ComCLIP: Training-Free Compositional Image and Text Matching

Kenan Jiang[*]
UC Berkeley
kenanj11@berkeley.edu

Xuehai He[*]
UC Santa Cruz
xhe89@ucsc.edu

Ruize Xu
Columbia University
rx2246@columbia.edu

Xin Eric Wang
UC Santa Cruz
xwang366@ucsc.edu

Abstract

Contrastive Language-Image Pretraining (CLIP) has demonstrated great zero-shot performance for matching images and text. However, it is still challenging to adapt vision-language pretrained models like CLIP to compositional image and text matching — a more challenging image and text matching task requiring the model's understanding of compositional word concepts and visual components. Towards better compositional generalization in zero-shot image and text matching, in this paper, we study the 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*** compositional CLIP model (ComCLIP). ComCLIP disentangles input images into subjects, objects, and action subimages and composes CLIP's vision encoder and text encoder to perform evolving matching over compositional text embedding and subimage embeddings. In this way, ComCLIP can mitigate spurious correlations introduced by the pretrained CLIP models and dynamically evaluate the importance of each component. Experiments on four compositional image-text matching datasets: Winoground, VL-checklist, SVO, and ComVG, and two general image-text retrieval datasets: Flick30K, and MSCOCO demonstrate the effectiveness of our plug-and-play method, which boosts the ***zero-shot*** inference ability of CLIP, SLIP, and BLIP2 even without further training or fine-tuning. Our codes can be found at `https://github.com/eric-ai-lab/ComCLIP`.

1 Introduction

Image and text matching (Plummer et al., 2015; Lin et al., 2014) is a fundamental task for vision-language research that involves multimodal reasoning and multi-level visual and text concept alignment. Recently, a growing number of pretrained vision-language foundation models (Radford et al., 2021; Jia et al., 2021; Li et al., 2022a,b) have shown encouraging results towards open-domain visual and language concept matching. Among these models, 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 compositional matching of disentangled concepts, especially for tasks that require the model's compositional understanding ability. For instance, Figure 1 shows some examples that CLIP fails at, which require a compositional generalization of the model to understand different subject, predicate, or object concepts.

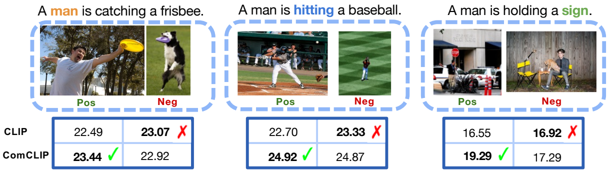

Figure 1: Examples of the compositional image-text matching problem, in which the positive and negative images have very similar semantics except for the only difference in subject, predicate/verb, or object. CLIP mistakenly connects the text prompts with the wrong images on the right (high similarity scores with negative images), while our ComCLIP model does compositional matching more effectively.

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

[*] Equally contributed.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6639–6659
June 16-21, 2024 ©2024 Association for Computational Linguistics

Therefore, there are primarily two challenges to address when adopting CLIP for compositional image and text matching. *Challenge 1*: the pre-trained language model in CLIP is biased and tends to rely on spurious relationships learned in pre-training. For example, in Figure 1 (A), CLIP associates “frisbee” with “dog” because of their more frequent co-occurrence and makes the wrong prediction. Meanwhile, the richness of entities in text descriptions brings *Challenge 2*: entity embeddings should contribute dynamically for compositional matching. In Figure 1, the subject/predicate/object entities “man/hitting/sign”, as identifiers for correct matching in each scenario, should be endowed with more importance. Based on the semantics of input images, CLIP should adjust the weights for these entity embeddings. Yet existing approaches often calculate the similarities merely based on the global embedding of images and texts and overlook fine-grained concept matching (Li et al., 2019).

To address the above limitations, we propose a new ***training-free*** framework based on CLIP-like models from the causal viewpoint, named ComCLIP. Specifically, we disentangle the visual scene into individual visual concepts and construct counterfactual subimages containing subject/object/predicate entities only. Then we utilize backdoor adjustment (Pearl et al., 2000a) to implement interventions over the disentangled subimages to mitigate the effect of spurious correlations. With this design, ComCLIP can bind the disentangled visual components with the correct word concept and avoid matching solely based on spurious correlations learned during pretraining and fine-tuning, achieving compositional generalization. To validate our approach, we formalize the compositional image and text matching task and construct a new Compositional Visual Genome (ComVG) dataset from the Visual Genome (Krishna et al., 2017) dataset for this task. We evaluated on multiple datasets: Winoground, VL-checklist, SVO-Probes (Hendricks and Nematzadeh, 2021), Flickr30K (Plummer et al., 2015), MSCOCO (Lin et al., 2014), and the ComVG dataset. Notably, ComCLIP gains an absolute accuracy improvement of 4.50% on the image score and 2.34% on the group score over CLIP and SLIP respectively on the challenging Winoground dataset.

Our contributions are summarized as follows:

- We formally define the compositional image and text matching problem and propose a novel approach named ComCLIP to address it from the causal perspective: disentangling the input image into counterfactual subimages and leverages the backdoor adjustment (Pearl et al., 2000a) to compose entity features and perform fine-grained compositional concept matching, mitigating the effect of spurious correlations introduced during training and achieving compositional generalization.
- The ComCLIP framework is ***training-free*** and can be applied to CLIP-like models for ***zero-shot*** inference without further training.
- We introduce a new dataset, the Compositional Visual Genome[1] , which contains 5400 image-text pairs with **s**ubject, **v**erb, and **o**bject annotations. This dataset was generated by creating image–sentence pairs from the Visual Genome (Krishna et al., 2017) in the same format as the SVO-Probes (Hendricks and Nematzadeh, 2021) dataset, to benchmark compositional image and text matching.
- We demonstrate the effectiveness of ComCLIP by outperforming CLIP on the Winoground, VL-checklist, SVO-Probes, and ComVG dataset over the compositional image-text matching task. We also shows its effectiveness over the general image-text retrieval task by testing Flickr30K and MSCOCO.

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 the relationship or interaction detection. (Gupta et al., 2020; Faghri et al., 2017) explore how creating hard negatives (e.g., 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., 2015a) incorporates a soft form of attention into their recurrent

[1] The dataset is available at https://drive.google.com/file/d/1rWHuq48paToXZs7_OT2Wko4l5YrAfFmR/view

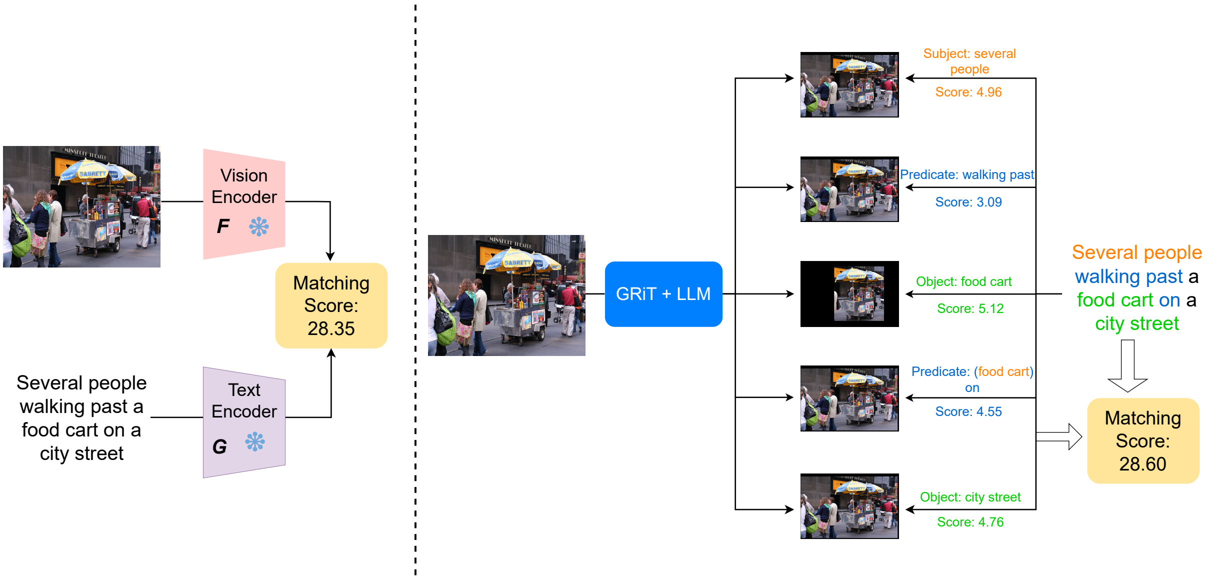

Figure 2: Overview of our ComCLIP framework using CLIP as the backbone. We disentangle the input image using GRiT (Wu et al., 2022) and the Large Language Model (LLM) by obeying the rules of encoding object, subject, and predicate respectively. The figure shows the case where multiple subjects/objects/predicates are involved (this is a positive example from Flickr30K).

model. (Ma et al., 2015) learns multiple networks that capture words, phrases, and sentence-level interactions with images 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. In this paper, we focus on the task of matching error-prone texts with images, requiring distinguishing words on a granular level — compositional image and text matching.

Pretrained Vision-Language Models Vision-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; Li et al., 2022b; Radford et al., 2021). Among them, CLIP (Radford et al., 2021) benefits from 400M curated data and defines various prompt templates to carry out zero-shot image classification. GLIP (Li et al., 2022b) has incorporated region-level alignment in its pretraining. However, these models can suffer from connecting verbs/subjects/objects concepts with visual components correctly (Hendricks and 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), (Zhang et al., 2020) executes 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 and compositonal generalization abilities of off-the-shelf pretrained vision-language models. We develop a new training-free paradigm that gains superior performance on compositional image and text 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. (Besserve et al., 2020) leverages the idea of independent mechanisms to identify modularity in pretrained generative models. (Niu et al., 2020) performs hierarchical alignments in three different granularities, i.e., global-global, global-local, and local-local alignments for description-based person re-id. (Chen et al., 2020) improves fine-grained video-text retrieval by decomposing video-text matching into global-to-local levels. (Zhang et al., 2022) proposes a multi-granularity semantic collaborative reasoning network and employs different granularity semantic representations of the question and dialog history to collaboratively identify the relevant information from multiple inputs based on attention mechanisms. (Sauer and Geiger,

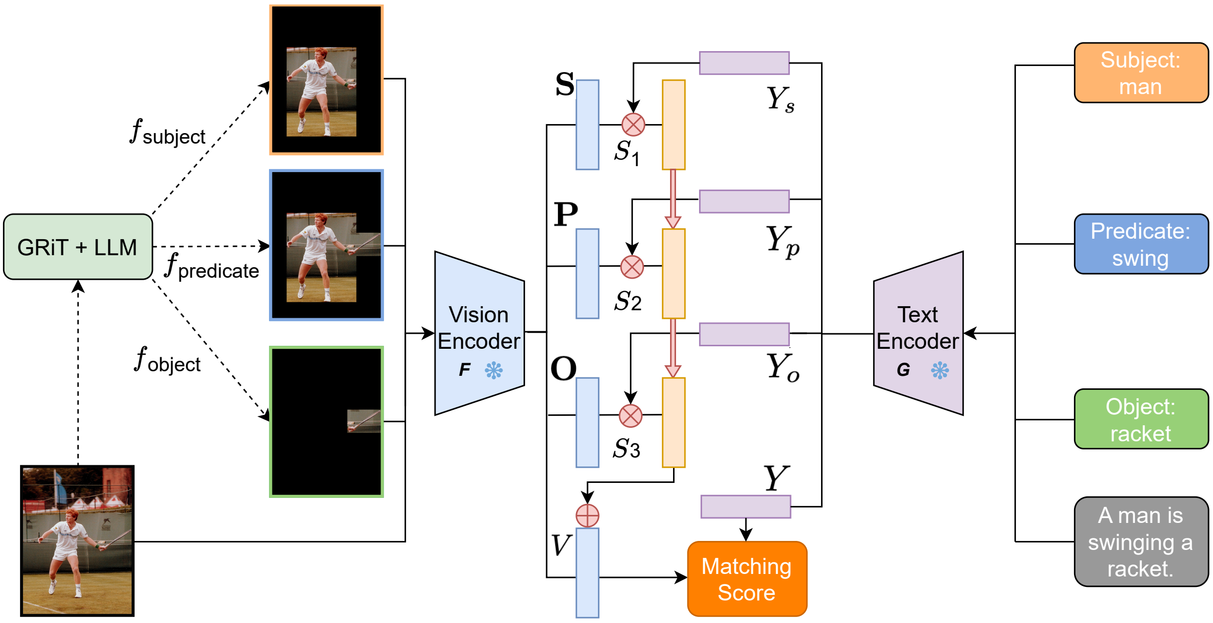

Figure 3: Overview of our ComCLIP framework using CLIP as the backbone. We disentangle the input image using three independent encoding mechanisms by obeying the rules of encoding object, subject, and predicate respectively. The entity information is introduced to the global embedding of the whole image. Module components from CLIP (vision encoder $F(\cdot)$, text encoder $G(\cdot)$) are always frozen. During implementation, the process for matching and calculating the score begins with the input image being processed into object, subject, and predicate subimages. This is followed by feeding both the original sentence and image, along with their parsed words and subimages, into the CLIP text and vision encoders. Subsequently, cosine similarity scores are computed for each pairing of subimage and word embeddings. These scores are then subjected to a Softmax layer, resulting in three positive weights. The next step involves adding the reweighted subimage embeddings to the embedding of the original image. Finally, the ultimate matching score is derived from comparing this aggregated image embedding and the global text embedding.The whole framework is ***training-free***.

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. None of these works consider the alignment of subjects, objects, and predicate entities. Different from them (Peters et al., 2017), we employ independent mechanisms to disentangle images and use generated subimages to improve fine-grained visual and language concept matching, which can mitigate spurious correlations introduced by the pretrained model.

3 Compositional Image and Text Matching

We first introduce the task of compositional image and text matching, where we are interested in improving the compositional understanding, more specifically, subject/object/predicate understanding of vision-language models. Compositional image and text matching is a task focused on enhancing the understanding of compositional elements such as subjects, objects, and predicates within CLIP-like models. This task requires an appreciation of fine distinctions between texts and their underlying compositional structure, as illustrated in Figure 1 with phrases like "man/hitting/sign." The model's ability to differentiate images that only vary by one conceptual element in their accompanying text highlights its comprehension of compositionality.

We formally define this task as follows: given text prompts Y (e.g., `"A man is hitting a baseball"`) and a set of entities $T^E = \{e^k\}_{k=1}^K$ such as hitting, where K denotes the total number of entities and e^k represents the k-th entity, the model's objective is to match the text prompts with the corresponding images. The challenge lies in the inclusion of negative images that contain mismatched entities $\{e^k\}_{k=1}^n$, where $n < k$. These negative images are designed to confuse the model, demanding a nuanced understanding of the entities within a sentence. Simply relying on nouns or spurious relations would not succeed at this task. To evaluate how well the model grasps this concept of compositionality in texts and matches them with the right images, we introduce an additional

ComVG dataset as an extended testing platform.

4 ComCLIP

We propose ComCLIP to incorporate a causal view into the CLIP-like models. We briefly introduce the background of ComCLIP in view of structured causal models in Section 4.1. Then, we present the overview of ComCLIP pipeline in Section 4.2. We introduce its critical components in depth in Section 4.3 and 4.4. Our objectives are: (i) We aim at disentangling visual input into subimages containing fine-grained compositional concepts. (ii) We intend to utilize those disentangled concepts to perform entity-level matching dynamically and mitigate the effect of spurious relations in the pretrained vision-language models learned during training.

4.1 Background

Causal inference aims to understand how changing one variable can affect another, often represented using concepts such as confounders, interventions, counterfactuals, and do-operations. In the realm of computer vision and natural language processing, the causal relationships can provide insights into the underlying generative processes.

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)). These factors, which we term confounders Z, may affect either X or Y. By disentangling these factors, we can achieve more granular image and text matching. This idea of disentanglement resonates with the principles of structural causal models (SCMs) (Pearl et al., 2000b) and independent mechanisms (IMs). An SCM is a mathematical formulation representing how variables influence one another, often composed of multiple IMs, the individual causal processes. Inspired by SCMs, our approach decomposes the subimage generation process into three independent mechanisms: object mechanism f_{object}, subject mechanism f_{subject}, and predicate mechanism $f_{\text{predicate}}$.

4.2 Method Overview

We introduce the overview of our method from a conceptual view. The pipeline is shown in Figure 2 and Figure 3. Our goal is to refine a pretrained vision-language model for fine-grained compositional image-text matching. This involves disentangling an input image to create entity-specific subimages, calculating similarity scores between these subimages and their textual counterparts, and integrating these weighted embeddings with the global image embedding. This process enables the model to capture non-spurious semantic entity information and conduct concept matching at the granular level.

4.3 Counterfactual Subimage Generation

Our method centers on the concept of causality, particularly, the Independent Mechanism (IM) assumption. In the realm of causality, the IM assumption posits that a system's variable generation process comprises autonomous modules that operate without mutual interference (Peters et al., 2017). We adopt this principle and tailor it to our context by considering three independent mechanisms for generating object, subject, and predicate subimages.

While our method is inspired by causal mechanisms, we do not make strong causal claims. Instead, we utilize the intuition that in a complex system, certain variables (or mechanisms) operate autonomously. Given the aforementioned setup, our structural causal model (SCM) takes the form: $\mathbf{O} := f_{\text{object}}\left(X\right), \mathbf{S} := f_{\text{subject}}\left(X\right), \mathbf{P} := f_{\text{predicate}}\left(X\right)$. Where $\mathbf{O}$ is the object image, $\mathbf{S}$ is the subject image, and $\mathbf{P}$ is the predicate image.

With the structural framework above, we answer counterfactual questions, a fundamental concept in causality. Specifically, we pose questions like "What if we retain only the subject/object/predicate in the original image?". The responses to such inquiries allow us to generate what we term as *counterfactual subimages*. The essence of these images is that they exclusively feature the entity in question (see Figure 3). This procedure leads to the disentanglement of the input image into three distinct and causally independent subimages.

With these foundational blocks in place, our method is geared to connect each disentangled image entity with its corresponding textual counterpart. When each entity is independently and aptly encoded, matching becomes streamlined and efficient. The remaining challenge is to craft a mechanism that effectively governs the composition process of distinct entity regions within an image.

4.4 Entity Composition

As mentioned, the pretrained CLIP-like model is prone to be biased toward specific subjects, objects or predicates, or even rely solely on one of them in the sentence.

From the causal perspective, to match image X with text prompt Y correctly, 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 $P(Y \mid X)$.

Our goal is to infer $P(Y \mid X)$ while mitigating the effects of detrimental confounders z. Leveraging Bayes Rule,

$$\begin{aligned} P(Y \mid X) &= \sum_z P(Y, z \mid X) && (1) \\ &= \sum_z P(Y \mid X, z) P(z \mid X), && (2) \end{aligned}$$

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]:

$$P(Y \mid do(X)) = \sum P(Y \mid X, z) P(z). \quad (3)$$

$do(X)$ refers to the process of mitigating the effect of harmful confounders z. These confounders z, as explained in Section 4.1, are lower-dimensional and semantically meaningful factors that include objects, subjects, and predicates within the image. By mitigating the impact of these confounders, we aim to refine our compositional matching process between the image and text. 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 entity semantics into attended regions.

To do concept matching over the text prompt Y and the entity set $T^E = \{e^k\}_{k=1}^K$, where K is the total number of entities, and e^k is the k-th entity. T^E represents a set of entities extracted from text prompts, during testing, both the image and its corresponding text, along with these parsed entities and their associated subimages, are processed through the CLIP text and vision encoders.

This interpretation motivates us to compute the 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., 2015b), we have: $P(Y \mid do(X)) \approx \text{Softmax}\left[\mathbb{E}_z\left(f_i(X, z)\right)\right]$. Specifically, to implement this on the ComVG dataset, given an 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, 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 concept-level similarity separately:

$$\begin{aligned} &S_1 = \mathcal{S}(F(f_{\text{object}}(X)), G(Y_s)), \\ &S_2 = \mathcal{S}(F(f_{\text{subject}}(X)), G(Y_o)), \\ &S_3 = \mathcal{S}(F(f_{\text{predicate}}(X)), G(Y_p)), \\ &\text{where } F(\cdot) = \text{CLIP}_{\text{vision}}(\cdot),\ G(\cdot) = \text{CLIP}_{\text{text}}(\cdot). \end{aligned} \quad (4)$$

The final visual feature is composed by:

$$\begin{aligned} V = F(X) + F(f_{\text{object}}(X))S_1 + F(f_{\text{subject}}(X))S_2 \\ + F(f_{\text{predicate}}(X))S_3. \end{aligned} \quad (5)$$

By adding compositional features back to the global image feature (as in Eq 5) and matching them with the global text features, we balance the need for detailed matching with overall context preservation.

We can compute the image-text matching score by: $O = S(G(Y), V)$. 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. During implementation, we calculate cosine similarity scores for each pair of subimage and word embedding. These scores are then transformed into weights using a Softmax layer. Subsequently, we enhance the original image embedding by adding these reweighted subimage embeddings. The final step involves computing the overall matching score by comparing this augmented image embedding with the global text embedding, thus finalizing our image-text matching process.

[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.

Our algorithm is summarized in Algorithm 1 in the Appendix, which requires ***no training or additional data***. Note that apart from CLIP, it can be easily adapted to other vision-language pretrained model with the two-stream encoder structure.

5 Experiments

5.1 Datasets

Winoground (Thrush et al., 2022) Designed to evaluate vision-language models, this dataset contains 400 instances with two image-text pairs per instance. The challenge is the differing arrangement of identical words across the pairs. Our evaluation spanned the entire dataset.

VL-checklist (Zhao et al., 2022) Distinguishing itself by combining multiple sources, VL-checklist classifies 410,000 images into three categories. We analyzed a subset of 2000 images from each category to gauge our method's effectiveness.

Flickr30K (Plummer et al., 2015) Each of the 1000 test images has 5 annotations; one annotation is selected randomly. CLIP is evaluated across the dataset; for ComCLIP, the top 10 similar images from CLIP are taken. We create subimages for the top 10 similar images and apply ComCLIP to them.

MSCOCO (Lin et al., 2014) Like Flickr30K, for each of the 1000 test images, one annotation is selected randomly. The top 10 images from CLIP undergo ComCLIP processing, and subimages are created based on parsed elements.

SVO-Probes (Hendricks and Nematzadeh, 2021) Built to assess language-image models on distinctions within image elements. From its initial 30,000 data points, we utilized 13,000 due to accessibility issues. We conducted tests using three random divisions and presented the average accuracy.

Compositional Visual Genome (ComVG) Derived from Visual Genome's (Krishna et al., 2017) 2.3 million relationships, we developed ComVG. These relationships, encompassing action and spatial aspects, are in subject-predicate-object triplets. Using these, we created image descriptions and selected 542 distinct relationship images from Visual Genome. Similar to SVO-Probes, we identified variants for each image with single discrepancies in subject, object, or predicate, resulting in 5400 curated test samples with grammatical corrections. ComVG stands out for its high-quality images and focus on text-to-image retrieval. For comprehensive dataset statistics, kindly refer Table 1. Our evaluation covered the entire ComVG.

Table 1: The number of data samples in the dataset that have one of their subjects, objects, or predicates changed between positive and negative images and the number of unique types of subjects, predicates, and objects across ComVG and SVO-Probes (SVO).

	Sub-Neg	Pred-Neg	Obj-Neg	Subjects	Predicates	Objects
ComVG	2,584	1,536	1,280	30	65	82
SVO	5,679	23,525	7,637	100	421	275

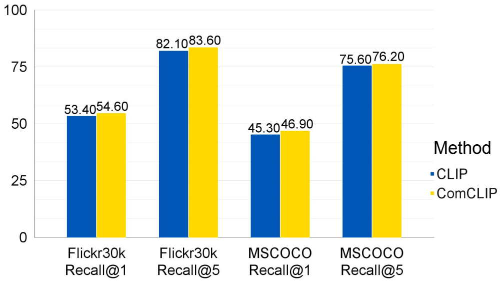

Figure 4: Comparison of Recall@1 (%) and Recall@5 (%) using CLIP and ComCLIP over the general image-text retrieval datasets.

More data examples are presented in Appendix.

5.2 Baselines

CLIP (Radford et al., 2021) We use standard CLIP, where image embeddings are generated by CLIP's vision encoder F; and text embeddings are generated by CLIP's text encoder G. The cosine similarity between them is computed to do matching.

SLIP (Mu et al., 2021) We use the SLIP ViT-L-16. Similar to CLIP, the cosine similarity between the image embeddings and text embeddings is computed to do matching.

GLIP (Li et al., 2022b) As GLIP has no global sentence and image embedding, we perform the following rule-based matching: 1) The image with more matched objects is predicted to be matching; 2) For images with the same set of objects, we compute the average confidence score of each object on both images. Larger score image is predicted.

BLIP2 (Li et al., 2023) We employed the official pretrained BLIP2. For the cosine similarity between image and text features, we adopted BLIP2's image-text contrastive learning match head as our BLIP2 baseline. Specifically, BLIP2 computes the cosine similarity score between each image embedding from each query output and the text embedding of the [CLS] token, selecting the highest similarity score as the ultimate outcome.

Table 2: Comparison of accuracy (%) on Winoground and VL-checklist using SLIP, and CLIP, and BLIP2. Results marked with ♠ are our methods.

Method	Winoground			VL-checklist			
	Text	Image	Group	Attribute	Object	Relation	Ave
SLIP	23.25	10.00	6.75	65.95	76.81	65.30	69.35
ComSLIP ♠	26.76 (+3.51)	12.12 (+2.12)	9.09 (+2.34)	67.64 (+1.69)	77.79 (+0.98)	67.02 (+1.72)	70.82 (+1.47)
CLIP	31.25	11.25	9.00	67.85	75.70	67.15	70.23
ComCLIP ♠	**34.00** (+2.75)	**15.75** (+4.50)	**10.50** (+1.50)	69.90 (+2.05)	79.00 (+3.30)	69.30 (+2.15)	72.73 (+2.50)
BLIP	29.25	12.00	8.75	79.00	84.05	73.55	78.87
ComBLIP ♠	28.75 (−0.50)	13.00 (+1.00)	10.00 (+1.25)	**79.15** (+0.15)	**84.70** (+0.65)	**73.95** (+0.40)	**79.27** (+0.40)

Table 3: Comparison of accuracy (%) on ComVG, and average accuracy (%) across the three splits on SVO-Probes using CLIP, GLIP, and ComCLIP. Results marked with ♠ are our methods. Ours could also beat GLIP, showing the superiority of our method compared with region-based vision-language pretrained models.

Method	ComVG				SVO-Probes			
	Sub	Pred	Obj	Ave	Sub	Pred	Obj	Ave
GLIP	65.95	57.50	65.75	63.85	68.91	65.14	74.94	67.81
SLIP	86.20	61.33	85.84	80.13	79.62	79.92	78.43	79.57
ComSLIP ♠	87.43	61.25	87.11	81.07	79.73	80.83	79.63	80.42
CLIP	88.61	68.52	93.85	86.38	85.53	80.77	90.53	85.60
ComCLIP ♠	**90.04**	**69.06**	**94.78**	**87.40**	**86.70**	**81.87**	**90.67**	**86.41**

Table 4: Comparison of accuracy (%) on Compositional Visual Genome and SVO-Probes using CLIP, OpenCLIP, and ComCLIP.

Vision Encoder	Compositional Visual Genome			SVO-Probes		
	CLIP	OpenCLIP	ComCLIP	CLIP	OpenCLIP	ComCLIP
ResNet-50	82.25	82.21	**83.73**	83.07	83.06	**84.17**
ViT-B-32	82.45	82.41	**84.75**	84.28	84.27	**85.18**
ViT-L-14	86.38	86.38	**87.40**	85.61	85.60	**86.41**

Table 5: Comparison of accuracy (%) on Compostional Visual Genome and SVO-Probes using different subimage configuration.

Subimage Configuraion	Compositional Visual Genome			SVO-Probes		
	ResNet-50	ViT-B-32	ViT-L-14	ResNet-50	ViT-B-32	ViT-L-14
ComCLIP	**83.73**	**84.73**	**87.40**	**84.17**	**85.18**	**86.41**
All black subimages	82.75	83.33	86.35	83.09	83.83	84.47
All original images	82.25	82.45	86.38	83.07	84.27	85.60
All subject subimages	82.46	82.55	86.46	83.18	84.10	85.24
All object subimages	83.28	83.73	86.48	83.85	84.53	85.72
All predicate subimages	82.79	83.33	86.37	83.30	84.22	85.34

5.3 Implementation Details

The process begins by processing the original image with the dense caption module of GRiT (Wu et al., 2022), producing dense image captions based on object. The input text sentence is then parsed using the large language model (LLM), `gpt-3.5-turbo`, extracting entity words and organizing them into a subject-predicate-object format. We provide the prompt for parsing sentences for entities: `Analyze the objects in this sentence, the attributes of the objects and how each object is connected.` The prompt to match objects to text entities: `Find labels of the image that refer to this object from the sentence.` The alignment between dense image captions and entity words is realized using the same LLM, mapping entity words to their image counterparts based on captions.

For creating a predicate subimage, related object and subject subimages are combined. The original sentence and image, along with their respective parsed words and subimages, are fed into the CLIP text and vision encoders. Cosine similarity scores between each image and word embedding are computed and processed through a Softmax (Jang et al., 2016) layer, yielding three positive weights. The weighted sum of the subimage embeddings is then added to the original image's global embedding to obtain the final image embedding. The methodology remains similar for SLIP (Mu et al., 2021) and BLIP2 (Li et al., 2023), termed as ComSLIP and ComBLIP respectively. Notably, for BLIP2, we project the final image embedding to the sentence embedding dimension for the score computation.

Evaluation Metrics We use Accuracy as the evaluation metric on the ComVG, SVO-Probes and VL-checklist datasets. For Winoground, we use three accuracy scores: text, image, and group score. The text score quantifies the proportion of both images correctly matched to their corresponding texts. The image score indicates the rate of both texts correctly matched to their corresponding images. Lastly, the group score signifies the accuracy of all texts and images matched correctly. We use Recall (Buckland and Gey, 1994) for Flickr30K and MSCOCO over the general image-text retrieval task.

5.4 Main Results

Compositional Image and Text Matching

Results on Winoground and VL-checklist From Table 2, ComCLIP and ComSLIP consistently outperforms CLIP and SLIP respectively across both datasets, emphasizing their ability to grasp complex image-text relationships. ComBLIP shows modest improvements, because BLIP2, pretrained on the Visual Genome dataset, already performs strongly. Overall, it shows that our method's capability to be generalized to other stronger vision-language pretrained models.

Results on ComVG and SVO-Probes In this subsection, we show the evaluation results on ComVG and SVO-Probes datasets in Table 3. Our ComCLIP can outperform zero-shot CLIP on both ComVG and SVO-Probes datasets. Separately reviewing the results, we see improvements in all negative types. This indicates that incorporating the information of subimages at inference time is helping CLIP attend to the semantic details of images and make fine-grained alignment. Apart from CLIP, we also validate the effectiveness of our method on SLIP (Mu et al., 2021), denoted by ComSLIP, with the results shown in Table 3. As presented, ours can beat SLIP on both the ComVG and SVO-Probes datasets, validating the effectiveness of our method on other CLIP-like models. In addition, we realize that our methods have lower performance improvement on the SVO-Probes dataset compared to ComVG on both CLIP and SLIP. This is because SVO-Probes contains sketchy data samples that we can not fully remove. We discuss some poor examples from SVO-Probes in the Appendix.

Comparison with GLIP We compare our methods with GLIP in Table 3. Ours outperforms GLIP by a large margin on the compositional image-text matching task, further suggesting the effectiveness of our method compared with other region-based vision-language pretrained models.

General Image-Text Retrieval Results on two image-text retrieval datasets are shown in Figure 4. CLIP and ComCLIP both perform well in Recall@5, particularly in general image-text retrieval tasks like those in the Flickr30K, where compositionality comprehension is not crucial. ComCLIP outperforms CLIP in Recall@1 on both Flickr30K and MSCOCO, due to its focus on entities and their relations, steering CLIP away from decisions based on single nouns or spurious associations. Overall, these results suggest that our method is also competitive for general image-text retrieval tasks.

5.5 Ablations and Analysis

Ablation of Different Vision Encoders The results of using different vision encoders are shown in Table 4. ComCLIP demonstrates its effectiveness on various vision encoders and also yields notable improvements over OpenCLIP (Ilharco et al., 2021), an open source implementation of CLIP.

Ablation of Different Subimage Configurations Furthermore, in Table 5, we show the efficacy of our method by comparing it against variations that employ either all black subimages or only one type of subimages. The results present that the amalgamation of subject, object, and predicate subimages achieved the highest accuracy across all vision encoders on both datasets. This implies that ComCLIP utilizes the specialized information conveyed by subimages to make accurate decisions.

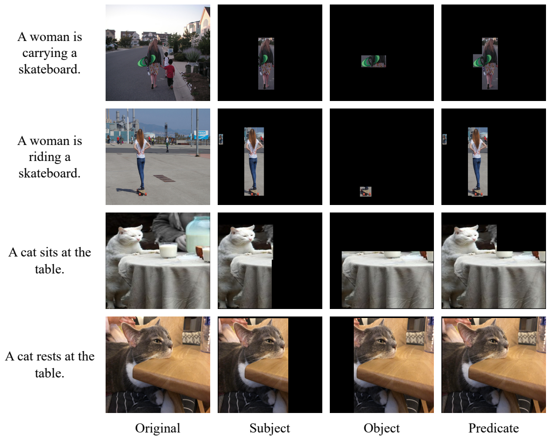
A woman is carrying a skateboard.
A woman is riding a skateboard.
A cat sits at the table.
A cat rests at the table.
Original
Subject
Object
Predicate

Figure 5: Examples of the generated subject, object, and predicate subimages. The first and third rows correspond to positive images and individual outputs of each IM for different entities. The second and fourth rows correspond to negative ones. **Top two rows**: examples from the ComVG dataset. (Woman, carrying, skateboard) is used as input (subject, predicate, object) to each IM. **Bottom two rows**: examples from the SVO-Probes dataset. (Cat, sits, table) 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 subimages, the subimage is replaced with the original image for entity composition.

5.6 Qualitative Comparison

We illustrate the individual outputs of each IM for different entities in Figure 5. In each row, we show from left to right: the original image X, subject image $\mathbf{S}$, object image $\mathbf{O}$, and predicate image $\mathbf{P}$.

6 Conclusion

In this work, we observe that CLIP-like model could struggle in situations that require object, subject, and verb/predicate understanding when performing compositional image and text matching. Based on this observation, we propose a training-free method for compositional image and text matching from the causal view, mitigating the effect of spurious relations and improving compositional generalization. We also propose a new dataset to facilitate future research in this direction. Our method is plug-and-play and could be applied to other vision-language pretrained model. We hope that our simple yet effective training-free approach could boost the development of more interpretable and principled methods for the compositional image and text matching task.

References

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

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

Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19.

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

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1218–1226.

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

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

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

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

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

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

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

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

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

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

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

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

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

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

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

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096.

Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 2022. 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*, pages 18051–18061.

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

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

Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556.

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

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

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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

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

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

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

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

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

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015a. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

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

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

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

Hongwei Zhang, Xiaojie Wang, Si Jiang, and Xuefeng Li. 2022. Multi-granularity semantic collaborative reasoning network for visual dialog. *Applied Sciences*, 12(18):8947.

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

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

This Appendix is organized as follows:

- Section A contains a detailed process of creating subimages for image-text pairs;
- Section B compares the cost for running ComCLIP and CLIP;
- Section C contains a description of a causal graph for image-text matching;
- Section D contains additional implementation details of ComCLIP;
- Section E contains additional results on varied SVO-Probes data splits;
- Section F contains case studies from Flickr30K and MSCOCO;
- Section G contains an error analysis on SVO-Probes;
- Section H contains experiments of ComCLIP's performance with different language parsers;
- Section I contains an evaluation of GRiT's robustness as our counterfactual subimage generator;
- Section J contains ablation experiments with additional counterfactual subimage generators;
- Section K contains ablation experiments on all except on one type subimages;
- Section L contains experiments presenting ComCLIP's superiority to simple entity-image matching;
- Section M contains data examples from MSCOCO;
- Section N contains data examples from Winoground, ComVG and SVO-Probes;
- Section O contains comparison between ComCLIP and finetuned ComCLIP;
- Section P contains experiments of applying our methods to Instance-level Image-text Matching Baselines;
- Section Q summarizes the detailed algorithm of ComCLIP.

A Counterfactual Subimage Generation

Figure 6 presents a visual guide to our subimage creation process for image-text pairs. For instance, GRiT analyzes the image, generating detailed captions for objects such as pizza, person, fork, and knife, along with their spatial references. Next, LLM extracts relation triplets from sentences, like *person cutting into pizza* and *person with a fork*. Utilizing LLM again, we identify all captions that could pertain to an object. To illustrate, for creating the pizza subimage, LLM recognizes that the dense caption *a pizza on a table* refers to pizza, so we use the corresponding image section of this caption. For generating the predicate cutting into subimage, we merely overlap the subimages of person and pizza, the subject and object of cutting into respectively.

B Inference Cost

This section offers a comparative analysis of the inference time for processing a single image-text pair using ComCLIP and the standard CLIP model. The evaluation, conducted over 10 trials with four V100 GPUs, incorporated pre-extracted subimages and entity words to optimize the process. The results indicate that the average inference time for the CLIP model is 0.24±0.01 seconds, while for our ComCLIP model, it is marginally higher at 0.25±0.03 seconds using the ViT-B/32 architecture. This minor increase is particularly noteworthy as it falls within the same order of magnitude, underscoring the efficiency of ComCLIP in maintaining comparable processing speeds.

Furthermore, the GPU memory consumption during inference was also assessed. The CLIP model utilized 2047±44 MB, and ComCLIP required slightly more at 2086±98 MB. This modest increment in memory usage is offset by the enhanced capabilities of ComCLIP, affirming its practicality for deployment in similar computational settings. Thus, ComCLIP stands out as an efficient solution, offering advanced functionalities with only a nominal increase in resource requirements.

C Causal Graph in the Context of Image-text Matching

We show the causal graph in the context of our image-text matching task in Figure 7. X are high-dimensional observations (i.e., images), and Y are

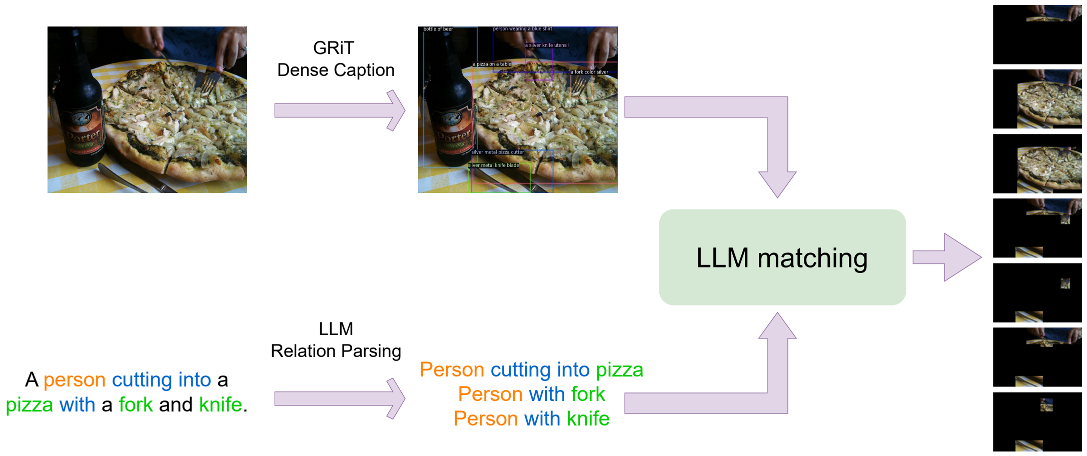

Figure 6: Design of the counterfactual subimage generation process. LLM matches the dense captions generated by GRiT from image to parsed subjects, objects, predicates from text.

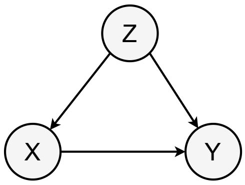

Figure 7: The causal graphs in the context of compositional image-text matching.

corresponding text prompts. 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)).

D Implementation Details

We introduce the implementation details of ComCLIP in this section. Our pipeline is training-free, so there are no parameters involved in ComCLIP. In the main paper, we use the CLIP model with a ViT-B-32 vision encoder for the results in Table 2, and a ViT-L-14 vision encoder for the results in Table 3. The masks for subjects/objects/predicates are generated using GRiT (Wu et al., 2022) with the dense caption version, which is pre-trained for 200 epochs.

E Experimental Results on SVO-Probes over Different Splits

In this section, we show additional results using three different data splits on SVO-Probes. We use random seeds $42, 11, 2$ to re-split the dataset, with the results of CLIP vs. ComCLIP shown in Table 6 and the results of other CLIP-based models shown in Table 7.

F Case Study: Generalized Scenario with Multiple SVO

In this section, we present the case study where the text contains multiple SVOs on Flickr30K and MSCOCO.

F.1 Cases Study on Flickr30K

In Figure 9, we first show the case where single SVO are involved.

In Figure 2 and Figure 10, we show the case where multiple SVOs are involved. Specifically, in this provided case, multiple objects (Food cart, City street) and subjects (Several People, Food cart) are involved. Figure 2 is a positive example, and Figure 10 is a negative example. As can be seen, ComCLIP can utilize multiple subjects/objects/predicates in the input texts to do the matching. The food cart object dominates the decision process and helps ComCLIP make the correct match.

F.2 Case Study on MSCOCO

In Figure 11 and 12, we provide a breakdown of how ComCLIP makes the correct decision when multiple SVOs are involved on MSCOCO. Both the negative image from Figure 11 and the positive image from Figure 12 are closely aligned with the text, featuring prominent visual entities such as a person and a pizza. ComCLIP integrates various subjects, objects, and predicates, effectively distinguishing the correct image match from a pair of visually analogous images.

Table 6: Comparison of ComCLIP with CLIP under three different splits on the SVO-Probes dataset.

	Seed 42		Seed 11		Seed 2	
Vision Encoder	CLIP	ComCLIP	CLIP	ComCLIP	CLIP	ComCLIP
ResNet-50	82.77	83.87	82.06	83.10	82.87	83.97
ViT-B-32	84.13	85.47	84.47	84.83	84.17	84.67
ViT-L-14	85.53	86.63	84.76	86.10	85.27	86.33

Table 7: Effectiveness of our method using SLIP under three different splits on the SVO-Probes dataset.

	Seed 42		Seed 11		Seed 2	
Vision Encoder	SLIP	ComSLIP	SLIP	ComSLIP	SLIP	ComSLIP
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

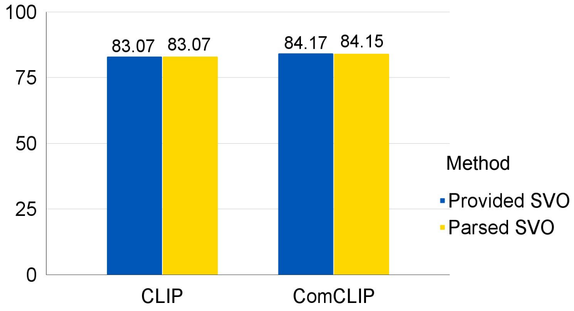

Figure 8: Comparison of accuracy (%) on SVO-Probes using parsed and ground-truth SVO triplets.

G Error Analysis

As shown in the main paper, we get higher improvements using ComCLIP on Compositional Visual Genome compared with SVO-Probes. This is mainly because our collected Compositional Visual Genome is cleaner and the SVO-Probes dataset tends to be noisy. Herein, we give a case study covering three major error-inducing issues found in SVO-Probes, as depicted in Figure 13: instances where the negative image aligns with the input sentence, object mismatches, and the presence of watermarks in images.

H Extracted Entities

Use Language Parser to Extract SVO The performance of ComCLIP is also dependent on the quality of the subject, object, and predicate entity provided. To study the effect of extracted entities, we analyze our methods on SVO-Probes since it has more complex sentence structures. Apart from the LLM approach shown in the main paper, we remove stop words from the sentence using NLTK (Loper and Bird, 2002) and then use a Subject Verb Object extractor developed based on (Honnibal and Montani, 2017) to extract the subject, predicate, and object from the original sentence. Figure 8 shows that our parsed entities have almost the same performance as that using the ground truth subjects, predicates, and objects.

I Robustness of Counterfactual Subimage Generator

To show the robustness of using GRiT (Wu et al., 2022) to generate counterfactual subimages, we quote the results and conclusions from (Wu et al., 2022). According to (Wu et al., 2022), GRiT is comparable to the closed-set object detector with a 0.8 AP gap. This result demonstrates GRiT's open-set framework can serve as a new promising formulation for object detection. GRiT also performs comparably with the state-of-the-art closed-set object detectors. This once again demonstrates GRiT can serve as the subimage generator in our pipeline.

J Additional Ablations on Counterfactual Subimage Generation

In this section, we show that ComCLIP is **robust** to the choice of counterfactual subimage generator. We use segmentation models, Lang-Seg and CLIPSeg (Lüddecke and Ecker, 2022) with the `clipseg-rd64-refined` version, to create segmentation masks and generate subimages. Specifically, given the input (subject, object, predicate)

Table 8: Compositional Visual Genome subset accuracy (%) with masks generated by Lang-Seg and CLIPSeg.

	ResNet-50	ViT-B-32	ViT-L-14
CLIP	79.38	79,94	83.70
ComCLIP (Lang-Seg mask)	82.41	**83.15**	**86.05**
ComCLIP (Lang-Seg mask with blur)	83.27	83.09	85.31
ComCLIP (CLIPSeg mask)	**83.27**	82.22	85.18
ComCLIP (CLIPSeg mask with blur)	82.78	82.90	85.31

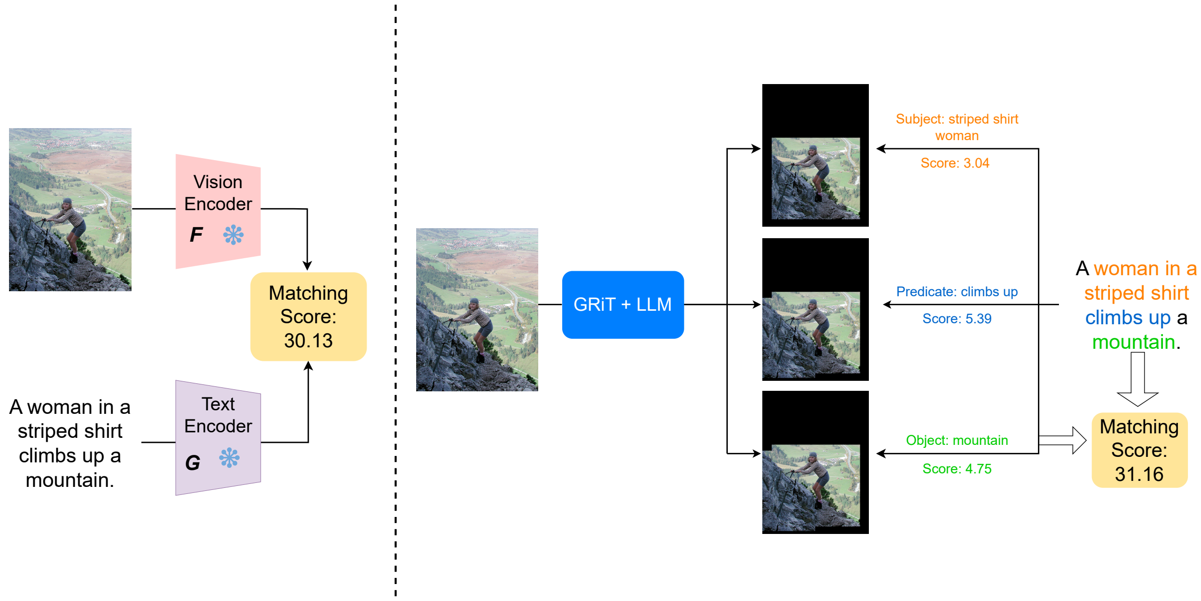

Figure 9: The comparison of CLIP (left) and ComCLIP (right) over the case where single subjects/objects/predicates are involved. Image and text examples are from Flickr30K.

triplet, we model the object mechanism f_{object} using a binary mask generated by Lang-Seg and CLIPSeg, which are both CLIP-based language-guided segmentation models. Given the segmentation results, the object part will be set to 1 while the remainder of the image is 0. 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 combining 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.

We test the masks on a randomly selected 30% subset of Compositional Visual Genome. The results in Table 8 indicate that ComCLIP continues to outperform CLIP across all vision encoders when the masks are generated by CLIPSeg. To further test its robustness, we add noise by applying Gaussian image blur to the backgrounds of generated subimages rather than using pure black backgrounds. Despite the blurring, ComCLIP using either Lang-Seg or CLIPSeg masks still performs better than CLIP and achieves similar performance to ComCLIP without blur as shown in Table 8. Thus, ComCLIP is shown to be resilient to the precision of generated subimages.

K Additional Ablations on All Except One Type Subimages

We test 3 combinations of subimages on a balanced randomly sampled 3000 subset of ComVG, presented in Table 9. As can be observed, ComCLIP outperforms all 3 scenarios in which all but one subimage are utilized, confirming that ComCLIP effectively leverages the composite information for reasoning.

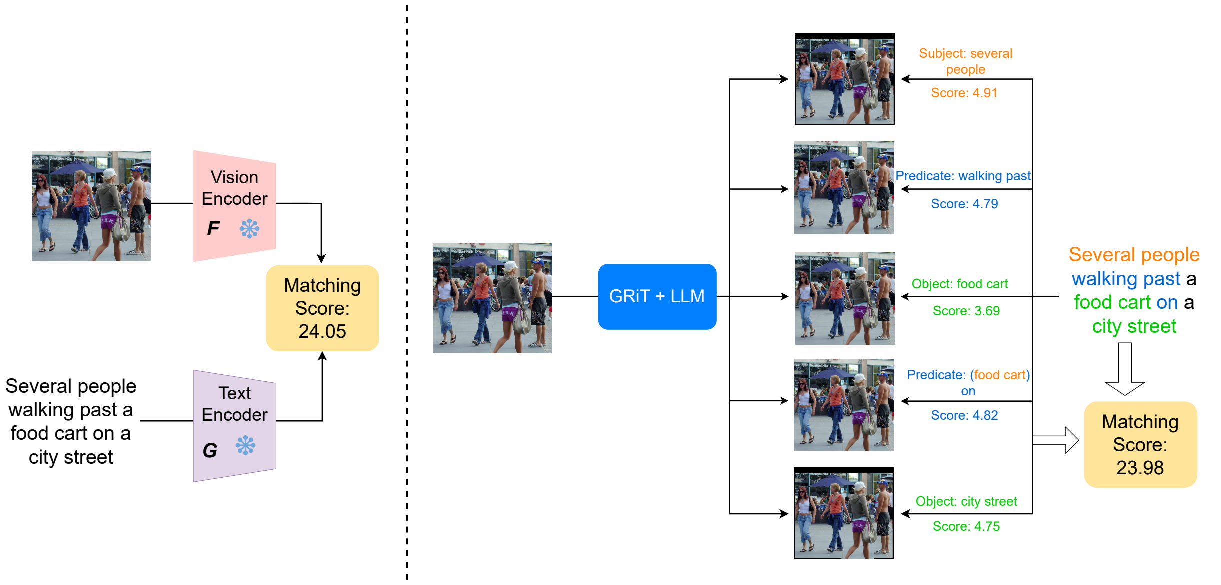

Figure 10: The comparison of CLIP (left) and ComCLIP (right) over the case where multiple subjects/objects/predicates are involved (this is a negative example from Flickr30K).

Table 9: Results of different subimg configurations (ComVG)

Vision Encoder	All Sub & Obj	All Sub & Pred	All Obj & Pred	ComCLIP ♠
ResNet-50	81.03 (**-0.04**)	80.47 (**-0.60**)	79.73 (**-1.34**)	**81.07**
ViT-B-32	80.70 (**-0.10**)	79.33 (**-1.47**)	80.60 (**-0.20**)	**80.80**
ViT-L-14	84.37 (**-0.06**)	83.33 (**-1.10**)	83.73 (**-0.70**)	**84.43**

L Additional Ablations on Comparisons with Fine-grained Similarity Matching Methods

In our additional analysis, detailed in Table 10, we explore the impact of matching individual parsed entity words with images, as opposed to full sentences, employing the CLIP architecture as our foundation. The results demonstrate that ComCLIP markedly surpasses the performance of three baseline models on four entity scenarios, which are based solely on the similarity between a single entity word and an image. This highlights the superior efficacy of ComCLIP.

M Data Examples from MSCOCO

In this section, we provide an example from the MSCOCO dataset that we constructed, as shown in Figure 14. The MSCOCO dataset typically incorporates adjectives to enhance query sentences, which CLIP tends to overlook. For instance, in the provided example, the orange road sign helps ComCLIP successfully identify the accurate image as the best match, while CLIP does not rank it among the top 5 matches.

N Data Examples from Winoground, ComVG and SVO-Probes

In this section, we show examples from Winoground in Figure 15. Winoground presents a challenging task, requiring precise match of two image-text pairs to successfully earn a group score. We also show the ComVG dataset constructed by us and the SVO-Probes in Figure 16. As can be seen, they are formatted similarly: Negative Types — Sentence — SVO Triplet — Positive Image — Negative Image. Visual Genome is licensed under a Creative Commons Attribution 4.0 International License. Compositional Visual Genome dataset is compatible with the original access conditions of Visual Genome.

O Compared with Finetuned ComCLIP

In addition to the original ComCLIP model, we explored the effects of finetuning ComCLIP using the MSCOCO dataset, subsequently evaluating its performance on the ComVG dataset and SVO-

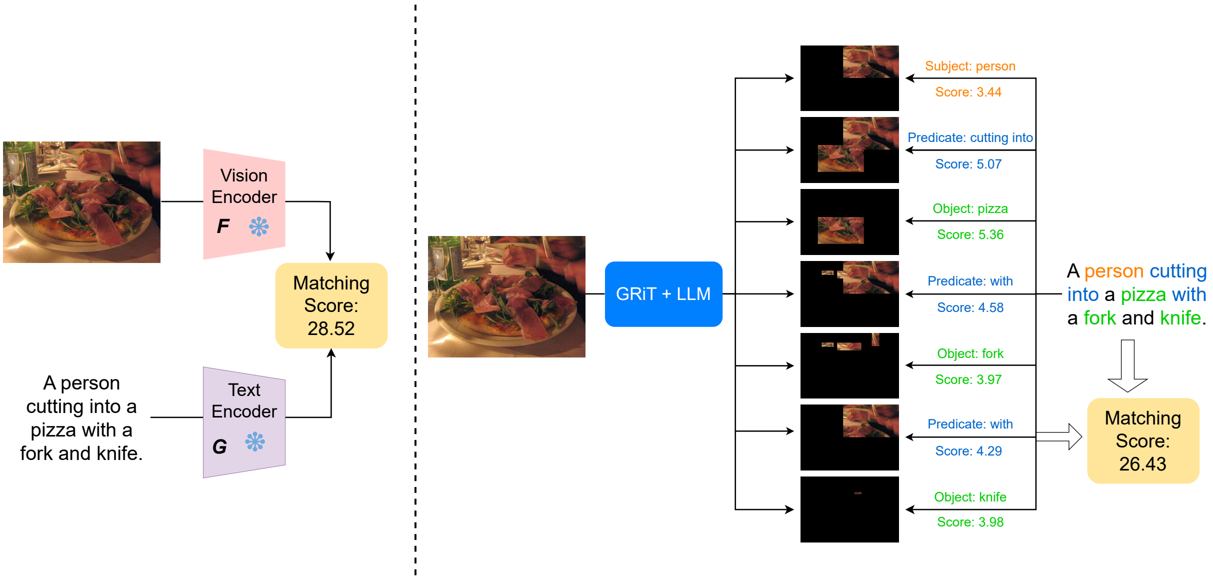

Figure 11: The comparison of CLIP (left) and ComCLIP (right) over the case where multiple subjects/objects/predicates are involved (this is a negative example from MSCOCO).

Table 10: Fine-grained similarity w/ parsed words (ComVG)

Vision Encoder	Subject Entity	Predicate Entity	Object Entity	All Entity	ComCLIP ♠
ResNet-50	61.37 (-22.36)	52.98 (-30.75)	68.76 (-14.97)	81.17 (-2.56)	**83.73**
Vit-B-32	60.57 (-24.18)	53.35 (-31.40)	69.91 (-14.84)	81.44 (-3.31)	**84.75**
Vit-L-14	62.17 (-25.23)	54.98 (-32.42)	70.52 (-16.88)	84.85 (-2.55)	**87.40**

Probes dataset. This process involved a approach to training example construction: for each query text, we utilized the CLIP model to identify the most challenging negative image from the MSCOCO training set, based on the highest similarity score. This method aimed to enhance the model's ability to discern subtle distinctions between closely related visual-textual pairs. The resulted finetuned CLIP is still evaluated in a zero-shot fashion on the target evaluation dataset, i.e., how well does finetuned ComCLIP trained on MSCOCO transfer to target datasets. The results, as outlined in Table 11, demonstrate notable improvements in the finetuned ComCLIP's performance compared to both the standard CLIP and the unfinetuned ComCLIP models.

The finetuned ComCLIP model exhibited significant gains across all categories on both the ComVG and SVO-Probes datasets. Particularly, the average accuracy on the ComVG dataset increased from 84.63% for ComCLIP to 86.98% for the finetuned version, underscoring the effectiveness of finetuning in enhancing model performance. Similarly, on the SVO-Probes, there was an increase from 86.41% to 87.99%. These improvements are most prominent in the 'Object' category of the ComVG dataset, where the finetuned ComCLIP achieved a 97.83% accuracy, indicating a substantial enhancement over the original model's performance.

These results suggest that finetuning on a dataset with diverse visual and textual representations, such as MSCOCO, significantly improves the model's capability to generalize and transfer learned features to different, yet related, datasets. The enhancements in accuracy, particularly in the 'Object' recognition category, could be attributed to the comprehensive and varied nature of objects represented in the MSCOCO dataset, which may have provided a more robust learning experience for the model.

This analysis indicates that while the original ComCLIP model is effective and can improve over the CLIP pipeline in zero-shot learning tasks, its performance can be further enhanced through finetuning on a suitably diverse dataset. This enhancement is critical for tasks requiring nuanced understanding of visual and textual data. Future work could explore the impact of finetuning on other

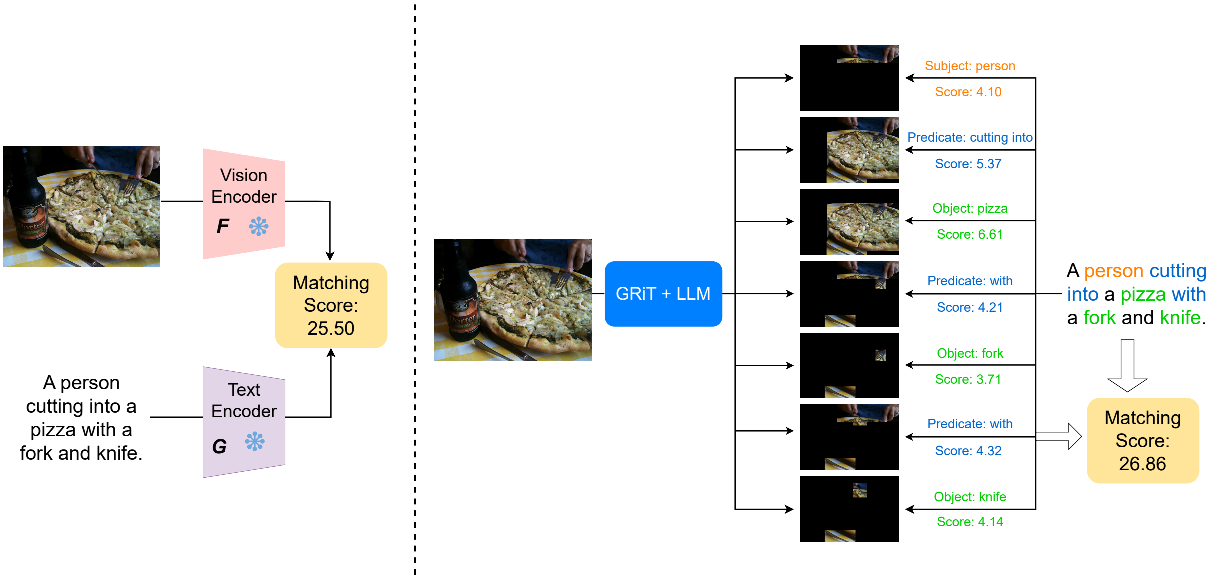

Figure 12: The comparison of CLIP (left) and ComCLIP (right) over the case where multiple subjects/objects/predicates are involved (this is a positive example from MSCOCO).

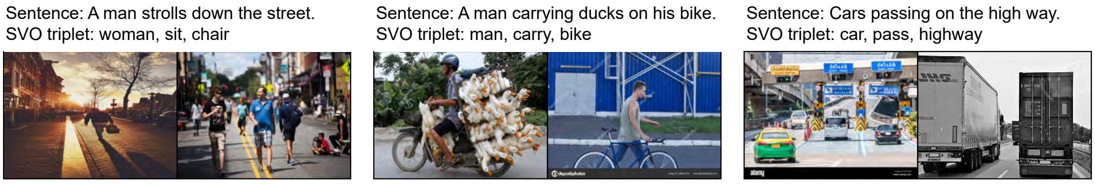
Sentence: A man strolls down the street.
SVO triplet: woman, sit, chair
Sentence: A man carrying ducks on his bike.
SVO triplet: man, carry, bike
Sentence: Cars passing on the high way.
SVO triplet: car, pass, highway

Figure 13: Selected bad quality examples from the SVO-Probes dataset.

Table 11: Comparison of accuracy (%) on ComVG, and SVO-Probes using ComCLIP and finetuned ComCLIP.

Method	ComVG				SVO-Probes			
	Sub	Pred	Obj	Ave	Sub	Pred	Obj	Ave
CLIP	88.61	68.52	93.85	83.66	85.53	80.77	90.53	85.61
ComCLIP	90.04	69.06	94.78	84.63	86.70	81.87	90.67	86.41
Finetuned ComCLIP	**92.14**	**69.74**	**97.83**	**86.98**	**87.44**	**81.90**	**92.48**	**87.99**

datasets or using different finetuning strategies to further understand the adaptability of the ComCLIP model.

P Instance-level Image-text Matching Baselines

We further evaluate ComCLIP's applicability to instance-level image-text matching models by integrating it with SGRAF (Diao et al., 2021) on the ComVG dataset. This implementation involves processing the input texts with the same parsing technique used in ComCLIP, coupled with the utilization of grounded image regions for computing the matching score, followed by a reweighting step. The integration of ComCLIP results in a notable performance enhancement: the matching accuracy increases from 76.79% without ComCLIP to 78.9% with ComCLIP.

Q Detailed Algorithm

The detailed ComCLIP algorithm is summarized in Algorithm 1.

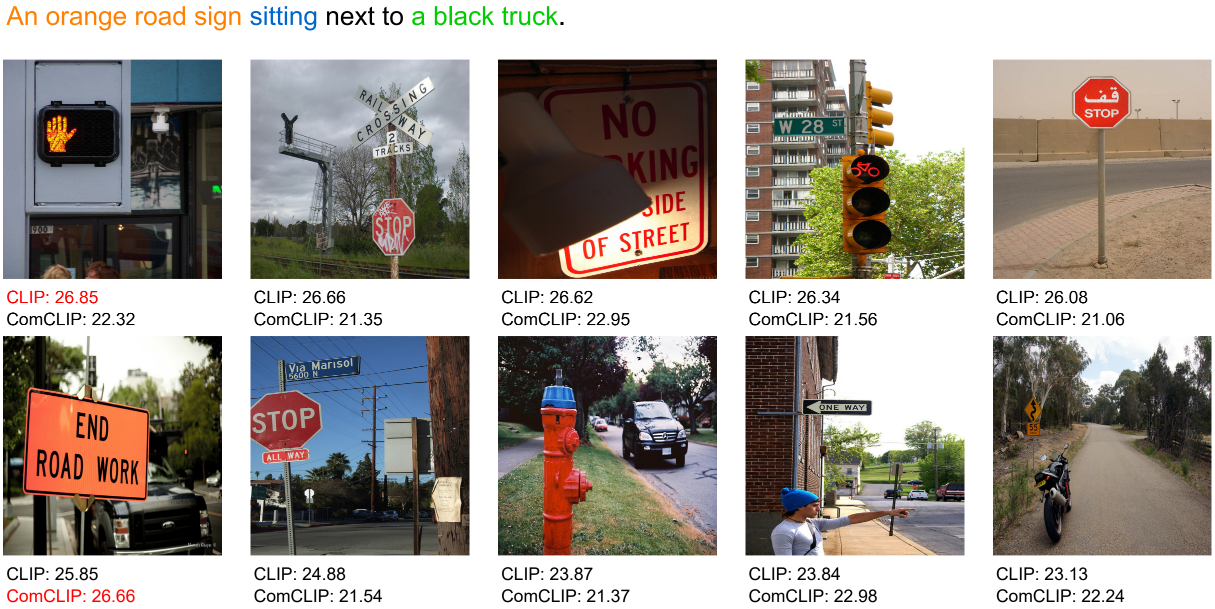

Figure 14: Example from MSCOCO dataset.

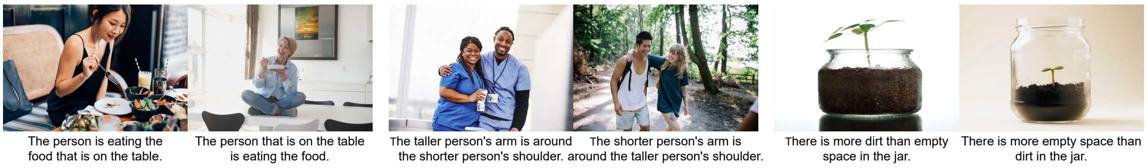
The person is eating the food that is on the table.
The person that is on the table is eating the food.
The taller person's arm is around the shorter person's shoulder.
The shorter person's arm is around the taller person's shoulder.
There is more dirt than empty space in the jar.
There is more empty space than dirt in the jar.

Figure 15: Examples from Winoground dataset.

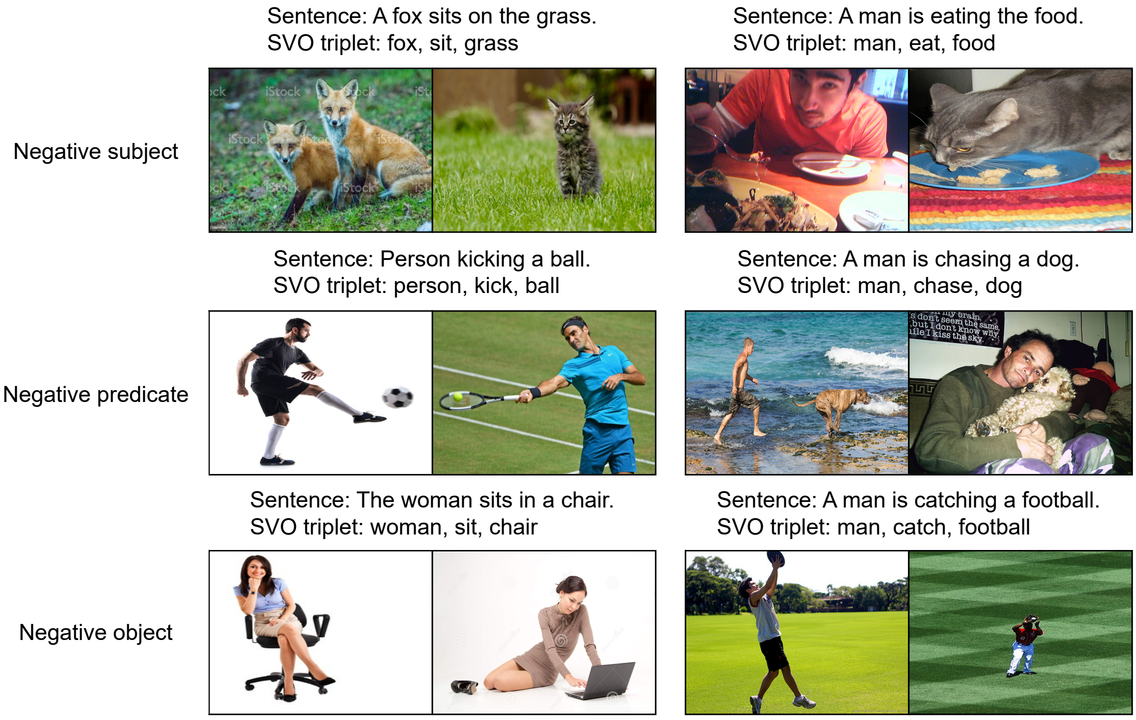
Sentence: A fox sits on the grass.
SVO triplet: fox, sit, grass
Sentence: A man is eating the food.
SVO triplet: man, eat, food
Negative subject
Sentence: Person kicking a ball.
SVO triplet: person, kick, ball
Sentence: A man is chasing a dog.
SVO triplet: man, chase, dog
Negative predicate
Sentence: The woman sits in a chair.
SVO triplet: woman, sit, chair
Sentence: A man is catching a football.
SVO triplet: man, catch, football
Negative object

Figure 16: Examples from the SVO-Probes dataset (left two) and Compositional Visual Genome dataset (right two). There are three negative types for a given triplet: a subject-negative, predicate-negative, or object-negative where respectively, the subject, predicate, or object in the triplet is replaced by a different word. Within each image pair, the positive image on the left represents the positive triplet, while the negative image on the right corresponds to the negative triplet.

Algorithm 1 Training-Free Compositional Image and Text Matching with ComCLIP.

```
Require:
    Input: image $X$, text prompt $Y$, 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 $O$.
Generate counterfactual subimages $\mathbf{O}, \mathbf{S}, \mathbf{P} \leftarrow f_{\text{object}}(X), f_{\text{subject}}(X), f_{\text{predicate}}(X)$;
Extract feature embeddings $F(\mathbf{O}), F(\mathbf{S}), F(\mathbf{B}) \leftarrow \mathbf{O}, \mathbf{S}, \mathbf{P}$;
Extract (subject, object, predicate) words $Y_s, Y_o, Y_p \leftarrow Y$;
Compute the concept-level similarity $S_1, S_2, S_3 \leftarrow G(Y_s), G(Y_o), G(Y_p), F(\mathbf{O}), F(\mathbf{S}), F(\mathbf{P})$; {Eq. (3)}
Extract sentence embeddings $G(Y) \leftarrow Y$;
Compose visual features $V \leftarrow S_1, S_2, S_3, f_{\text{object}}(\cdot), f_{\text{subject}}(\cdot), f_{\text{predicate}}(\cdot), F(\cdot), X$; {Eq. (4)}
Compute the matching score $O \leftarrow Y, V$
```