This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Compositional Image-Text Matching and Retrieval by Grounding Entities

Madhukar Reddy Vongala, Saurabh Srivastava, Jana Košecká
Department of Computer Science,
George Mason University,
Fairfax, VA, USA
{mvongala, ssrivas6, kosecka}@gmu.edu

Abstract

Vision-language pretraining on large datasets of images-text pairs is one of the main building blocks of current Vision-Language Models. While with additional training, these models excel in various downstream tasks, including visual question answering, image captioning, and visual commonsense reasoning. However, a notable weakness of pretrained models like CLIP, is their inability to perform entity grounding and compositional image and text matching [8, 27, 28, 36]. In this work we propose a novel learning-free zero-shot augmentation of CLIP embeddings that has favorable compositional properties. We compute separate embeddings of sub-images of object entities and relations that are localized by the state of the art open vocabulary detectors and dynamically adjust the baseline global image embedding. The resulting embedding is then utilized for similarity computation with text embedding, resulting in a average 1.5% improvement in image-text matching accuracy on the Visual Genome and SVO Probes datasets [5, 11]. Notably, the enhanced embeddings demonstrate superior retrieval performance, thus achieving significant gains on the Flickr30K and MS-COCO retrieval benchmarks [19, 25], improving the state-of-the-art Recall@1 by 12% and 0.4%, respectively. Our code is available at https://github.com/madhukarreddyvongala/GroundingCLIP.

1. Introduction

Vision-language models (VLMs) have demonstrated significant success in image-text matching and retrieval, achieving strong zero-shot generalization across various vision-language tasks through pretraining on large multi-modal datasets. Models like CLIP [26] and BLIP-2 [16] can effectively retrieve and align images with textual descriptions without requiring task-specific fine-tuning. However, despite their impressive performance, these models face challenges in compositional generalization. During the pretraining stage, CLIP focuses on aligning images and text as a whole, disregarding the compositional structure of the caption. As a result, its final embeddings often struggle to accurately associate subject and object noun phrases with their corresponding locations in images and face difficulties in handling predicates and relationships in novel configurations [27]. Although pretraining involves ingesting large amounts of image-text pairs, it also captures unintended correlations present in the datasets. When certain objects frequently co-occur, the learned representations develop biases, making it difficult to reason about compositional structures that were not seen during training. This often results in errors and hallucinations in downstream VQA tasks [4] and poor alignment of image and text embeddings. For instance, the similarity between the caption "a man holding a sign" and the image depicting "a man holding the glasses" (see Figure 1) is high—primarily due to dataset priors rather than compositional computation of similarity. Similarly, the baseline CLIP embedding of the phrase "a man holding a sign" will be closer to the image of



Figure 1. For the caption A man is holding a sign (positive image on right side and negative image on left side) Grounding CLIP scores are **0.2259** and 0.2151 ✓; CompCLIP are 0.2152 and 0.2226 ✗, CLIP are 0.2112, 0.2123 ✗. Note that according to Grounding CLIP the positive images associated with the caption have higher similarity score than negative images.

"a man holding the glasses", due to a more confident grounding of the subject "a man" in the right image. This will cause the baseline model to incorrectly rank the similarity of images and text, affecting the performance of image-text matching and retrieval tasks. To address these limitations, we propose Grounding CLIP, a training-free approach that dynamically refines the global image embedding $\boldsymbol{I}$ using noun phrase grounding. This enhances the model's ability to better capture the compositional structure of the images and captions, resulting in a more accurate similarity score $\mathbf{sim}(I, T)$ between image and text.

We utilize the GPT-3.5 language model to break down the text description into object entities and relations, and leverage a state-of-the-art open-vocabulary detector Grounding DINO [21] to localize the phrases corresponding to entities and relations in the image. The localized sub-images embeddings are used to dynamically adjust the global image embedding $\boldsymbol{I}_c$, which is used in the image-text cosine similarity computation. We term the approach Grounding CLIP (GCLIP) and evaluate its performance on several image-text matching and retrieval benchmarks. Our contributions are summarized as follows:

- We introduce Grounding CLIP, a training free enhancement for compositional image-text matching and image retrieval, leveraging object detection based grounding for improved subject-object alignment.
- We achieve state-of-the-art results on ComVG and SVO Probes, consistently outperforming the ComCLIP model and demonstrating better zero-shot compositional understanding using image-text matching evaluation.
- We achieve significant improvements in retrieval tasks on the Flickr30K and MS-COCO datasets using the proposed embeddings improving the state-of-the-art Recall@1 by 12% and 0.4%, respectively.

2. Related Work

Vision Language Models Foundational Vision and Language Models (VLMs) have demonstrated impressive performance in various vision and language tasks, including visual question answering (VQA), retrieval, image-text matching, referring expression comprehension and captioning. While the earlier models [3, 18, 22, 32, 39] achieved impressive performance using smaller amounts of training data (∼9.7 million image-text pairs) from specially curated datasets, the later generation of VLM's such as ALIGN, BLIP-2, ALBEF, FLAVA, Flamingo, PALI, LLaVA [1, 2, 7, 14, 16, 20, 30] relies on the effective pre-training of image and text representations using larger web scale datasets (∼ 400 million image-text pairs) followed by multi-modal learning techniques that combine contrastive learning, masked token modeling, and frozen large language models (LLMs), using another ∼ 0.5 million image-text pairs. These models have been shown to improve vision-text alignment, cross-modal reasoning, and zero-shot generalization on variety of datasets and downstream tasks.

Probing Studies Multiple probing studies of fine-grained understanding capabilities of VLM's have demonstrated poor performance on tasks that require more compositional reasoning. For example, [31] showed on the CLEVR [9] and RefCOCO datasets that if a spatial clause is added to the caption, the performance of image-text matching is at the level of random chance. Additional probing studies using specially curated datasets demonstrated that these models lack the understanding of attribution, relations, order, verbs, and counting [5, 10, 33, 38].

Grounding and Compositionality The lack of grounding and compositional generalization of the pretrained representations has been pointed out in several previous works. Authors in [13] focused on CLIP and basic types of compositions (adjectives and nouns, two adjectives and two nouns) in a controlled setting on CLEVR dataset. The linguistic structure and parsing has been used to do zero-shot referring expression comprehension in [31]; where CLIP was repurposed by region-scoring method that isolates object proposals via cropping and blurring, and passes them to CLIP. This demonstrated the effectiveness of grounding individual noun phrases, relations and combining the evidence using structured probabilistic model. Similar approach was proposed in [27] for special case of spatial relations understanding. Poor grounding and localization ability of referring expressions was studied in [36], where notable improvements were achieved by proposing novel loss function for training multi-modal model with bounding box supervision. Additional recent works AMC, OWL, GLIP, Grounding DINO or FLORENCE [17, 21, 23, 35, 36] used explicit bounding box supervision to improve localization and grounding performance. Authors in [8] used GRiT model [34] trained for dense captioning tasks to localize the phrases and dynamically adjusted the global CLIP embeddings to improve similarity score computation.

Our approach is more closely related to Jiang et al. [8], with few notable departures. We use an LLM for guided decomposition of text into object entities and relations,

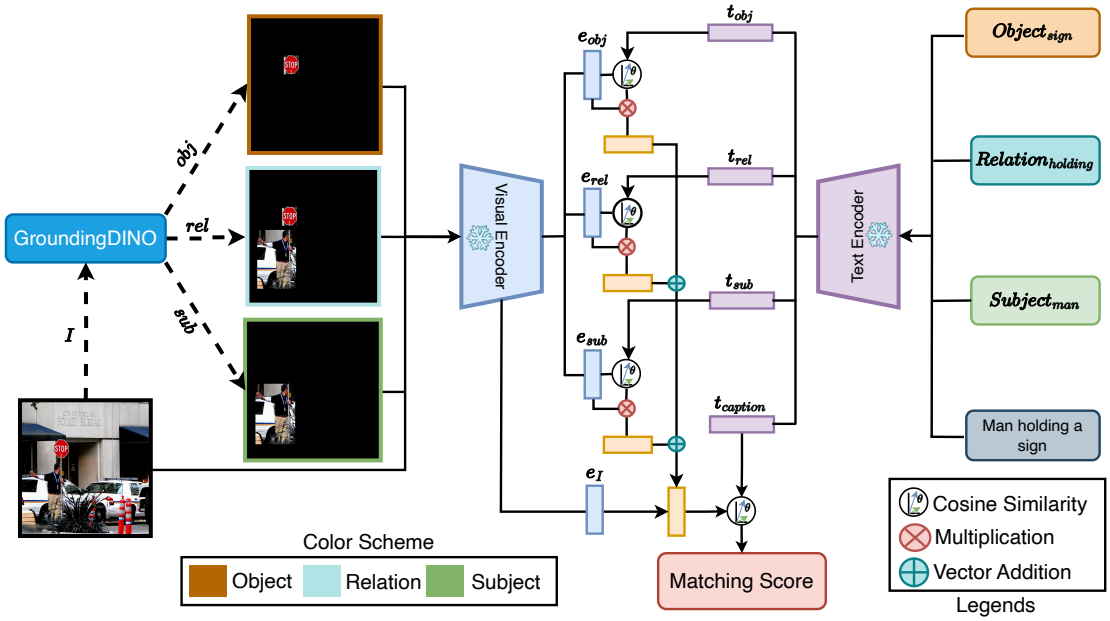


Figure 2. Overview of the proposed Grounding CLIP (GCLIP) framework. Given a caption such as "A man holding a sign," we prompt the GroundingDINO with 'man' and 'sign' to localize relevant regions in the image. The resulting sub-images are embedded using CLIP and fused with the global image embedding through a weighted sum based on cosine similarity with their corresponding text phrases. This adjusted image embedding enables more accurate compositional alignment with the caption.

followed by prompting the state-of-the-art open vocabulary grounding model [21] for localization. The use of normalized similarity scores instead of `Softmax` has also been observed as a contributing factor to the improved performance, due to the fact that individual components of the sentence are not independent.

3. Grounding CLIP

Our approach, shown in Figure 2, builds upon Jiang et al. [8], that dynamically adjusts image embeddings $\boldsymbol{I}_e$ through text query decomposition, to improve grounding and compositionality of image and text alignment.

To address the compositionality aspects, we decompose the text into $(subject, relation, object)$ tuples or multiple entities and relations $(ent_1, \ldots, ent_K, rel_1, \ldots, rel_L)$ using GPT-3.5-turbo. First, we prompt the language model to identify all objects in the caption, extract their associated attributes if present and describe how the objects are connected using subject-verb-object relations. We guide the model with detailed descriptions and also with few in-context examples. In the second step, we again prompt the same GPT-3.5-turbo model to generate short natural phrases by combining each object with its attribute as shown in Figure 3. These phrases are later used as prompts to Grounding DINO [21] that localizes the corresponding phrases in the image. For SVO probes [5] and ComVG datasets [8], which are used to test the image-text matching, the image captions have simpler triplet structure of the form subject-relationship-object. For more complex captions as encountered in Flickr30K [37] and MS-COCO [19] captions, GPT-3.5-turbo returns K phrases corresponding to K entities present in the captions and L relations.

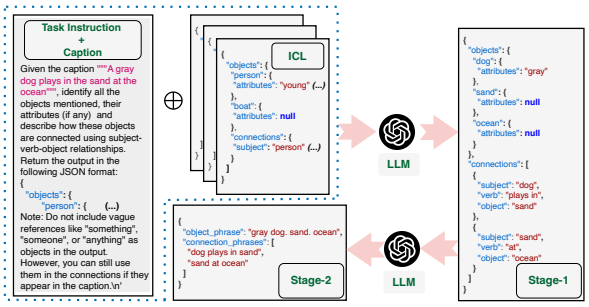


Figure 3. Two-stage decomposition of the caption "A gray dog plays in the sand at the ocean" using GPT-3.5. In Stage 1, the model identifies objects, attributes, and their relations in structured JSON format. In Stage 2, it generates short natural phrases combining attributes and objects (e.g., "gray dog", "sand", "ocean") and extracts subject-verb-object connections.

Grounding DINO is a zero-shot open-set object detection model that integrates DINO (Distillation with No Labels) visual transformer (ViT) trained with self-supervised with language grounding capabilities. The

Algorithm 1 Compositional Image-Text Alignment

Require: Caption C, Image I, LLM $\mathcal{M}_L$, Object Detector $\mathcal{M}_O$, Vision-Language Model $\mathcal{M}_V$
Ensure: Aligned multimodal embeddings

1. Text Decomposition
Query $\mathcal{M}_L$:
$O, A, R \leftarrow \mathcal{M}_L(C)$
$O = \{o_1, o_2, ..., o_n\}$ ▷ Objects
$A = \{a_1, a_2, ..., a_n\}$ ▷ Attributes
$R = \{(s_1, r_1, o_1), ..., (s_m, r_m, o_m)\}$ ▷ Relations
2. Phrase Generation
$P \leftarrow \{(a_i, o_i) \mid a_i \in A, o_i \in O\}$
$T \leftarrow \{(s_j, r_j, o_j) \mid (s_j, r_j, o_j) \in R\}$
3. Object Localization
for $p_i \in P$ **do**
  $L_i \leftarrow \mathcal{M}_O(p_i, I)$ ▷ Detect and localize object
end for
4. Alignment Computation
Compute global image embedding: $\phi_I \leftarrow \mathcal{M}_V(I)$
for $t_j \in T$ **do**
  Extract textual embedding: $\phi_T \leftarrow \mathcal{M}_V(t_j)$
  Initialize weighted image embedding: $\phi'_I \leftarrow \phi_I$
  for $p_i \in P$ **do**
    Sub-image embedding: $\phi_{L_i} \leftarrow \mathcal{M}_V(L_i)$
    Alignment score: $s_i = \cos(\phi_{L_i}, \phi_T)$
    Update embedding: $\phi'_I \leftarrow \phi'_I + s_i \cdot \phi_{L_i}$
  end for
  Compute final similarity: $\mathcal{S}_j = \cos(\phi'_I, \phi_T)$
end for
return $\{\mathcal{S}_j \mid t_j \in T\}$

model can detect objects based on arbitrary text queries. It uses contrastive training techniques and cross-modal matching to align textual descriptions with visual features, leading to improved detection of objects, their attributes and referring expressions. It is trained on massive datasets to ensure generalization across various domains (Figure 4 and 5).

Sub-image generation. Given the subject and object noun phrases and their associated bounding boxes localized by Grounding DINO $b_i = (x_i, y_i, w_i, h_i)$ and $b_j = (x_j, y_j, w_j, h_j)$, we mask out the remainder of the image and compute the associated CLIP embeddings $\mathbf{e}_{subj}$ and $\mathbf{e}_{obj}$. In case of more complex captions that involve K entities, we prompt Grounding DINO with K phrases returned by GPT-3.5-turbo, obtaining K bounding boxes and their associated CLIP sub-image embeddings $\boldsymbol{e}_k$. The sub-images associated with relations are determined by bounds of bounding boxes of entities associated with relations. See Figure 6 for more detail.

Similarity Computation. We use the normalized form of cosine similarity between constituent sub-images and phrase embeddings $\boldsymbol{e}_i$ and $\boldsymbol{t}_j$, i.e.,

$$\mathbf{sim}(\boldsymbol{e}_i, \boldsymbol{t}_j) = \frac{\boldsymbol{e}_i^T \boldsymbol{t}_j}{\|\boldsymbol{e}_i\| \|\boldsymbol{t}_j\|} \tag{1}$$

For ComVG and SVO image-text matching benchmarks, the image captions are made of triples $(subject, relation, object)$. Corresponding similarity scores with textual embeddings of the associated noun phrases $\mathbf{t}_{subj}$ and $\mathbf{t}_{obj}$ are obtained as:

$$s_{obj} = \text{sim}(\boldsymbol{e}_{obj}, \boldsymbol{t}_{obj}) \text{ and } s_{subj} = \text{sim}(\boldsymbol{e}_{subj}, \mathbf{t}_{subj}) \tag{2}$$

For relation grounding, we compute embedding of sub-image bounded by the extreme coordinates of b_{obj} and b_{subj} and compute its similarity score with the relation, i.e.,

$$s_{rel} = \text{sim}(\boldsymbol{e}_{rel}, \boldsymbol{t}_{rel}) \tag{3}$$

The dynamically adjusted embedding of the entire image $\boldsymbol{I}_c$ will be:

$$\boldsymbol{I}_c = \boldsymbol{e}_I + s_{obj}\boldsymbol{e}_{obj} + s_{subj}\boldsymbol{e}_{sub} + s_{rel}\boldsymbol{e}_{rel} \tag{4}$$



Figure 4. Examples of bounding boxes returned by Grounding DINO, when prompted by `man` and `sign` on the left and `man` and `umbrella` on the right

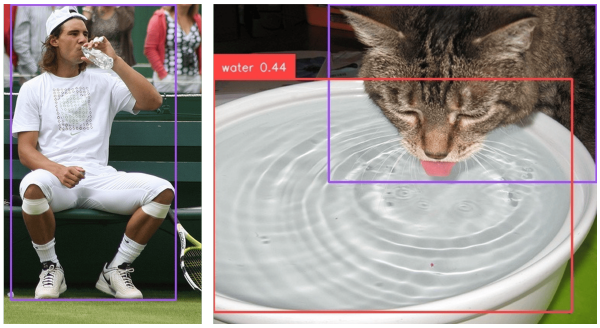


Figure 5. Examples of bounding boxes returned by Grounding DINO, when prompted by `man`, and `water` on the left and `cat` and `water` on the right.

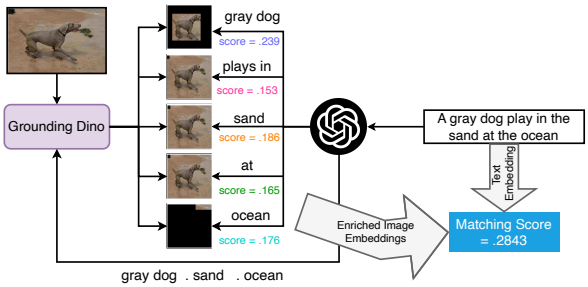


Figure 6. Illustration of sub-image generation and compositional fusion in GCLIP. Given the caption from Flickr, "A gray dog plays in the sand at the ocean," individual noun phrases such as "gray dog," "sand," and "ocean" are grounded using GDINO, and corresponding sub-images are extracted. Each sub-image is embedded with CLIP and weighted by its similarity to the associated phrase. These embeddings are combined with the global image embedding to form an enriched representation that improves alignment with the caption.

where $\boldsymbol{e}_I$ is the baseline CLIP embedding of the entire image. In case of more complex captions that involve k entities corresponding to k noun phrases returned by LLM and associated similarity scores $s_j = \text{sim}(\boldsymbol{e}_j, \boldsymbol{t}_j)$ for all $j = 1 \ldots k$, the new image embedding is given as

$$\boldsymbol{I}_c = \boldsymbol{e}_I + \sum_{k=1}^{K} s_k \boldsymbol{e}_k \tag{5}$$

The final similarity score will be $\text{sim}(\boldsymbol{I}_c, \boldsymbol{T})$, where $\boldsymbol{T}$ is the text embedding of the entire caption.

We found the score s_k computed using normalized embeddings to be more effective than unnormalized scores followed by `Softmax` normalization used in Jiang et al. [8] that assumes the scores are independent. In summary, the overall process entails breaking down the input image into entity-specific sub-images, computing similarity scores between these sub-images and their corresponding textual representations, and then combining these weighted embeddings with the global image embedding.

CompCLIP [8] achieves compositional matching through a two stage process, where a dense captioning model (GRiT) [34] generates region descriptions, and then a language model (GPT-3.5) filters the captions that best corresponds to the entities in the image. While this approach improves compositionality, it raises two key challenges. First, relying on automated captioning and text heuristics introduces potential errors, due to incomplete or misleading captions that affect the alignment. Second, dense captioning is prone to generating numerous captions for a single subject/object, creating redundant sub-images that increase computational cost and induce inconsistency in similarity scoring. Such challenges make ComCLIP computationally expensive and less interpretable, thus limiting its efficiency in real-world scenarios where precise and effective compositional reasoning is critical.

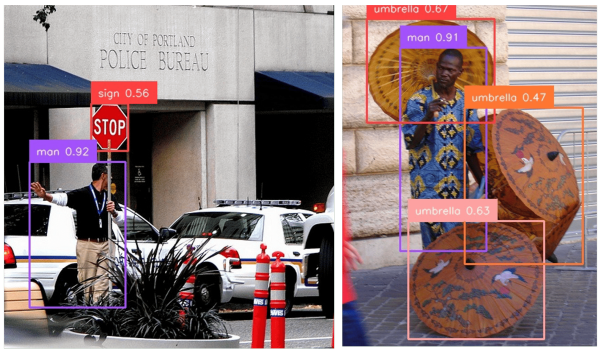


Figure 7. Positive and negative image for caption `a man is holding a sign`. Positive and negative scores for GCLIP are 0.2052/0.2069 ✗and ComCLIP are 0.2140/0.2080 ✓

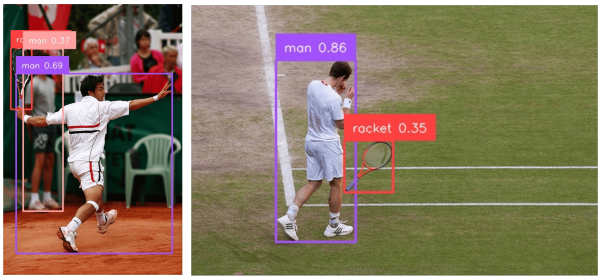


Figure 8. Positive and negative image for caption `a man is swinging a racket`, negative image differs in verb; Positive/negative scores for GCLIP are 0.2260/0.2101 ✓and ComCLIP are 0.2153/0.2182. ✗

4. Experiments

To analyze fine grained compositional image embeddings, we evaluate our approach on selected image-text matching and retrieval benchmarks.

Compositional Visual Genome (ComVG) [8] dataset is derived from Visual Genome [12] and focuses on subject-verb-object (SVO) triplets. This dataset comprises of 524 distinct images, each varying by either the subject, verb, or object. Across these images, there are 5400 relations capturing fine grained variations in SVO combinations. This dataset focuses on understanding structured triplets rather than natural captions. We evaluate our approach on the entire dataset.

SVO Probes [5] dataset is constructed from Conceptual Captions [29] and is designed to evaluate a model's

sensitivity to changes in the compositional structure of subject-verb-object triplets. While the dataset contains 36,842 samples, we evaluate our approach on a filtered subset of 13,767 samples, as many of the original images are either unreadable or no longer accessible.

Flickr30K [37] is a widely used image-text retrieval benchmark containing 31,783 images each annotated with 5 human written captions. The captions are natural, free form descriptions that capture objects, actions and interactions in the scene. Following the evaluation protocol in [8], we randomly select one caption for the test set that has 1000 images. First, CLIP [26] is used to retrieve the top 10 similar images for each caption. For these top ranked candidates, we apply our Grounding CLIP (GCLIP) approach.

MS-COCO [19] is another large benchmark for image-text retrieval with over 120,000 images paired with five human written captions. Similar to Flickr30K, the captions in MS-COCO describe objects, actions and spatial relationships in the scene using natural language. We follow the same evaluation protocol similar to Flickr30K. Out of 40,000 images from the validation set, we randomly select 1000 images and for each image we select one random caption out of the five captions. CLIP is used to retrieve the top 10 similar images for each caption from this validation set. Then we apply Grounding CLIP (GCLIP) on the top 10 candidates.

ComVG and SVO probes are used to evaluate image-text matching as they consist of structured subject-verb-object triplets. Flickr30K and MS-COCO are used to assess image-to-text retrieval capabilities based on natural language descriptions.

Vision Encoder	CLIP	OpenCLIP	ComCLIP*	$\text{GCLIP}_{\text{ours}}$
ResNet-50	82.25	82.21	83.26	**84.79**
ViT-B-32	82.45	82.41	83.12	**85.61**
ViT-L-14	86.38	86.38	86.93	**87.85**

Table 1. Image-text matching accuracy % of different models on the Compositional Visual Genome dataset.(**The numbers are reproduced by us.*)

Vision Encoder	CLIP	OpenCLIP	ComCLIP*	$\text{GCLIP}_{\text{ours}}$
ResNet-50	81.21	81.56	82.06	**82.67**
ViT-B-32	81.96	80.67	82.68	**83.88**
ViT-L-14	84.79	85.23	85.32	**86.29**

Table 2. Image-text matching accuracy % of different models on the SVO-Probes dataset. (**The numbers are reproduced by us.*)

4.1. Baselines

We compare our approach against several strong vision-language baselines to evaluate improvements in both compositional understanding and retrieval performance.

CLIP [26]: CLIP is a widely used vision language model trained with contrastive learning on 400M image-text pairs. It is our primary baseline for both image-text matching and retrieval tasks.

OpenCLIP [6]: An open source implementation of CLIP with similar architecture and training objectives. We include it to assess consistency across CLIP variants.

ComCLIP [8]: A training free method that enhances CLIP for compositional reasoning by leveraging dense captioning and localization. This is the most closely related prior work and serves as a key comparison for compositional image-text matching.

SLIP [24] and ComSLIP [8]: SLIP extends contrastive vision-language training with self-supervised objectives. ComSLIP integrates compositional information into SLIP embeddings similar to ComCLIP.

BLIP [15] and BLIP2 [16]: These models improve multimodal alignment and reasoning by combining Vision Encoder with Large Language Models. We also compare with their compositionally enhanced variants ComBLIP and ComBLIP2.

4.2. Image-Text Matching

In image-text matching, for each triplet, we localize the subject and object nouns in the image using Grounding DINO. Instead of cropping these regions, we create masked sub-images by blacking out all areas of the image except the detected bounding box for the target entity. This ensures that each sub-image has same dimension as the original image while separating the visual signal associated with a specific entity. To create a relation sub-image, we combine the related subject and object sub-images. This combined sub-image effectively captures the context and interaction between the subject and object.

The original image and its corresponding caption along with the masked sub-images and their associated words (subject, object, relation) are processed through the CLIP's vision and text encoders. We then calculate the cosine similarity between each sub-image embedding and its respective word embedding. These similarity scores are used as weights to compute a weighted sum of the sub-image embeddings. This weighted sum is added to the global embedding of the original image to

produce an adjusted image representation $\boldsymbol{I}_c$. Finally, we compute the cosine similarity between $\boldsymbol{I}_c$ and the embedding of the entire caption. The methodology remains same for the other baselines as well, termed as GSLIP, GBLIP and GBLIP2.

Method	Sub	Rel	Obj	Total
SLIP	86.2	61.33	85.81	80.13
ComSLIP*	86.11	**63.05**	89.51	82.27
GSLIP$_{\text{ours}}$	**87.17**	62.97	**89.67**	**82.63**
CLIP	88.61	68.52	93.85	86.38
ComCLIP*	89.84	**69.92**	93.61	86.93
GCLIP$_{\text{ours}}$	**92.38**	69.45	**94.27**	**87.85**
BLIP2	96.48	83.52	95.63	93.00
ComBLIP2*	96.28	**82.81**	95.39	92.66
GBLIP2$_{\text{ours}}$	**97.14**	82.73	**95.94**	**93.15**

Table 3. Image-text matching accuracy of different models and approaches on the ComVG dataset. (**The numbers are reproduced by us.*)

Evaluation Metric: We use accuracy as the evaluation metric for both the ComVG and SVO-Probes. Each test example consists of a caption paired with one positive image and one negative image. A prediction is considered correct if the similarity score between the caption and the positive image is higher than that of the negative image. The final accuracy is computed as the percentage of correct predictions across the dataset. In addition to overall accuracy, we report accuracy separately based on which component (subject, predicate, or object) differs between the two images.

4.2.1. Results on ComVG and SVO-Probes

Table 1 and 2 present image-text matching accuracy on ComVG and SVO Probes across different vision encoder backbones: ResNet-50, ViT-B/32, and ViT-L/14. Across all baselines: CLIP, OpenCLIP, ComCLIP and our proposed GCLIP, we observe that ViT-L/14 consistently achieves the highest accuracy, demonstrating stronger visual representations for compositional tasks. Based on this observation, we use ViT-L/14 as the default vision encoder in all subsequent evaluations.

Table 3 and 4 provide a more detailed breakdown of performance across compositional categories: subject, relation, object. These results offer direct evidence in support of our central claim that explicit grounding of entities enables more accurate visual-textual alignment. Note that CLIP embeddings continue to struggle with verb disambiguation in **Rel** column and BLIP2 does notably better in all columns suggesting the effectiveness of an additional training stage when compared to CLIP. Examples of correct and incorrect matches can be found in Figures 7, and 8.

Method	Sub	Rel	Obj	Total
SLIP	81.71	72.80	89.14	77.59
ComSLIP*	82.73	73.39	90.38	78.39
GSLIP$_{\text{ours}}$	**86.54**	**73.76**	**91.90**	**80.09**
CLIP	87.46	82.31	90.27	84.79
ComCLIP*	88.34	82.74	90.90	85.32
GCLIP$_{\text{ours}}$	**89.54**	**83.27**	**91.75**	**86.29**
BLIP	83.13	82.54	89.14	83.99
ComBLIP*	85.09	83.63	90.41	85.25
GBLIP$_{\text{ours}}$	**85.49**	**84.21**	**90.52**	**85.85**
BLIP2	92.75	89.46	96.27	91.39
ComBLIP2*	92.97	89.18	96.70	91.33
GBLIP2$_{\text{ours}}$	**93.78**	**89.87**	**97.36**	**92.24**

Table 4. Image-text matching accuracy of different models and approaches on the SVO-Probes dataset.(**The numbers are reproduced by us.*)

In Table 3 on the ComVG dataset, GCLIP consistently improves performance across all components. In the subject category, GCLIP outperformed ComCLIP (92.38% vs. 89.84%) and CLIP (92.38% vs. 88.61%). This is crucial because, dataset biases often affect subject recognition and models struggle when the same objects appear in different positions across training data. By localizing and comparing grounded embeddings, our method more reliably distinguishes between semantically distinct configurations. The object category follows a similar trend, where GroundingCLIP reaches 95.94%, surpassing all prior methods. This suggests that grounding improves not only fine-grained subject recognition but also reinforces object-level distinctions in scenes with visually similar elements.

In Table 4, on the SVO-Probes dataset, which features more diverse and unconstrained image-text pairs, GCLIP outperformed both CLIP and ComCLIP across all categories. This suggests that the benefits of grounding are not just limited to structured or synthetic datasets, but also generalize well to real-world compositions and small variations in phrasing and visual layout.

ComCLIP's minor advantage on ComVG relation accuracy likely comes from GRiT captioning model which is partly being trained on Visual Genome dataset. SVO Probes images do not come from Visual Genome dataset and the verbs are more ambiguous. Hence, GCLIP's direct box-based relation grounding provides a more robust and accurate signal, resulting in better performance.

4.3. Image Text Retrieval

To evaluate the generalization ability of GCLIP beyond structured triplets, we have conducted retrieval experiments on Flickr30K and MS-COCO, two benchmarks featuring natural, free form human written captions.

Our retrieval pipeline begins by prompting GPT-3.5 to analyze the input caption in two stages:

- In the first stage, we prompt the model to identify all objects mentioned in the caption, extract their associated attributes if any (eg: "red shirt") and describe how these objects are connected through subject-verb-object relations.
- In the second stage, we again prompt the same model to combine the objects with their attributes and generate short natural phrases. These noun phrases are then passed to GroundingDINO to detect and localize the corresponding visual regions in the image.

For each caption, we apply this process to the top 10 retrieved images. Each image is treated as an individual image-text matching instance, where we use the localized bounding boxes and compositional fusion (Section 4.2) to compute a refined Image Text Matching score (the cosine similarity between the enriched image embedding and the full caption embedding.). These scores are then used to re-rank the 10 candidates, improving the precision of final retrieval.

An added advantage of our approach is the use of global image embedding in combination with the grounded region embeddings. The global embedding offers contextual information that may not be captured by a single region, while the grounded sub images allow the model to focus on specific entities and relations mentioned in the caption. This is very crucial in retrieval where the free form captions can contain overall scene details such as background or scene context that span beyond individual objects. Using the weighted sum of grounded region embeddings and the global image embedding, our method achieves a balance between scene-level semantics and fine-grained compositional alignment.

The results on Flickr30K and MS-COCO benchmarks from Figure 9 reveal significant improvement introduced by GCLIP and how it aligns images with free form captions. Unlike traditional models that solely rely on global image representation, GCLIP integrates localized entity specific visual details guided by textual phrases enabling it to distinguish between visually similar scenes based on fine grained compositional details. A 12% boost in Recall@1 on Flickr30K indicates a meaningful shift in retrieval behavior. These results underscore that retrieval performance can be significantly enhanced by redefining the image representation through grounding, without requiring any additional training.

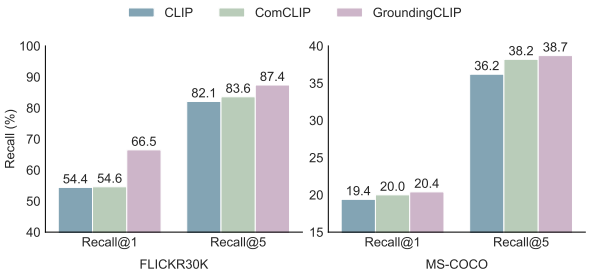


Figure 9. Comparison of Recall@1 (%) and Recall@5 on Flickr30K and MS-COCO datasets

5. Conclusions

We proposed Grounding CLIP (GCLIP), a training-free method for enhancing compositional image-text alignment by dynamically refining global image embeddings using grounded sub-regions detected via an open-vocabulary detector. Our method consistently outperformed CLIP and ComCLIP across both structured (ComVG, SVO Probes) and natural (Flickr30K, MS-COCO) benchmarks, achieving significant improvements in image-text matching and retrieval tasks. The results highlight the importance of grounding individual entities and relations in improving fine-grained visual-textual alignment. In our Future work, we plan to explore additional benchmarks, models and fine-tuning strategies to further improve the model's ability to distinguish between different verbs and relations and consider scalability aspects on large scale retrieval datasets.

6. Limitations

While GCLIP achieves strong performance without additional training, it has a few limitations. First, the reliance on large language models and open-vocabulary detectors which introduces additional inference time overhead, which can impact scalability in real-time applications. Second, the quality of the final image embedding is sensitive to the accuracy of entity grounding; failure to detect or correctly localize entities can negatively affect similarity computation. Finally, although GCLIP significantly improves alignment for subjects and objects, the performance on verb (relation) disambiguation still lags behind, likely due to the complexity involved in identifying actions and interactions that span across multiple visual regions.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur

Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[2] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[4] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024. 1

[5] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, 2021. 1, 2, 3, 5

[6] 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. Openclip, 2021. 6

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[8] Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric Wang. Comclip: Training-free compositional image and text matching. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. 1, 2, 3, 5, 6

[9] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[10] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, 2023. 2

[11] 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*, 123:32–73, 2017. 1

[12] 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*, 123:32–73, 2017. 5

[13] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500, St. Julian's, Malta, 2024. Association for Computational Linguistics. 2

[14] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 6

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2, 6

[17] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[18] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 3, 6

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2, 3

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representa-

tions for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[23] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, page 728–755, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[24] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022. 6

[25] 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 (ICCV)*, 2015. 1

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 6

[27] Navid Rajabi and Jana Kosecka. Towards grounded visual spatial reasoning in multi-modal vision language models. *CoRR*, abs/2308.09778, 2023. 1, 2

[28] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 5

[30] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[31] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5198–5215, 2022. 2

[32] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 2

[33] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[34] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224. Springer, 2024. 2, 5

[35] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 2

[36] Ziyan Yang, Kushal Kafle, Franck Dernoncourt, and Vicente Ordonez. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19165–19174, 2023. 1, 2

[37] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014. 3, 6

[38] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*. 2

[39] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 2