓	100%	74.12	76.35	85.40	66.44	75.58
#9	✓	✓	20%	75.90	77.56	85.78	64.44	75.92
#9	✓	✓	50%	77.90	79.38	86.71	64.21	77.05
#9	✓	✓	100%	77.15	79.16	87.11	66.12	77.38

gies. We compare three variants with distinct margin approaches: (1) *Fixed (Variant #8)*: a fixed margin without adaptive adjustments; (2) *Adaptive (Variant #9)*: the margin is calculated according to Eq. (4); and (3) *Adaptive Inversed (Variant #10)*: a larger margin is applied to samples with higher similarity differences, representing an inverse version of Eq. (4) that prioritizes learning from easier samples. Comparing Variant #8 with Variant #7, we observe that #8 improves by over 1% for ViT-B/32 and 0.6% for ViT-L/14 in average accuracy, suggesting that margin loss effectively aids in compositional understanding. Between Variants #8, #9, and #10, we find that Variant #9 outperforms Variant #8 across all four benchmarks, with an average boost of over 0.3% for both ViT-B/32 and ViT-L/14. In contrast, Variant #10 results in a performance decline on three benchmarks and no improvement in average accuracy. These findings highlight the superior effectiveness of adaptive margins.

Is applying adaptive margin loss only to hard samples sufficient? To explore whether applying margin loss solely to hard samples is adequate, we construct Variant #11, which uses only the positive set $\mathbb{P}$ and the hard negative set $\mathbb{N}_h$ for margin loss calculation in Eq. (3), which is similar to learning strategies in [49, 109]. When comparing Variant #11 with Variant #8 and #9, we observe that Variant #11 performs better on ARO and VL-CheckList, but worse on SugarCrepe and SugarCrepe++. We hypothesize that by focusing exclusively on the positive and hard negative sets, the model may learn nonsensical artifacts present in the hard negatives. These artifacts improve performance on ARO and VL-CheckList, where the test samples exhibit similar patterns, but hinder performance on SugarCrepe and SugarCrepe++, which are designed to avoid such patterns [32]. In contrast, Variant #9, which incorporates all sets in Eq. (3) for margin loss calculation, achieves better results on SugarCrepe and SugarCrepe++ with only a minor drop in performance on ARO and VL-CheckList. This suggests that easy negative samples serve as regularization, preventing overfitting to the artifacts in hard negative samples.

Effects of synthetic positive samples. We compare the performance of training with only synthetic negative samples versus using both synthetic negative and positive samples

in Table 3. Based on Variant #7, we train ViT-B/32 with these two combinations of training data. We observe nearly 3% decrease in average accuracy when using only synthetic negative samples, compared to using both synthetic negative and positive samples. This decrease is particularly noticeable on SugarCrepe++, ARO, and VL-CheckList. These results underscore the importance of incorporating synthetic positive samples in training, as they provide essential information about variations that maintain semantic consistency, which is critical for compositional understanding.

Influence of model size and training data size. To evaluate the impact of model size on compositional understanding, we compare the performance of ViT-B/32 and ViT-L/14 in Table 2. We find that ViT-L/14 only provides a modest improvement of about 0.5% of average accuracy over ViT-B/32. This suggests that increasing model size does not significantly enhance compositional understanding. To investigate the effect of training data size, we train ViT-B/32 using 20% and 50% of randomly sampled training data and report the results in Table 3. We observe that performance improves as the proportion of training data increases, indicating that a larger training set helps the model better capture compositional knowledge. However, the performance gain diminishes when the training data size increases from 50% to 100%. Perhaps generating or selecting high-quality data would be more effective than simplistic data scaling.

Additional experiments on hyperparameter analysis can be found in Appendix A3.

5. Conclusion

In this paper, we introduce SPARCL to enhance the compositional understanding capabilities of VLMs by generating and training with synthetic data. To tackle two key challenges in using synthetic data—namely, the difficulty of generating accurate variations and the inconsistency in cross-modal alignment quality—SPARCL integrates image feature injection into a T2I model to improve the quality of synthetic variations and introduces an adaptive margin loss to account for varying levels of cross-modal alignment for effectively learning nuanced distinctions. Experiments on four visual-language compositional understanding benchmarks demonstrate the effectiveness of SPARCL.

6. Acknowledgments

This work has been supported by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF- NRFF13-2021-0006), Singapore. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

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 Sasstry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcreeper++ 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 mini-batch 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, Jinhung 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, Jianjun 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. Tripleclip: 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 Kaffle, 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, Han-shu 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