

Improving Vision and Language Concepts Understanding with Multimodal Counterfactual Samples

Chengen Lai, Shengli Song*, Sitong Yan, and Guangneng Hu

School of Computer Science and Technology, Xidian University, Xi'an, China
{laice, styan}@stu.xidian.edu.cn, shlsong@xidian.edu.cn, njuhgn@gmail.com

Abstract. Vision and Language (VL) models have achieved remarkable performance in a variety of multimodal learning tasks. The success of these models is attributed to learning a joint and aligned representation space of visual and text. However, recent popular VL models still struggle with concepts understanding beyond bag-of-objects in images & texts, suffering from compositional reasoning about relationship between objects & attributes and word order. To address the above issues, we create a synthetic multimodal counterfactual dataset (COCO-CF) and propose a novel contrastive learning framework (COMO). We contribute the COCO-CF dataset which is automatically generated from MS-COCO by injecting concepts from off-the-shelf language models and diffusion models to reduce the bias of bag-of-objects. We contribute the COMO framework for effectively leveraging COCO-CF to treat the counterfactual samples as hard negatives and reweight their importance during contrastive learning. Extensive experiments and ablations show COMO achieved a significant improvement of VL concept understanding on the two VL-Checklist and Winoground benchmarks over five strong VL baselines in their zero-shot setting evaluations. Dataset is available at <https://github.com/laichengen/COMO>.

Keywords: Concepts Understanding · Contrastive learning · Multimodal Counterfactual Samples

1 Introduction

Vision and Language (VL) models have recently achieved impressive performance in various challenging multimodal learning tasks, including image-text retrieval [6, 45, 58], visual question answering [1, 2, 55], and captioning [22, 53], thanks to the availability of large paired image-text corpora [25, 37]. These cross-modal tasks are heavily dependent on joint representations of multi-modalities which are typically learned by building interactions between vision and language features [10, 32, 33, 59]. The goal is to learn an effective aligning representation spaces of images and text, typically under a contrastive learning framework [4, 5, 52].

* Corresponding author

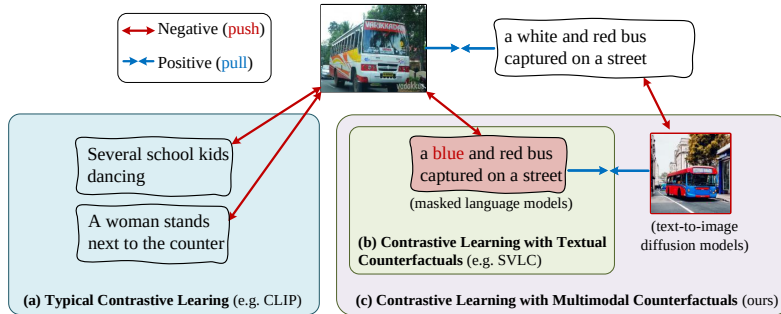


Fig. 1: Overview of our proposed multimodal counterfactual generation. (a) Standard contrastive text-to-image loss (e.g. CLIP [33]) is easy to solve “bag-of-objects”; (b) Contrastive learning with textual counterfactuals (e.g. SVLC [11])) teaches VL models by contrasting between counterfactual texts and original image. (c) Our proposed contrastive learning with multimodal counterfactuals (COMO) generate both textual and visual counterfactuals to improve the VL models’ concept understanding by contrasting between counterfactual image (and text) and original text (and image), and matching the counterfactual image with counterfactual text.

Despite these great advances on various vision and language tasks, they often exploit spurious correlations in datasets as shortcuts during training to fit the dataset [11, 12, 40]. Recent works have found popular VL models have difficulty in understanding structured concepts beyond bag-of-objects such as object attributes, inter-object relations, and word order in the sentence [27, 28, 49]. As a result, these models are vulnerable to text-domain or image-domain adversarial attacks: the model can be easily deceived by counterfactual captions constructed from original captions by adding small perturbations [41, 54, 57]. For example, large VL models such as CLIP [33] and CyCLIP [13] get confused by these counterfactual captions and fail to distinguish the difference between factual and counterfactual concepts, suffering from compositional reasoning beyond bag-of-objects (see Figure 1(a)).

In this paper, we propose a novel **CO**ntrastive learning framework with **M**ultimodal **cO**unterfactuals (COMO) to improve the VL model’s concepts understanding and compositional reasoning. We propose a way of leveraging the existing VL pre-training source to improve concepts understanding without any expensive and time-consuming human annotations on object attributes, relations, and states. The idea is to automatically generate a synthetic dataset by injecting counterfactual concepts from powerful pre-trained language and text-to-image stable diffusion [35]. On top of the recently proposed text augmentation [3, 11, 44] (see Figure 1(b)), we go deeper to study the effects of multimodal augmentation—i.e., generating both text counterfactual and image counterfactual examples—on the compositional reasoning performance of VL models trained on such synthetic multimodal counterfactual dataset (see Figure 1(c)).

⁰ <https://huggingface.co/CompVis/stable-diffusion-v1-4>

We empirically find that both text counterfactual and image counterfactual contribute to the improved concepts understanding, demonstrating the necessity of multimodal augmentation.

Our multimodal counterfactual generation not only substitutes the concept in the text by exploiting pretrained masked language models, but also synthesizes the corresponding image by exploiting text-to-image stable diffusion. Compared to textual counterfactual generation, our multimodal counterfactual samples match the counterfactual text and image together, to learn out-of-domain concepts and alleviate the data scarcity in multimodal learning task. The generated multimodal counterfactual dataset improves the quality of image-caption alignment and reduces the bias of bag-of-objects. To effectively enforce the VL model to distinguish between original concepts and substituted concepts, we treat the counterfactual samples as hard negatives by modifying the traditional contrastive loss to dynamically reweigh their importance. This new contrastive loss pushes the concept representation of counterfactual samples (which are very similar to the factual samples) far away from their corresponding factual samples).

Our main contributions can be summarized as:

- We contribute a multimodal counterfactual dataset (COCO-CF) which is automatically generated from MS-COCO by injecting concepts from off-the-shelf language models and stable diffusion to improve the VL model’s concept understanding beyond bag-of-objects and compositional reasoning.
- We propose a novel contrastive framework (COMO) for effectively leveraging COCO-CF to treat the multimodal counterfactual samples as hard negatives and reweight their importance during contrastive learning to enforce the VL model differentiating the original concepts from substituted ones.
- We demonstrate the compositional reasoning performance of COMO on the two VL-Checklist and Winoground benchmarks with 3.17% (on VL-Checklist’s attribute) and 4.35% (on Winoground’s image) improvement over the current SOTA models, respectively. We also perform detailed ablation on the importance of counterfactual text, counterfactual image, and the new contrastive loss.

2 Related Work

Vision and Language Models Pretrained VL models [16, 23, 32] have made impressive performance in various zero-shot downstream tasks, such as image-text retrieval [58]. However, recent research [3, 11, 57] show that existing VL models often exploit spurious correlations between non-causal features as shortcuts during training to fit the dataset [12, 54] and still struggle with vision and language concepts understanding. Popular VL models have difficulty in understanding structured concepts beyond bag-of-objects such as object attributes, inter-object relations, and word order in the sentence [27, 28, 49]. As a result, these models are vulnerable to text-domain or image-domain adversarial attacks:

the model can be easily deceived by counterfactual captions constructed from original captions by adding small perturbations [41, 54]. For example, large VL models such as CLIP [33] and CyCLIP [13] get confused by these counterfactual captions and fail to distinguish the difference between factual and counterfactual concepts, suffering from compositional reasoning beyond bag-of-objects. In this paper, we propose a data-driven technique to improve concepts understanding for typical VL models.

Counterfactual Samples Generation Counterfactual samples are widely used in various computer vision and natural language processing tasks [46, 60]. Counterfactual texts have been shown to improve the robustness of models through random word substitution [36, 47] and swap [54, 56]. Counterfactual captions are generated by rule-based [11, 36, 39] or pretrained language models [11, 44, 51] to enhance semantic coherence and ground the mismatched words [50]. Diffusion models [15] are exploited to generate text-image pairs [38, 43, 48]. Synthetic videos and corresponding text captions are generated by using 3D graphic engines [3]. A counterfactual dataset created by substituting nouns only, ignoring the “beyond nouns” concepts [20]. In contrast, we create a multimodal counterfactual dataset (COCO-CF) which is automatically generated from existing VL pre-training source (i.e., MS-COCO) by injecting concepts from off-the-shelf language models and stable diffusion. COCO-CF can be used to improve the VL model’s concept understanding beyond bag-of-objects and compositional reasoning without any expensive and time-consuming human annotations on object attributes, relations, and states.

Hard-Negative Contrastive Learning Contrastive learning (CL) [14] is exploited to learn aligned representations of text and image in most VL models [9, 18, 42]. Recently, some studies have investigated the selection of hard negative examples [17, 44] and accounted for the importance of different negative samples [7, 32, 34]. We extend the hard negative loss to vision-language concepts understanding by treating the counterfactual samples as hard negatives by modifying the traditional contrastive loss. This new contrastive loss pushes the concept representation of counterfactual samples (which are very similar to the factual samples) far away from their corresponding factual samples, so as to improve the compositional reasoning about object relation, attributes, state, and word order in texts.

3 Method

In this section, our **C**Ontrastive framework with **M**ultimodal **c**ounterfactual examples (COMO) are presented for improving the VL model’s understanding of concepts and enhancing robustness of VL models. As shown in Figure 2, COMO consists of an effective way of generating the multimodal counterfactual text and image samples (see Figure 2(a-b)), and a novel loss of counterfactual guided and weighted contrastive learning (see Figure 2(c)). In the following, we firstly describe the counterfactual text and image samples generation in Section 3.1.

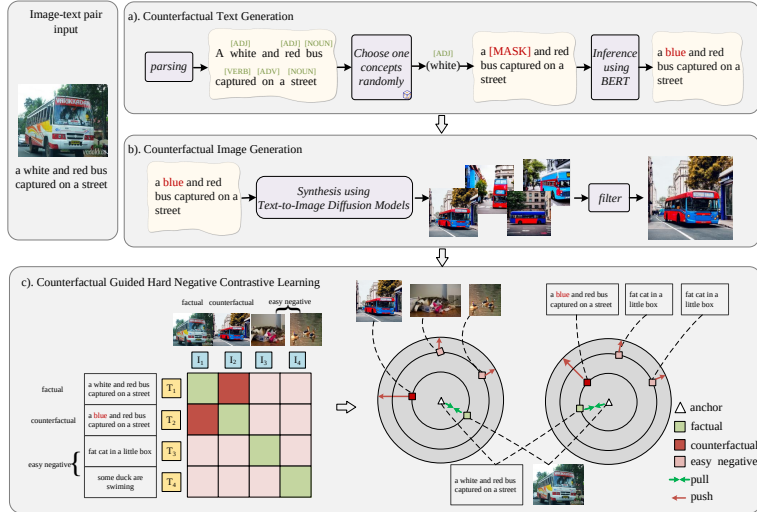


Fig. 2: The proposed contrastive learning framework with multimodal counterfactual samples (COMO). (a) Generating counterfactual text uses masked language models such as BERT¹. (b) Generating counterfactual image uses text-to-image diffusion models². (c) Counterfactual guided and weighted hard negative contrastive learning. This module feeds the generated counterfactual text and image into the same mini-batch and exploits a hard negative loss to improve the VL models’ concept understanding. The longer the red arrow, the further it pushes away.

Then, we introduce the counterfactual guided and weighted contrastive loss in Section 3.2, respectively.

3.1 Generation of Multimodal Counterfactual

Recent works [11, 49] have shown the effectiveness of injecting concepts to VL models by textual-modality augmentation. Compared to these works, we improve the VL models’ concepts understanding by automatically constructing both textual and visual counterfactual examples in a way of multimodal augmentation. **Counterfactual Text Generation** To improve the clear understanding of VL models on concepts beyond bag-of-objects and rather than spurious correlated features, we manipulate the original caption T^o to construct its corresponding counterfactual caption T^c with subtle but completely different semantics. Counterfactual text T^c is generated by replacing a concept in T^o , while most of its original details remain the same. In this way, the generated counterfactual caption represents the changed causal concept while remaining preserved

¹ <https://huggingface.co/bert-base-uncased>

details from the original caption can be viewed as potentially spurious correlated features.

Large or pre-trained language models [8] are capable of suggesting multiple words that fit the context given a sentence with one missing word. We can automatically create a plausible negative sentence which is difficult for VL models to distinguish from its corresponding actual caption by randomly selecting a masked word. Such masked word should be useful for compositional reasoning about objects relation, attributes, and states. This can be achieved by NLP parsing technique to identify all objects (nouns), relations (verbs and adverb), and attributes (adjectives) in the original caption sentence. In specific, we randomly choose a word among these concepts, replace the selected concept in T^o with the [MASK] token, and then retrieve the most probable replacements by the language models unmasking as our generated counterfactual caption T^c (see Figure 2(a)). These counterfactual/negative examples enforce the VL model focusing on the important details that affect the concepts understanding and compositional reasoning.

Counterfactual Image Generation After creating a counterfactual caption T^c , the VL models can understand the original text-image pair (T^o, V^o) but have no idea of what the corresponding image of counterfactual caption T^c looks like. Therefore, we explicitly generate a corresponding counterfactual image V^c by feeding the counterfactual caption T^c into stable diffusion models (see Figure 2(b)). To control the quality of generated images and enable the changed concept (causal concept) to be learned in the presence of unchanged concepts (spurious concepts) between T^c and T^o , we firstly generate multiple images $\{V_i\}_{i=1}^n$. Then we choose the counterfactual image V^c which has the highest similarity score with both the counterfactual caption T^c and the original image V^o by:

$$score_i = \varepsilon_T(T^c) \cdot \varepsilon_V(V_i) + \varepsilon_V(V^o) \cdot \varepsilon_V(V_i) \quad (1)$$

where $\varepsilon_T(\cdot)$ and $\varepsilon_V(\cdot)$ are text and image encoders in CLIP³, respectively. The first part of the score measures the consistency between the generated image and counterfactual caption T^c , and the second part measures the similarity between the generated image and the original image V^o .

3.2 Counterfactual Guided and Weighted Contrastive Learning

Contrastive learning [29] is an effective approach for multimodal alignment, which pulls the positive pairs to anchor nearby locations, while pushes negative pairs further away. A dual-encoder VL model like CLIP [33] admits text-image pair (T, V) and computes their similarity score by:

$$S(T, V) = \exp(\mathbf{t}^\top \mathbf{v} / \tau) \quad (2)$$

where $\mathbf{t} = \varepsilon_T(T)$ and $\mathbf{v} = \varepsilon_V(V)$ are the encoded features of the image-text pair respectively. Temperature $\tau > 0$ is learnable.

³ <https://openaipublic.azureedge.net/clip/ViT-B-32>

Counterfactual Guided Contrastive Learning The naive contrastive learning with uniform sampling from large-scale datasets can often provide negative samples that are not necessarily discriminative. In our counterfactual guided contrastive learning, we feed the original example with its corresponding counterfactual example into the same mini-batch so as to enforce the VL model differentiates the original concept from substituted concept (see Figure 2(c)).

Given a mini-batch $\mathcal{B} = \{(T_i, V_i)\}_{i=1}^n$ of containing the factual and its counterfactual image-caption pairs, the contrastive CLIP-loss [29] is defined as:

$$\mathcal{L}_{\text{HN}} = - \sum_{i=1}^n \left[\log \frac{S(T_i, V_i)}{\sum_j S(T_i, V_j)} + \log \frac{S(T_i, V_i)}{\sum_j S(T_j, V_i)} \right] \quad (3)$$

The hard negative for factual image-caption (T^o, V^o) (causal concept) is its corresponding counterfactual image-caption (T^c, V^c) (spurious concept), which needs to be pushed far away (see Figure 2(c)).

Counterfactual Weighted Contrastive Learning To emphasize hard negative pairs (i.e., between the factual example and its corresponding counterfactual example) and push the embedding of counterfactual sample (which is close to the factual example) far away from the anchor, we modify the un-weighted contrastive loss in Eq.3 as the weighted version in the following:

$$\begin{aligned} \mathcal{L}_{\text{HNW}} = - \sum_{i=1}^n & \left[\log \frac{S(T_i, V_i)}{S(T_i, V_i) + \sum_{j \neq i} \alpha_{ij} \cdot S(T_i, V_j)} \right. \\ & \left. + \log \frac{S(T_i, V_i)}{S(T_i, V_i) + \sum_{j \neq i} \beta_{ji} \cdot S(T_j, V_i)} \right] \end{aligned} \quad (4)$$

where the weights are computed by:

$$\alpha_{ij} = \frac{(n-1) \cdot S(T_i, V_j)}{\sum_{k \neq i} S(T_i, V_k)} \quad \beta_{ji} = \frac{(n-1) \cdot S(T_j, V_i)}{\sum_{k \neq i} S(T_k, V_i)} \quad (5)$$

Weights α_{ij} and β_{ji} are designed that difficult negative pairs (e.g., between original examples and corresponding counterfactual examples) are emphasized, and easier pairs (e.g., between original examples and other “bag-of-objects” examples) are ignored. So, the VL model can more easily understand the concepts beyond objects and improve the compositional reasoning about relation, attribute, and word order. Observe that we get contrastive objective of Eq.3 when setting weights to be all ones. The form of weights is an unnormalized von Mises-Fisher distribution [32].

4 Experiments

4.1 Datasets

To test the effectiveness of our proposed COMO for improving VL models’ vision language concepts understanding, we train the model using our counterfactual image-text pairs generated from MS-COCO [25]. We evaluate on the

Table 1: The statistics of factual and counterfactual examples in the generated COCO-CF dataset.

	Factual		Counterfactual		Total	
	# images	# text	# images	# text	# images	# text
COCO-CF	113K	567K	567K	567K	680K	1,134K

VL-Checklist [57], Winoground [41] and Zero-Shot Classification [21] datasets over strong VL models. We introduce the training and testing datasets in detail below.

Training datasets. We start from MS-COCO [25] dataset as our factual samples, where an image has multiple corresponding captions. We automatically generate the counterfactual image-text pairs from COCO with our multimodal counterfactual generation pipeline (see Figure 2(a-b)). The augmented dataset is named as COCO-CF, meaning COCO counterfactual examples. As shown in Table 1, we obtain additional 567K counterfactual images and captions based on the original 113K images and 567K text. Finally, COCO-CF has total 680K images and 1,134K captions.

Evaluation datasets. We evaluate our models on VL-Checklist, Winoground and Zero-Shot classification datasets.

VL-Checklist [57] is a large-scale dataset comprised of Visual Genome [19], SwiG [31], VAW [30], and HAKE [24]. Each image is associated with two captions, a positive and a negative. The positive caption corresponds to the image and is taken from the source dataset. The negative caption is generated from the positive caption by changing one word only, so the resulting sentence no longer corresponds to the image. Depending on the word that was changed, VL-Checklist evaluates seven types of VL concepts divided into three categories: (i) Attributes (color, material, size, state, and action); (ii) Relations (spatial or action relation between two objects and/or humans); and (iii) Objects (spatial location and size).

Winoground [41] is a dataset that evaluates the ability of VL models for compositional reasoning, specifically understanding the meaning of the sentence after changing the order of its words. An example comprises of two images and two texts. The texts have the same set of words but in a different order, each text corresponding to one image in the paired sample. The Winoground evaluation divide into three metrics: (i) image score - percent of samples where the model picks the correct image for each text; (ii) text score - percent of samples where the model picks the correct text for each image; (iii) group score - percent of samples where both text and image score conditions are satisfied jointly.

Zero-Shot Classification [21] is a dataset that includes 21 different classification datasets, we evaluate our model and report the average result over the dataset.

Table 2: Performance comparison with strong VL baselines on the VL-Checklist, Winoground and 21 classification task datasets in zero-shot setting evaluation.

Models	Fine-tuned Dataset Size	VL-Checklist			Winoground			21-Zero-Shot Task
		Relation	Attribute	Object	Text	Image	Group	
CLIP (Radford et al, 2021)	None	61.80	67.32	82.91	30.50	11.00	8.75	56.37
CyCLIP (Goel et al, 2022)	None	61.15	66.96	80.87	30.50	10.75	9.00	55.99
NegCLIP (Yuksekgonul et al, 2023)	None	63.52	72.23	81.35	29.50	10.50	8.00	-
SyVic (Cascante-Bonilla et al, 2023)	767K	69.39	70.37	84.62	30.00	11.50	9.50	54.77
SVLC (Doveh et al, 2023)	3M	69.68	71.18	84.75	29.75	11.00	9.00	55.27
COMO (ours)	1.1M	71.16	73.44	86.20	31.00	12.00	9.75	55.87
Improvement over best baseline	-	2.12%	3.17%	1.70%	1.60%	4.35%	2.63%	-0.89%

4.2 Implementation details

For the multimodal counterfactual samples generation (Sec. 3.1), we use NLP spacy⁴ to parse the sentence into its components including nouns, verbs, adjectives, adverbs, etc. Then we utilize the popular BERT⁵ as the masked LMs to generate the counterfactual text. To ensure the words sampled from the LM are true negatives, we obtain the synonyms set of the original word from the WordNet [26], and generate n candidates from the language model. Then we filter out the candidates that belong in the original word’s synonyms set and variants (i.e., plural and tense). Finally, we choose the one which has the highest score among the remaining candidates as our counterfactual text. After we generate the counterfactual captions, we use the text-to-image stable diffusion model⁶ to synthesize multiple images (size set to 10). For the model architecture, following [3, 11], we utilize the original CLIP implementation that initialized with the checkpoints of ViT-B/32⁷ released by the OpenAI. We modify the naive contrastive loss in their codebase to our counterfactual guided-and-weighted hard negative contrastive loss (see Eq.4). For the model training, we use Adam optimizer with a 1e-6 initial learning rate and a 1e-7 weight decay for finetuning, where the batch size is set to 16 and the training epoches to 5. We conduct all experiments on one NVIDIA GTX A100-PCIE-40GB GPUs with PyTorch 1.9.0. Our code and model checkpoints will be released upon acceptance together with the generated counterfactual examples COCO-CF dataset.

4.3 Main Results

We compare our COMO method with five state-of-the-art methods on three benchmark datasets and the results are shown in Table 2, where the best results are in boldface. We have following observations.

⁴ https://github.com/explosion/spacy-models/en_core_web_trf-3.7.2

⁵ <https://huggingface.co/bert-base-uncased>

⁶ <https://huggingface.co/CompVis/stable-diffusion-v1-4>

⁷ <https://openaipublic.azureedge.net/clip/ViT-B-32>

Table 3: Ablated performance of COMO trained on combinations of factuels and counterfactuals. For training on full factuels and counterfactuals (i.e. COCO-CF), COMO has two learning strategies, Random and In-Batch. The In-Batch strategy enforces the factual and its corresponding counterfactual feeded into the same mini-batch (see Figure 2(c)), while the Random strategy does not.

Fine-tuned Datasets	VL-Checklist				Winoground			
	Relation	Attribute	Object	Average	Text	Image	Group	Average
None (pre-trained CLIP)	61.80	67.32	82.91	70.67	30.50	11.00	8.75	16.75
Factuals Only (i.e. COCO)	62.12	67.43	82.98	70.84	30.25	10.50	8.75	16.50
Factuals + Counterfactuals (Random)	61.95	67.79	83.12	70.95	30.25	11.00	9.00	16.75
Factuals + Counterfactuals (In-Batch)	71.16	73.44	86.20	76.93	31.00	12.00	9.75	17.58

We observe that our COMO achieves the new state-of-the-art performance on both concepts understanding datasets. Specifically, our COMO outperforms the best baseline with 3.17% relative improvement in terms of VL-Checklist’s attribute metric, and with 4.35% relative improvement in terms of image score on Winoground. This shows that our COMO can better understand the concepts beyond bag-of-objects and conduct compositional reasoning about relation, attribute, and word order.

Furthermore, our COMO gets the best result over all six settings on both datasets. It shows our model can distinguish the differences between causal and spurious concepts. The improvement of COMO over baselines could be attributed to reasons below: i) COMO admits multimodal counterfactual examples by injecting concepts from pretrained masked language models and text-to-image diffusion model, which explicitly improve the clear understanding of VL models on semantic concepts; ii) COMO adopts a modified contrastive loss to reweight the importance of counterfactual samples in the mini-batch, which enforces the model to differentiate the original concepts in factuels from substitute concepts in counterfactuals.

We display relative gains on fine-grained “Attribute” and “Relation” tests of VL-Checklist as shown in Figure 3a. It is clear that COMO gets gains across over all tests (Attribute: size, material, color, state, action; Relation: spatial, action). These improvements can be contributed to the multimodal counterfactual samples that enforce the model to attend to the small concepts changes in the vision and text, and hard negative contrastive loss that push the embedding of counterfactual samples close to the anchor to be far away by reweighting the importance of counterfactual examples.

4.4 Ablation Studies

Ablation on Mixing Factuals and Counterfactuals We have demonstrated that augmenting factual datasets (e.g. MS COCO) could improve the concept understanding and compositional reasoning of VL models with the help of multimodal counterfactual examples. We now conduct study on ablated contribution from factual, counterfactual, and their learning strategies. As shown in Table 3,

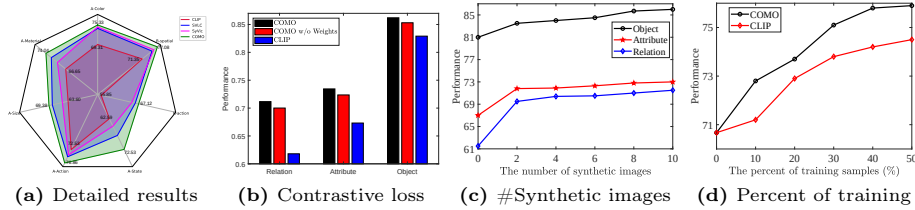


Fig. 3: (a) Detailed results of baselines (CLIP, SVLC, SyVic) and our COMO on VL-Checklist. A: Attribute, R: Relation. (b) Impact of guided-and-weighted contrastive loss; (c) Influence of the number of generated images; (d) Performance varying with the training percent. (all evaluated on VL-Checklist)

Table 4: Ablated contribution from counterfactual images and counterfactual texts in COMO.

Counterfactual		VL-Checklist				Winoground
Image	Text	Relation	Attribute	Object	Average	Average
	CLIP	61.80	67.32	82.91	70.67	16.75
✗	✗	62.12	67.43	82.98	70.84	16.50
✗	✓	69.56	71.11	84.55	75.07	16.13
✓	✗	69.34	70.85	84.22	74.80	16.58
✓	✓	71.16	73.44	86.20	76.93	17.58

we have the following findings. Firstly, compared to the pretrained CLIP, finetuning COMO on the factual only dataset (i.e., MS-COCO) has a negligible impact on Winoground dataset. It shows that the COMO that finetuning on only factual dataset still has the difficulty in concepts understanding beyond “bag-of-objects”. Secondly, there is almost no improvement of performance when finetuning the COMO on the factual and counterfactual dataset with a random sampling strategies. It shows the necessity of enforcing the contrastive learning between counterfactuals and the factuais explicitly to reduce the VL model from the bias that taking spurious correlations between captions and images as shortcuts during training. Thirdly, compared to the random sampling strategy, sampling the factual and counterfactuals together (i.e. in a mini-batch) could improve VL models’ robustness and the ability of concepts understanding and compositional reasoning.

Ablation on Multimodal Counterfactuals Text and Image We have ablated the contributions from multimodal counterfactuals, and now we investigate the impact of individual counterfactual text and individual counterfactual image. As shown in Table 4, removing either of them degrades the performance while removing both of them degrades a lot, demonstrating the usefulness of both counterfactual texts and counterfactual images to improve VL models’ concepts understanding. On the one hand, for the effectiveness of counterfactual text, COMO trained with counterfactual text performs much better than that trained with factual samples only. For examples, the relative improvement is

Table 5: Ablated Guided-and-Weighted Contrastive Loss on training from scratch and pretrained setting.

Pretrained	Weighted loss	VL-Checklist			
		Relation	Attribute	Object	Average
✓	✓	71.16	73.44	86.20	76.93
✗	✓	64.42	64.75	75.04	68.07
✓	✗	70.02	72.86	85.37	76.08
✗	✗	62.13	62.48	73.81	66.14

11.9% in terms of VL-Checklist’s relation metric. It shows that it is necessary to force the VL model to attend to the small changes in the text, which could improve the VL model’s understanding on concepts. On the other hand, for the effectiveness of counterfactual image, COMO trained with counterfactual image performs much better than that trained with factual samples only. For example, the relative improvement is 11.6% in terms of VL-Checklist’s relation metric. As for the relative importance of counterfactual texts and counterfactual images, COMO trained with counterfactual texts is better than that trained with counterfactual images when testing on the VL-Checklist, while it is on the contrary when testing on Winoground.

Ablation on Guided-and-Weighted Contrastive Loss As shown in Table 5, on pretrained setting, the performance of COMO w/o the weighted loss consistently degrades on all of VL-Checklist metrics (relation, attribute, object). Although the performance of COMO in training from scratch drops by a large margin, to explore the effectiveness of weighted loss in training from scratch, we compare the COMO with naive loss and weighted loss in training from scratch settings. The table shows that our weighted loss still performs better than the naive contrastive loss in training from scratch settings, which shows the design of weighted loss is robust. These ablation results show that the full guided-and-weighted contrastive loss is beneficial compared to the standard (i.e. CLIP-loss) and unweighted (i.e. Eq.3) ones.

4.5 Hyper-parameter Analysis

Influence of the number of generated images To improve the quality of generated images by text-to-diffusion models, and to reduce the inconsistencies in generated images and texts, we synthesize multiple images for a single caption at different time-step. In Figure 3c, we show the impact of the number of generated images on the three VL-Checklist evaluation settings (i.e. relation, attribute, object). We can see that, the performance increases as the number of generate images increases on all three evaluation settings, until getting saturated at around 8. This shows that filtering among the multiple generated images as the most satisfactory counterfactual image can improve the overall quality of our generated counterfactuals.

Low-resource scenario We conduct experiments in low-resource setting by randomly sampling part of the full training set to simulate a low-resource sce-

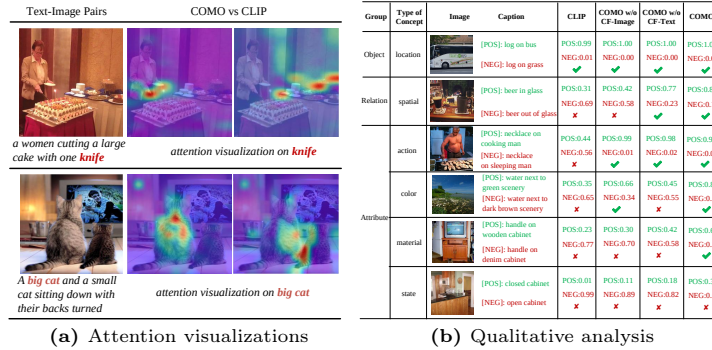


Fig. 4: (a) Attention visualizations on COMO and CLIP. (b) Qualitative analysis on COMO and its three variants. Note, CLIP i.e. COMO w/o both of counterfactual images and texts.

nario. The results are shown in Figure 3d where the CLIP and COMO are both trained on the corresponding low-resource training set. We can see that both CLIP and COMO get better performance with the increasing of more training examples. Moreover, our COMO achieves much advantage under the extreme low-resource scenario. In detail, relative improvements of COMO over CLIP are 2.24% at training percent 10%. It shows that our COMO is effective in improving the VL concepts’ understanding under data scarcity scenario.

4.6 Case study

Attention visualizations We conduct attention visualizations on COMO and CLIP to observe the ability of visual and language concept understanding. As shown in Figure 4a, our COMO can capture object “knife” and attribute “big” while CLIP wrongly attends to “plates” and “small” cat. It shows that our COMO improve the clear understanding of VL models on semantic concepts and learn the difference between VL concepts.

Qualitative analysis We conduct qualitative experiments to understand the working of COMO and its three variants, i) COMO without counterfactual images (COMO w/o CF-image), ii) COMO without counterfactual texts (COMO w/o CF-text), and iii) COMO without both of them (i.e. CLIP). The results are shown in Figure 4b. Firstly, for frequently occurring VL “bag-of-objects” concepts, such as bus and grass in Figure 4b (Object-location), all VL models including CLIP, two variants of COMO, and COMO predict the caption correctly. However, CLIP can not identify the difference between “cooking” and “sleeping” in Figure 4b (Attribute-action), while all other VL models with counterfactuals match the image to the ground-truth caption correctly. It shows that the counterfactuals can help the model to learn the differences between concepts.

Secondly, as for the importance of counterfactual texts (i.e. by observing the variant COMO w/o CF-text), as shown in Figure 4b (Attribute-color), removing

counterfactual texts leads to the COMO wrongly matching the image to the ground-truth caption. It is shown that the counterfactual text can help the model to learn the general concepts, especially when the factual description is coarse.

Thirdly, as for the importance of counterfactual images (i.e. by observing the variant COMO w/o CF-image), as shown in Figure 4b (Relation-spatial), removing counterfactual images leads to the COMO wrongly predict the caption when the image has lots of visual concepts unrelated to the text. It shows that VL models with counterfactual images can reduce the influence of spurious concepts in image.

Fourthly, as for the importance of multimodal counterfactuals, as shown in Figure 4b (Attribute-material), removing multimodal counterfactuals leads the COMO can not identify the material of cabinet from the image. The improvement of the VL model’s concept understanding could be attributed to enforcing the model to attend to the small concept changes in the counterfactual image and text, and to match the counterfactual image and counterfactual text together to learn out of domain concepts.

We observe that, as shown in Figure 4b (Attribute-state), all VL models fail in matching the image to its ground-truth caption “closed cabinet”. The failing reason could be that an “open door” appears in the image, which guides the model to trust that “open” is matching to the image. Though multimodal counterfactual samples improve the model’s understanding of concepts, VL models still suffer from the problems of “bags of objects” in images and texts. It is worthy of investigating in improving concepts understanding and compositional reasoning of VL models.

5 Conclusions

We have presented a data-and-algorithm driven technique for enhancing the performance of VL models to understand beyond “bag-of-objects” and reason about compositional concepts including attributes, relations, and word order. Our proposed contrastive learning framework with multimodal counterfactuals attains significant gains over five strong VL models on two benchmark datasets. It builds upon the modeling strength and knowledge of nowadays language models and diffusion models by injecting concepts to VL models, suggesting generalizations to future VL models. We demonstrated the necessity of both counterfactual texts and counterfactual images. We also show the necessity of effective learning with guided-and-weighted contrastive loss. Extensive ablation and case study help understand the working of the proposed approach as well its limitations, suggesting the future work of investigating other factors on improving compositional reasoning.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62306220).

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Basu, A., Addepalli, S., Babu, R.V.: Rmlvqa: A margin loss approach for visual question answering with language biases. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11671–11680 (June 2023)
3. Cascante-Bonilla, P., Shehada, K., Smith, J.S., Doveh, S., Kim, D., Panda, R., Varol, G., Oliva, A., Ordonez, V., Feris, R., et al.: Going beyond nouns with vision & language models using synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20155–20165 (2023)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Chen, Y., Yuan, J., Tian, Y., Geng, S., Li, X., Zhou, D., Metaxas, D.N., Yang, H.: Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15095–15104 (2023)
6. Chen, Y., Ma, Z., Zhang, Z., Qi, Z., Yuan, C., Shan, Y., Li, B., Hu, W., Qie, X., Wu, J.: Vilem: Visual-language error modeling for image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11018–11027 (June 2023)
7. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. *Advances in neural information processing systems* **33**, 8765–8775 (2020)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186 (2019)
9. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems* **27** (2014)
10. Doveh, S., Arbelle, A., Harary, S., Alfassy, A., Herzig, R., Kim, D., Giryes, R., Feris, R., Panda, R., Ullman, S., et al.: Dense and aligned captions (dac) promote compositional reasoning in vl models. *NeurIPS* (2023)
11. Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., Feris, R., Panda, R., Ullman, S., Karlinsky, L.: Teaching structured vision & language concepts to vision & language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2657–2668 (2023)
12. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
13. Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., Grover, A.: Cyclic: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems* **35**, 6704–6719 (2022)
14. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06). vol. 2, pp. 1735–1742. IEEE (2006)

15. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., QI, X.: Is synthetic data from generative models ready for image recognition? In: The Eleventh International Conference on Learning Representations (2022)
16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
17. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems* **33**, 21798–21809 (2020)
18. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. *Advances in neural information processing systems* **33**, 18661–18673 (2020)
19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **123**, 32–73 (2017)
20. Le, T., Lal, V., Howard, P.: Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *arXiv preprint arXiv:2309.14356* (2023)
21. Li, C., Liu, H., Li, L., Zhang, P., Aneja, J., Yang, J., Jin, P., Hu, H., Liu, Z., Lee, Y.J., et al.: Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems* **35**, 9287–9301 (2022)
22. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML* (2023)
23. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International Conference on Machine Learning*. pp. 12888–12900. PMLR (2022)
24. Li, Y.L., Xu, L., Liu, X., Huang, X., Xu, Y., Chen, M., Ma, Z., Wang, S., Fang, H.S., Lu, C.: Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539* (2019)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. pp. 740–755. Springer (2014)
26. Miller, G.A.: Wordnet: a lexical database for english. *Communications of the ACM* **38**(11), 39–41 (1995)
27. Moltisanti, D., Keller, F., Bilen, H., Sevilla-Lara, L.: Learning action changes by measuring verb-adverb textual relationships. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 23110–23118 (June 2023)
28. Momeni, L., Caron, M., Nagrani, A., Zisserman, A., Schmid, C.: Verbs in action: Improving verb understanding in video-language models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15579–15591 (2023)
29. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
30. Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Learning to predict visual attributes in the wild. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 13018–13028 (2021)
31. Pratt, S., Yatskar, M., Weihs, L., Farhadi, A., Kembhavi, A.: Grounded situation recognition. In: *European Conference on Computer Vision*. pp. 314–332 (2020)

32. Radenovic, F., Dubey, A., Kadian, A., Mihaylov, T., Vandenhende, S., Patel, Y., Wen, Y., Ramanathan, V., Mahajan, D.: Filtering, distillation, and hard negatives for vision-language pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6967–6977 (2023)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
34. Robinson, J.D., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: *International Conference on Learning Representations* (2020)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695 (2022)
36. Roth, K., Kim, J.M., Koepke, A.S., Vinyals, O., Schmid, C., Akata, Z.: Waffling around for performance: Visual classification with random words and broad concepts. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 15746–15757 (October 2023)
37. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021)
38. Shi, C., Yang, S.: Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2932–2941 (2023)
39. Shi, H., Mao, J., Xiao, T., Jiang, Y., Sun, J.: Learning visually-grounded semantics from contrastive adversarial samples. In: *Proceedings of the 27th International Conference on Computational Linguistics*. pp. 3715–3727 (2018)
40. Smith, J.S., Cascante-Bonilla, P., Arbelles, A., Kim, D., Panda, R., Cox, D., Yang, D., Kira, Z., Feris, R., Karlinsky, L.: Construct-vl: Data-free continual structured vl concepts learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14994–15004 (2023)
41. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5238–5248 (2022)
42. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? *Advances in neural information processing systems* **33**, 6827–6839 (2020)
43. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. In: *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models* (2023)
44. Wang, W., Yang, Z., Xu, B., Li, J., Sun, Y.: Vilt: Enhancing vision-language pre-training through textual augmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3158–3169 (2023)
45. Wang, Z., Gao, Z., Guo, K., Yang, Y., Wang, X., Shen, H.T.: Multilateral semantic relations modeling for image text retrieval. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2830–2839 (June 2023)
46. Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In: *Proceedings of the 2019 Conference on Empirical*

- Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 6382–6388 (2019)
47. Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 6382–6388 (2019)
 48. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15943–15953 (October 2023)
 49. Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., Ma, W.Y.: Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6609–6618 (2019)
 50. Wu, Y., Wei, Y., Wang, H., Liu, Y., Yang, S., He, X.: Grounded image text matching with mismatched relation reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2976–2987 (2023)
 51. Xie, C.W., Sun, S., Xiong, X., Zheng, Y., Zhao, D., Zhou, J.: Ra-clip: Retrieval augmented contrastive language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19265–19274 (June 2023)
 52. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15671–15680 (2022)
 53. Yang, K., Deng, J., An, X., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Alip: Adaptive language-image pre-training with synthetic caption. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2922–2931 (October 2023)
 54. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: 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)
 55. Zhang, X., Zhang, F., Xu, C.: Vqacl: A novel visual question answering continual learning setting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19102–19112 (June 2023)
 56. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. *Advances in neural information processing systems* **28** (2015)
 57. Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., Yin, J.: Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221* (2022)
 58. Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10394–10403 (2019)
 59. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)
 60. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13001–13008 (2020)