RELATIONCLIP: TRAINING-FREE FINE-GRAINED VISUAL AND LANGUAGE CONCEPT MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Language-Image Pretraining (CLIP) has demonstrated great zero-shot performance for image-text matching, because of its holistic use of natural language supervision that covers large-scale, unconstrained real-world visual concepts. However, it is still challenging to adapt CLIP to fined-grained image-text matching between disentangled visual concepts and text semantics without training. Towards a more accurate zero-shot inference of CLIP-like models for fine-grained concept matching, in this paper, we study the image-text matching problem from a causal perspective: the erroneous semantics of individual entities are essentially confounders that cause the matching failure. Therefore, we propose a novel training-free framework, RelationCLIP, by disentangling input images into subjects, objects, and action entities. By exploiting fine-grained matching between visual components and word concepts from different entities, RelationCLIP can mitigate spurious correlations introduced by the pretrained CLIP models and dynamically assess the contribution of each entity when performing image and text matching. Experiments on SVO-Probes and our newly-introduced Visual Genome Concept datasets demonstrate the effectiveness of our plug-and-play method, which boosts the zero-shot inference ability of CLIP even without pre-training or fine-tuning. Our code is available at `https://anonymous.4open.science/r/Relation-CLIP`.

1 INTRODUCTION

Image and text matching (Plummer et al., 2015; Lin et al., 2014) is a fundamental task for vision and language research that involves multimodal reasoning and multi-level visual and text concept alignment. Recently, a growing number of pretrained vision and language foundation models (Radford et al., 2021; Jia et al., 2021; Li et al., 2022b) have shown encouraging results towards open-domain visual and language concept matching. Among them, CLIP (Radford et al., 2021) can be easily transferred to image and text matching under zero-shot and few-shot scenarios. However, CLIP treats the image and the text as a whole for alignment and ignores the fine-grained matching of disentangled concepts. For instance, Figure 1 shows some examples that CLIP fails at, which require accurate subject, verb, or object concept matching.

In fact, it is widely observed that current pretrained vision and language models struggle in recognizing actions from the input image, distinguishing objects from subjects (Hendricks & Nematzadeh, 2021), or failing to identify objects in unseen surroundings (Rosenfeld et al., 2018). They may be ascribed to shortcut learning (Geirhos et al., 2020) and dataset biases during pretraining, where the models learn the correspondence between entities and images implicitly and are thus prone to spurious correlations, incurring biases toward particular objects/subjects/predicates as well as their combinations.

Therefore, there are mainly two challenges to address when adopting CLIP for fine-grained visual and language concept matching. *Challenge 1*: the pretrained language model in CLIP is biased and tends to rely on spurious relationships learned during pretraining. For example, in Figure 1 (A), CLIP connects "frisbee" with "dog" (as they often appear together) and makes the wrong prediction. Meanwhile, the diversity of entity embeddings gives rise to *Challenge 2*: entity embeddings should contribute dynamically to fine-grained concept matching. Still taking Figure 1 (A) as an example, the object entity "woman" should be allocated with more attention from the model between those two images. Yet existing approaches often calculate the similarities merely based on the global embedding of images and texts and ignore fine-grained concept matching (Li et al., 2019).

Figure 1: Three challenging examples of the fine-grained image-text matching problem. CLIP fails to match the text prompts and the images correctly, while our RelationCLIP can deal with subject, predicate/verb, and object matching more effectively.

To address the aforementioned limitations, we propose a new training-free framework based on CLIP-like models, named RelationCLIP. We disentangle the visual scene into individual visual concepts and construct counterfactual sub-images containing subject/object/predicate entities only. Then we utilize backdoor adjustment (Pearl et al., 2000a) to implement interventions over the disentangled sub-images to mitigate the effect of spurious correlations. With this design from the causal perspective, RelationCLIP can bind the visual concepts with the correct text semantics and avoid matching solely based on spurious correlations. To validate our approach, we focus on the fine-grained image and text matching problem and evaluate it on two datasets: the SVO-Probes dataset (Hendricks & Nematzadeh, 2021), and the newly-introduced Visual Genome Concept dataset built upon Visual Genome Krishna et al. (2017). RelationCLIP gains an absolute accuracy improvement of 3.28% and 1.24% over CLIP using ViT-L-14 on Visual Genome Concept and SVO-Probes respectively.

Our primary contributions are summarized as follows:

- We propose a novel approach RelationCLIP to address the visual and language concept matching problem from the causal view: it disentangles the input image into counterfactual sub-images and leverages the idea of backdoor adjustment (Pearl et al., 2000a) to compose entity features and perform fine-grained concept matching, in order to mitigate the spurious correlations introduced during pretraining.
- The RelationCLIP framework is training-free and can be applied to CLIP-like models for zero-shot inference without pretraining or fine-tuning.
- We introduce a new dataset — Visual Genome Concept[1], containing 5400 image-text pairs with (**s**ubject, **v**erb, **o**bject) annotations, by generating image–sentence pairs from Visual Genome (Krishna et al., 2017) in the form of SVO-Probes (Hendricks & Nematzadeh, 2021) dataset, to benchmark fine-grained visual and language concept matching.
- We demonstrate the effectiveness of RelationCLIP on fine-grained concept matching and outperform CLIP on the SVO-Probes and Visual Genome Concept datasets.

2 Related Work

Image-Text Matching Most existing image-text matching datasets are evaluated in a classification setting. For example, Chao et al. (2015); Lu et al. (2016) focus on relationship or interaction detection. V-COCO (Gupta et al., 2020b) and ImSitu (Yatskar et al., 2016) include verbs in their data for evaluating the model's understanding ability. Gupta et al. (2020a); Faghri et al. (2017) explore how creating hard negatives (, by substituting words in train examples) leads to better test performance. FOIL benchmark (Shekhar et al., 2017) tests if vision-language models can differentiate between sentences that vary with respect to only one noun. SVO-Probes adds hard evaluation examples to test the model's understanding of verbs as well as subjects and objects in a controlled way. To associate local regions in an image with texts to do matching, Xu et al. (2015b) incorporates a soft form of attention into their recurrent model. Karpathy & Fei-Fei (2015) proposes an image-sentence ranking approach in which the score between an image and sentence is defined as the average over correspondence scores between each sentence fragment and the best corresponding image region; Ma et al. (2015) learns multiple networks that capture words, phrases, and sentence-level interactions

[1]The dataset is available at `https://drive.google.com/file/d/1rWHuq48paToXZs7_OT2Wko4l5YrAfFmR/view?usp=sharing`. We will release it publicly to facilitate future research.

with an image and combines the scores of these networks to obtain a whole image-sentence score. Hu et al. (2016) leverages spatial information and global context to predict where objects are likely to occur. Wang et al. (2016) formulates a linear program to localize all the phrases from a caption jointly, taking their semantic relationships into account. In this paper, we focus on the task of matching text prompts with images, which requires the model to distinguish error-prone words on a granular level — visual and language concept matching.

Pre-trained Vision and Language Models Vision and Language models pretrained on large-scale image-text pairs have demonstrated great potential in multimodal representation learning (Jia et al., 2021; Yao et al., 2021; Yuan et al., 2021). Among them, the representative model — CLIP (Radford et al., 2021) benefits from 400M curated data and defines various prompt templates to carry out zero-shot image classification. Most recent works seek to improve the zero-shot inference ability of CLIP via (Zhou et al., 2021; 2022; Ju et al., 2021; Song et al., 2022). However, these models can suffer from connecting verbs/subjects/objects concepts with visual components correctly Hendricks & Nematzadeh (2021) and bias towards spurious relations they have seen in the pretraining data, referred to as "confounders" (Zhang et al., 2020). By modeling using a structural causal model (SCM) network (Pearl et al., 2000b), the authors in Zhang et al. (2020) execute a hard intervention to eliminate dataset bias via a backdoor intervention during pretraining. Different from them, in this work, we focus on mitigating the effect of spurious relations and improving the zero-shot inference ability of off-the-shelf pretrained vision and language models, for visual and language concept matching. We develop a new training-free paradigm that gains superior performance on visual and language concept matching.

Disentangled Representation Learning It is often assumed that real-world observations like images can be disentangled Bengio et al. (2013); Peters et al. (2017). Li et al. (2020) disentangles background, texture, shape, etc, and uses object bounding boxes as supervision to synthesize images. Recent research in image synthesis seeks to learn disentangled features for managing the image generation process. Besserve et al. (2020) leverages the idea of independent mechanisms to identify modularity in pretrained generative models. Sauer & Geiger (2021) utilizes independent mechanisms to generate images to improve image classification. Ma et al. (2022) disentangles word entities from the conventional meanings of special entities encoded in the pretrained language model. Different from these works, we employ independent mechanisms to disentangle images and use generated sub-images to improve fine-grained visual and language concept matching.

3 RELATIONCLIP

In this section, we propose a simple yet effective approach incorporating a causal view into the CLIP-like models. We briefly introduce the background of RelationCLIP in view of structured causal models in Sec. 3.1. Then, we present the overview of RelationCLIP pipeline in Sec. 3.2. We introduce its critical components in detail in Sec. 3.3 and 3.4. Our dual objectives are: (i) We aim at disentangling visual input into sub-images containing fine-grained concepts. (ii) We intend to utilize those disentangled concepts to perform entity-level matching and mitigate the effect of spurious relations in the pretrained vision and language models learned during pretraining.

3.1 BACKGROUND

Consider a dataset comprised of (high-dimensional) observations X (i.e. images), and corresponding text prompts Y. Assume that each X can be described by lower-dimensional, semantically meaningful factors of variation z (e.g., objects, subjects, or action relations between objects and subjects (i.e, predicates in the image)). We consider the semantics of these special entities as confounders Z, which may affect either X or Y. If we can disentangle these factors, we are able to perform fine-grained concept matching. We argue that rather than directly computing the similarity only based on the global embedding of X and Y, the mapping should be decomposed into several functions. Each of these functions is autonomous, e.g., replacing the object in the images results in different object encoding while all subject encoding remains unchanged. These criteria align with the principals of structural causal models (SCMs) (Pearl et al., 2000b) and independent mechanisms (IMs), where an SCM is defined as a collection of n independent mechanisms (IMs) $f_i, i = 1, \ldots, n$. Inspired by

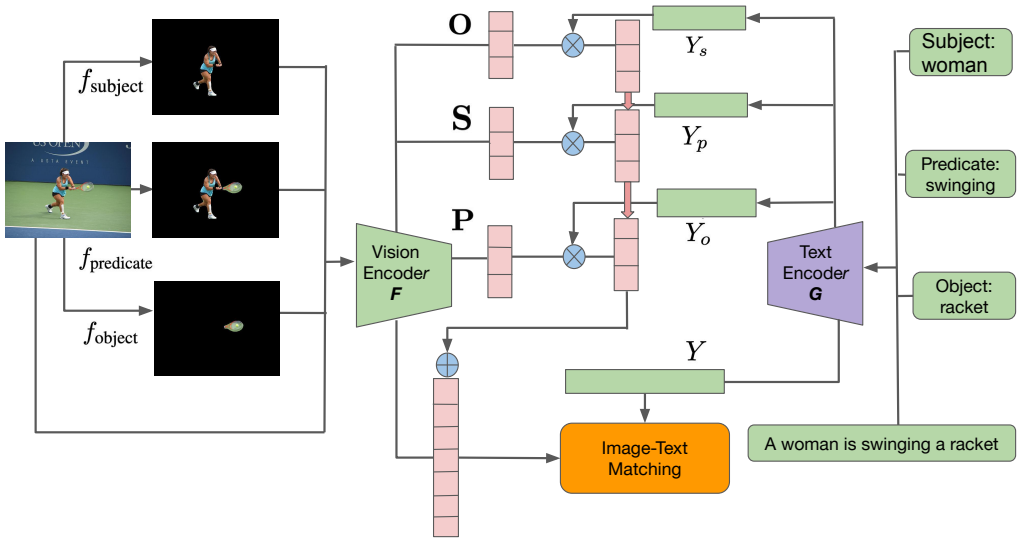


Figure 2: Overview of our RelationCLIP framework. We disentangle the input image using three independent encoding mechanisms by obeying the rules of encoding object, subject, and predicate respectively. The entity information is introduced to the global embedding of the whole image.

this, we can decompose the sub-image generation process into three independent mechanisms (IMs): object mechanism f_{object} , subject mechanism f_{subject}, and predicate mechanism $f_{\text{predicate}}$.

3.2 Method Overview

We first introduce the overview of our method from a conceptual view. An overview of our pipeline is shown in Figure 2. The goal is to steer the pretrained vision and language model to do fine-grained concept matching. Given an input image-prompt pair, we first disentangle the input image and generate sub-images only containing the given entity. Then we compute the pairwise similarity between the embedding of the sub-image with the entity embedding and adopt the similarity score to weight the sub-image embedding. The weighted embedding will be added to the global image embedding for final image-text matching, allowing the model to capture non-spurious semantic entity information and conduct concept matching at the granular level.

3.3 Counterfactual Sub-image Generation

We assume the causal structure to be known and consider three learned IMs (independent mechanisms) for generating object, subject, and predicate sub-images, respectively. An explicit formulation of the structural causal model (SCM) is:

$$\begin{aligned} \mathbf{O} &:= f_{\text{object}}\ (X, U_1) \\ \mathbf{S} &:= f_{\text{subject}}\ (X, U_2) \\ \mathbf{P} &:= f_{\text{predicate}}\ (X, U_3) \end{aligned} \tag{1}$$

where $\mathbf{O}$ is the object image, $\mathbf{S}$ is the subject image, $\mathbf{P}$ is the predicate image, $[U_1, U_2, U_3]$ are exogenous noises, X is the input image, and f_{object} , f_{subject}, $f_{\text{predicate}}$ are the independent mechanisms.

Specifically, given the input (subject, object, predicate) triplet, we model the object mechanism f_{object} using a binary mask generated by Lang-Seg (Li et al., 2022a), a CLIP-based language-guided segmentation model. The remainder of the image will be set to 0 while the object part is 1. In a manner similar to the object mechanism, the subject mechanism f_{subject} is achieved by setting the background to 0 while the subject region is set to 1. The predicate mechanism $f_{\text{predicate}}$ is implemented by combing the binary mask generated by f_{object} and f_{subject} together: the object and subject regions will be 1 while the remaining regions will be 0. Examples can be found in Figure 4.

Algorithm 1 Visual and Language Concept Matching with RelationCLIP.

Require:
Input: text prompt Y, image X, vision encoder $F(\cdot)$, text encoder $G(\cdot)$, independent mechanisms $f_{\text{object}}(\cdot), f_{\text{subject}}(\cdot), f_{\text{predicate}}(\cdot)$.
Output: Matching score O.

$\mathbf{O}, \mathbf{S}, \mathbf{P} \leftarrow f_{\text{object}}(X), f_{\text{subject}}(X), f_{\text{predicate}}(X)$; {Eq. 1}
Extract feature embeddings $F(\mathbf{O}), F(\mathbf{S}), F(\mathbf{B}) \leftarrow \mathbf{O}, \mathbf{S}, \mathbf{P}$;
Extract (subject, object, predicate) words $Y_s, Y_o, Y_p \leftarrow Y$;
$S_1, S_2, S_3 \leftarrow G(Y_s), G(Y_o), G(Y_p), F(\mathbf{O}), F(\mathbf{S}), F(\mathbf{P})$; {Eq. 5}
$G(Y) \leftarrow Y$;
$V \leftarrow S_1, S_2, S_3, f_{\text{object}}(\cdot), f_{\text{subject}}(\cdot), f_{\text{predicate}}(\cdot), F(\cdot), X$; {Eq. 6}
$O \leftarrow Y, V$ {Eq. 7}

With these IMs, giving the input image X, we can do counterfactual intervention by answering counterfactual questions such as "what if we only keep the subject/object/predicate in the original image?", and thereby we can generate *counterfactual images*, i.e., images which only contain the given entity (with examples shown in Figure 2). With the counterfactual sub-images generated, we can seek a way beyond its original image input to connect each disentangled entity concept with its corresponding text prompt. It is fair to anticipate appropriate matching results if each entity is encoded independently and connects correctly. The problem now boils down to how to develop a method to steer the composition process of different entity regions within an image.

3.4 Entity Composition

As mentioned earlier, the pretrained vision and language model is prone to be biased towards the specific subject, object or predicate, or even relied solely on one of them in the given sentence (Hendricks & Nematzadeh, 2021). From the causal perspective, given image X to match the correct Y, we want to infer $P(Y|X)$ while at the same time mitigating the effect of detrimental confounders z. The confounders may introduce spurious correlations in the model when directly inferring from $P(Y \mid X)$. Leveraging Bayes Rule,

$$P(Y \mid X) = \sum_z P(Y, z \mid X) = \sum_z P(Y \mid X, z) P(z \mid X), \tag{2}$$

where the confounder z introduces the bias of word concept via $P(z \mid X)$. To adjust the effect of confounder z, we can intervene X by first disentangling it and then intervening with it using do-operation [2]:

$$P(Y \mid do(X)) = \sum P(Y \mid X, z) P(z). \tag{3}$$

We now seek an implicit way to compute $P(Y \mid X, z)$ and $P(z)$. Considering the SCMs mentioned above, we interpret $f_{\text{object}}(X), f_{\text{subject}}(X), f_{\text{predicate}}(X)$ as incorporating the entity semantics into attended regions of the images.

To do concept matching over the prompt Y and the entity set $T^E = \left\{e^k\right\}_{k=1}^K$, where K is the total number of entities, and e^k is the k-th entity. This interpretation motivates us to compute similarity between $f_{\text{object}}(X), f_{\text{subject}}(X), f_{\text{predicate}}(X)$ with different word entity embeddings to achieve concept-wise semantic fusion and guidance. The prediction $P(Y \mid X, z)$ can be regarded as a classifier: $P(Y \mid X, z) = \text{Softmax } f_i(X, z)$. Similar to Wang et al. (2020), using the approximation of NGSM (Normalized Weighted Geometric Mean) (Xu et al., 2015a), we have:

$$P(Y \mid do(X)) \approx \text{Softmax}\left[\mathbb{E}_z\left(f_i(X, z)\right)\right]. \tag{4}$$

Specifically, to implement this on the SVO dataset, given a input image X and IMs $f_{\text{object}}(\cdot), f_{\text{subject}}(\cdot), f_{\text{predicate}}(\cdot)$, we first extract a collection of visual concepts from input images as $f_{\text{object}}(X), f_{\text{subject}}(X), f_{\text{predicate}}(X)$. For the language side, given a prompt Y and its entity set T^E,

[2] $P(Y \mid do(X)$ uses the do-operator Glymour et al. (2016). Given random variables X, Y, we write $P(Y = y \mid do(X = x))$ to indicate the probability that $Y = y$ when we intervene and set X to be x.

we extract all (subject, object, predicate) words (Y_s, Y_o, Y_p) from the input text prompts. Using cosine similarity score $\mathcal{S}$ as an example, we compute the similarity separately:

$$\begin{aligned} S_1 = \mathcal{S}(F(f_{\text{object}}(X)), G(Y_s)), S_2 = \mathcal{S}(F(f_{\text{subject}}(X)), G(Y_o)), \\ S_3 = S(F(f_{\text{predicate}}(X)), G(Y_p)), \text{ where } F(\cdot) = \text{CLIP}_{\text{vision}}(\cdot),\ G(\cdot) = \text{CLIP}_{\text{text}}(\cdot) \end{aligned} \tag{5}$$

The final visual feature is generated by:

$$V = F(X) + F(f_{\text{object}}(X))S_1 + F(f_{\text{subject}}(X))S_2 + F(f_{\text{predicate}}(X))S_3. \tag{6}$$

We can compute the image-text matching score by:

$$O = S(G(Y), V). \tag{7}$$

With this design, the language part of CLIP is aware of connections between entities from both the visual and language input when doing the concept matching.

Our algorithm can be summarized as 1, which requires no training or additional data. It is also simple enough to be adapted to any other vision and language pretrained model that implements the two-stream encoder structure.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

We evaluate RelationCLIP on SVO-Probes (Hendricks & Nematzadeh, 2021). We also collect a new dataset named Visual Genome Concept as a complementary testbed to SVO-Probes.

Visual Genome Concept Visual Genome Krishna et al. (2017) has around 2.3 million relationships annotated on its 108,007 images. These relationships contain both action relationships and spacial relationships, and are stored in a subject, predicate, and object triplet. With these relation triplets, we construct the Visual Genome Concept. We first create descriptions for all images by adding definite or indefinite articles that connect the relation triplets together. Since Visual Genome provides a rich number of images annotated with their relationships, we took this advantage and select a subset of it to form a dataset that can test a model's ability in differentiating detailed differences. We first pick out 542 images from Visual Genome that have clear relationships in them. Then inspired by SVO-Probes, for each key image, we iterate over the rest of the images and pick out its mutated images that only have one different value in either subject, object, or predicate. We can create multiple data points by treating the key image as the positive image, and each mutated image as a negative image. We manually created 5400 data samples like this for testing. The dataset is manually checked and grammatical mistakes are fixed.

SVO-Probes was designed to evaluate language-image models' capacity to distinguish fine-grained variations of the subject, object, or the relationship between subjects and objects in images. Each data sample contains a sentence, a positive image matching to the given sentence, and also a negative image that is different from the positive image in only the area among subjects, object, or actions. The model should match the sentence with the positive image. Originally, SVO-Probes has 30,000 data points. However, we can only evaluate our methods on 13,000 data points from it, because some images are no longer accessible, and some are failed to have effective and clear segmetation extracted from them.

Note that for both datasets, we used the same subject, object, and predicate from the only sentence to extract subject, object, and predicate images from both positive and negative images. No ground truth knowledge was used in the prediction. More examples from the two datasets can be found in Appendix A.1 and A.3. The dataset statistics are shown in Table 1. We evaluate our methods on the entire Visual Genome Concept. For SVO-Probes, we tested with 3 random splits and reported an average accuracy.

4.1.2 Implementation Details

Feature Fusion During inference, we pass the original image, subject image, object image, and predicate image into CLIP's visual encoder, and pass the original sentence, subject, object, and

	Subject-negative	Predicate-negative	Object-negative	Subjects	Predicates	Objects
Visual Genome Concept	2,584	1,536	1,280	30	65	82
SVO-Probes	5,679	23,525	7,637	100	421	275

Table 1: The number of data samples in the dataset that have one of their subject, object, or predicate changed between positive and negative images and the number of unique types of subjects, verbs, and objects across Visual Genome Concept and SVO-Probes.

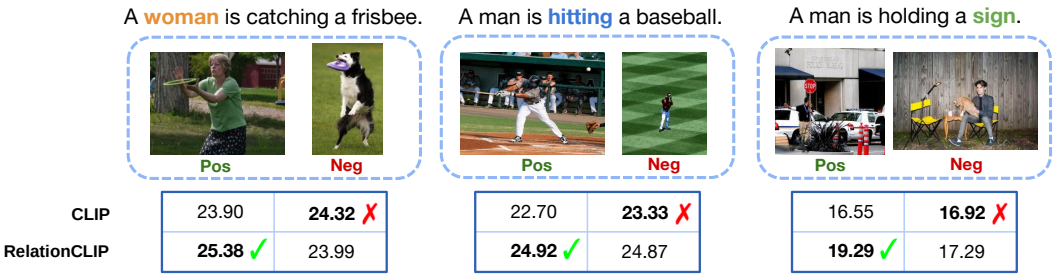


Figure 3: Examples showing the matching score between CLIP and RelationCLIP using ResNet-50 as the vision encoder. With entity embedding involved, RelationCLIP could match the text prompts with the correct image while CLIP makes the wrong prediction.

predicate entities into CLIP's text encoder. For each sub-image's embedding, we calculate a cosine similarity score with its corresponding word embedding. Three cosine similarity scores were fed into a softmax layer, yielding three positive weights. Finally, we would use the weights on the subject image embedding, object image embedding, and predicate image embedding and get a weighted sum of these three embeddings. We would add this embedding back to the original image's embedding. Then use it as a final embedding on the image side.

Evaluation Metrics We use accuracy as the evaluation metric, where we use the text input as the query and measure the accuracy of matching the correct images. In our experiments, both datasets follow the pattern of one text prompts — two images, and the model is actually selecting from the two images.

4.2 Main Results

In this subsection, we show the evaluation results on Visual Genome and SVO dataset in Table 2. Our RelationCLIP can outperform zero-shot CLIP on both Visual Genome Concept and SVO-Probes dataset. This indicates that incorporating the information of sub-images at inference time is helping CLIP grow attention to the details in images. From Table 2, our methods also work on different types of vision encoders. We noticed that with the relatively weak vision encoder ViT-L-14, our methods have the highest improvement compared to other vision encoders. CLIP with ViT-L-14 has low accuracy 74.35% on Visual Genome Concept, which means its ability in distinguishing fine-grained differences alone is limited. After employing our methods, its accuracy grows to 77.63%, even though the sub-images we added are still encoded by the same ViT-L-14. This shows that our methods are not simply stressing the major objects in the image by adding their representations to the original image embedding one more time. We are also guiding the image embedding process with the subject, object, and predicate words to make the image stand out from its negative counterpart.

In addition, we realize that our methods have lower performance improvement on SVO-Probes dataset compared to Visual Genome Datasets. We hypothesize this is due to two reasons: 1. SVO-Probes has a most portion of its data samples testing the fine-grained difference in predicates, while our methods are relatively weak in predicate differentiation. 2. SVO-Probes has a number of sketchy data samples that we can not remove completely. We present and analyze some bad examples from SVO-Probes in Appendix A.4.

	Visual Genome Concept		SVO-Probes	
Vision Encoder	CLIP	RelationCLIP	CLIP	RelationCLIP
ResNet-50	82.20	**83.91**	81.06	**82.05**
ViT-B-32	82.41	**84.33**	82.13	**83.00**
ViT-L-14	74.35	**77.63**	71.98	**73.22**

Table 2: Comparison of accuracy (%) on Visual Genome Concept and average accuracy (%) across the three splits on SVO-Probes using RelationCLIP and CLIP.

	Visual Genome Concept		SVO-Probes	
Vision Encoder	SLIP	Ours	SLIP	Ours
SLIP (ViT-B-32)	78.89	79.56	79.27	79.27
SLIP (ViT-L-14)	79.91	80.85	79.57	80.42

Table 3: Comparison of accuracy (%) on Visual Genome Concept and average accuracy (%) across the three splits on SVO-Probes using our method on SLIP (Mu et al., 2021) and original SLIP.

	Subject	Predicate	Object
CLIP	86.85	65.23	87.85
RelationCLIP	88.34	67.27	89.51

Table 4: Visual Genome Concept accuracy (%) on each negative type (ResNet50 as the vision encoder)

	CLIP	RelationCLIP
Provided SVO	81.07	82.05
Parsed SVO	81.07	82.03

Table 5: Comparison of accuracy (%) on Visual Genome Concept and SVO-Probes using parsed and ground-truth SVO triplets.

4.3 Ablations and Analysis

Performance on Other Pretrained Vision and Language Models Apart from CLIP, we also validate the effectiveness of our method on SLIP (Mu et al., 2021), with the results shown in Table 3. As can be seen, ours can beat SLIP using both ViT-B-32 and ViT-L-14, validating the effectiveness of our method on other CLIP-like models.

Visual Genome Concept Accuracy on Different Negative Types We categorize the results of the Visual Genome Concept into specific problem types (negation in subjects, objects, and predicates). Separately reviewing our results, we see an improvement in all negative types from Table 4. On negative predicate, RelationCLIP has the highest gain of 2.04% accuracy, suggesting our RelationCLIP can help to improve the verb/predicate understanding capability.

Use Language Parser to Extract SVO The performance of RelationCLIP is also dependent on the quality of the subject, object, and predicate entity provided. We analyze our methods on SVO-Probes since it have more complex sentence structures. We remove stop words from the sentence using NLTK (Bird & Loper, 2004) and then use a Subject Verb Object extractor developed based on Honnibal & Montani (2017) to extract the subject, predicate, and object from the original sentence. The results in Table 5 show that our parsed entities have almost the same performance as that using the ground truth subjects, predicates, and objects.

Compared with Finetuned CLIP on Visual Genome To further evaluate the effectiveness of our method, we utilize the abundant relations in the entire Visual Genome dataset. Excluding the images that occurred in our Visual Genome Concept, there are 1,129,818 image-text pairs left, and we randomly took 56,490 such pairs to finetune the CLIP. The detailed setting can be found in Table 9. We then evaluate the result on Visual Genome Concept and compare it with RelationCLIP, with the results shown in Table 6. Unexpectedly, finetuned CLIP performs worse than zero-shot CLIP by a large margin, suggesting that CLIP may further learn spurious relations during finetuning on the biased dataset. This further demonstrates the superiority of our method — training-free, effective, and can mitigate the effect of spurious correlations.

Vision Encoder	CLIP	Finetuned CLIP	RelationCLIP
ResNet-50	86.85	76.43	87.85

Table 6: Comparison with finetuned CLIP on Visual Genome Concept.

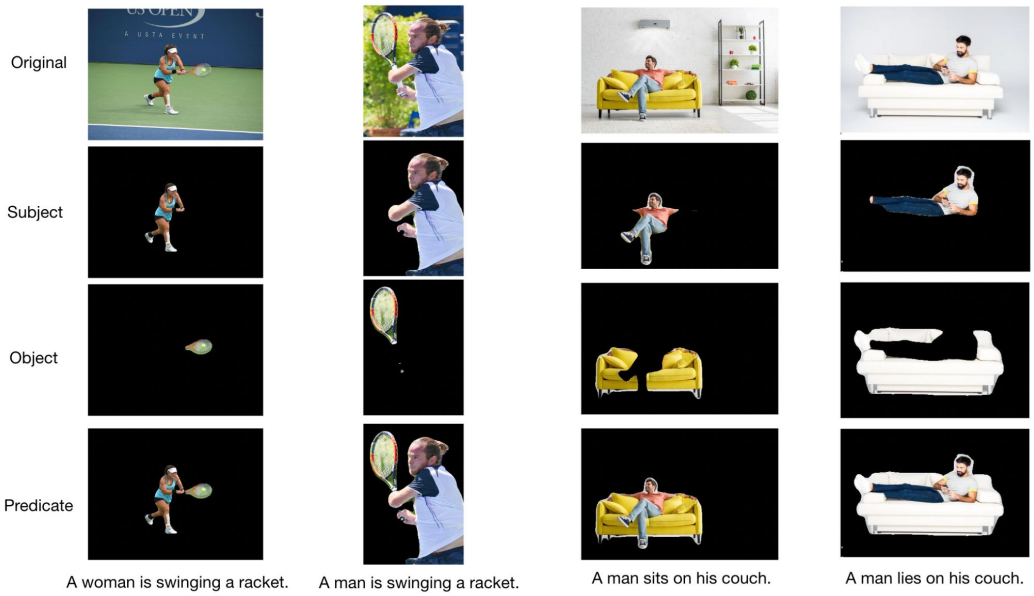


Figure 4: Examples showing the generated subject, object, and predicate sub-images. The first column and third column correspond to positive images and individual outputs of each IM for different entities. The second column and fourth column correspond to negative images and individual outputs of each IM for different entities. **Left two columns**: examples from the Visual Genome Concept dataset. (Woman, swinging, racket) is used as input (subject, predicate, object) to each IM. **Right two columns**: examples from the SVO-Probes dataset. (Man, sits, couch) is used as input to each IM. Note that for negative images, when IM could not accept the given (subject, predicate, object) and generate output sub-images, the sub-image will be replaced with the original image for entity composition.

4.4 Qualitative Comparison

Fine-grained Retrieved Samples To have a more intuitive comprehension of the proposed pipeline, we compare fine-grained matched samples by RelationCLIP, and that identified by standard CLIP in Figure 3. CLIP makes wrong prediction on all three examples, while RelationCLIP is able to discern between the two confusing images.

Extracted Entities We illustrate the individual outputs of each IM for different entities in Figure 4. In each column, we show from top to bottom: the original image X, subject image $\mathbf{S}$, object image $\mathbf{O}$, and predicate image $\mathbf{P}$.

5 Conclusion

In this work, we first make the observation that as a type of multimodal image–language transformer, CLIP could struggle in situations that require object, subject, and verb/predicate understanding when performing visual and language concept matching. Based on this observation, we propose a fine-grained training-free method for visual and language concept matching from the causal view, that could mitigate the effect of spurious relations. We also propose a new dataset to facilitate future research in this direction. We hope that our simple yet effective training-free approach could boost the development of more interpretable and principled methods for the visual and language concept matching task.

REPRODUCIBILITY STATEMENT

We release our codebase containing the methodology implementation, settings, and dataset in our submitted files along with the paper together.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

M Besserve, A Mehrjou, R Sun, and B Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *Eighth International Conference on Learning Representations (ICLR 2020)*, 2020.

Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P04-3031.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pp. 1017–1025, 2015.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pp. 752–768. Springer, 2020a.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pp. 752–768. Springer, 2020b.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017. URL https://spacy.io/.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4555–4564, 2016.

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*, pp. 4904–4916. PMLR, 2021.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=RriDjddCLN.

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

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4654–4662, 2019.

Yuheng Li, Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Mixnmatch: Multifactor disentanglement and encoding for conditional image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8039–8048, 2020.

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 *ECCV*, 2014.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pp. 852–869. Springer, 2016.

Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18051–18061, 2022.

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pp. 2623–2631, 2015.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19 (2), 2000a.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19 (2), 2000b.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

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. *arXiv preprint arXiv:2103.00020*, 2021.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046*, 2021.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022.

Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *European Conference on Computer Vision*, pp. 696–711. Springer, 2016.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual Commonsense R-CNN. *arXiv:2002.12204 [cs]*, April 2020.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2015a.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. 2015b.

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.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5534–5542, 2016.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4373–4382, 2020.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *arXiv preprint arXiv:2203.05557*, 2022.

Negative Types	Sentence	SVO triplet	Positive Image	Negative Image
Negative subject	A fox sits on the grass	<fox, sit, grass>		
Negative predicate	Person kicking a ball.	<person, kick, ball>		
Negative object	The woman sits in a chair.	<woman, sit, chair>		

Figure 5: Examples from the SVO dataset. There are three negative types for a given triplet: a subject-, verb-, or object-negative where respectively, the subject, verb, or object in the triplet are replaced by a different word.

	Seed 42		Seed 11		Seed 2	
Vision Encoder	CLIP	RelationCLIP	CLIP	RelationCLIP	CLIP	RelationCLIP
ResNet-50	80.45 %	81.30%	82.15%	83.15%	80.60%	81.70%
ViT-B-32	81.65 %	82.15%	82.55%	84.00%	82.20%	82.85%
ViT-L-14	72.70 %	73.80%	72.20%	72.45%	71.05%	73.40%

Table 7: Comparison of RelationCLIP with CLIP under three different splits on the SVO-Probes dataset.

A Appendix

A.1 Examples from SVO-Probes

In this section, we show examples from the SVO in Figure 7.

A.2 Examples from Visual Genome Concept

In this section, we show examples from the Visual Genome Concept dataset constructed by us in Figure 6.

A.3 Experimental Results on SVO-Probes over Different Splits

In this section, we show additional results using three different data splits. We use random seed $42, 11, 2$ to re-split the dataset, with the results of CLIP vs. RelationCLIP shown in Table 7 and the results of other CLIP-based model shown in Table 8.

A.4 A Case Study of Bad Examples from SVO-Probes

The improvement of our methods is relatively smaller on SVO compared with on our collected Visual Genome Concept mainly because the SVO-Probes tends to be noisy. Here, we present a case study to cover bad examples from SVO-Probes. As shown in Figure 7.

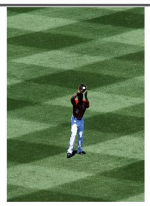
Negative Types	Sentence	SVO triplet	Positive Image	Negative Image
Negative subject	A man is eating the food.	<man, eat, food>		
Negative predicate	A man is chasing a dog.	<man, chase, dog>		
Negative object	A man is catching a football.	<man, catch, football>		

Figure 6: Examples from our constructed Visual Genome Concept dataset. There are three negative types for a given triplet: a subject-, verb-, or object-negative where respectively, the subject, verb, or object in the triplet are replaced by a different word.

	Seed 42		Seed 11		Seed 2	
Vision Encoder	Origin	Relation	Orgin	Relation	Orgin	Relation
SLIP (ViT-B-32)	77.70	77.90	79.10	79.75	81.00	80.15
SLIP (ViT-L-14)	78.90	79.70	79.70	80.15	80.10	81.30

Table 8: Effectiveness of our method using SLIP under three different splits on the SVO-Probes dataset.

A.5 Experimental Settings for Finetuning CLIP

The experimental settings for finetuning CLIP is shown in Table 9.

A.6 Additional Ablations on Using Single Entity

To validate the effectiveness of each component, we also evaluate RelationCLIP using only one entity. We sample 3000 SVO-Probes with 1000 each on subject, predicate, and object negation pairs. For Visual Genome Concept, we were able to directly tested on the entire dataset with single encoders. The results are shown in Table 10. We can see each specialized encoder has brought a certain performance increase in their corresponding area of interests compared to zero-shot CLIP. This indicates that each encoder contributes to the overall performance. However, we also noticed that among the three added encoders, the predicate encoder has the worst accuracy score in all three areas (subject, object, and predicate negated pairs) when applied alone. This may be because that CLIP is stronger at matching between the predicate semantic concept and image entities compared with matching physically visible image entities and object semantics (subjects and objects) to their name. However, we are still convinced that the predicate encoder is useful because it captures the information that connects subjects and objects, which will prevent models from making predictions solely based on objects.

Poor Quality Reason	Input Sentence	SVO triplet	Positive Image	Negative Image
Negative image also matches input sentence	A man strolls down the street.	<man, stroll, street>		
Object mismatch	A man carrying ducks on his bike.	<man, carry, bike>		
Negative image also matches input sentence; Images contain watermark	Cars passing on the highway.	<car, pass, highway>		

Figure 7: Selected bad quality examples along with reasons from the SVO-Probes dataset.

Name	Value
Optimizer	AdamW
Learning rate	0.003
Weight decay	0.05
Max epoch	10
Batch size	10

Table 9: Hyperparameter settings for finetuning CLIP.

	Visual Genome Concept			SVO-Probes		
	Subject	Predicate	Object	Subject	Predicate	Object
CLIP (ResNet-50)	86.85	65.23	87.85	82.50	76.60	88.80
RelationCLIP (Subject encoder)	89.26	66.72	87.23	83.70	77.70	88.50
RelationCLIP (Predicate encoder)	88.02	66.95	87.73	81.80	77.10	88.40
RelationCLIP (Object encoder)	86.39	66.64	91.06	83.00	78.40	91.20
RelationCLIP (All encoders)	88.34	67.26	89.51	83.02	78.10	89.80

Table 10: Ablations on using single entity.