Structure-CLIP: Towards Scene Graph Knowledge to Enhance Multi-Modal Structured Representations

Yufeng Huang[1] *, Jiji Tang[2] *, Zhuo Chen[3], Rongsheng Zhang[2,3], Xinfeng Zhang[2], Weijie Chen[2], Zeng Zhao[2], Zhou Zhao[3], Tangjie Lv[2], Zhipeng Hu[2], Wen Zhang[1] †

[1]School of Software Technology, Zhejiang University
[2]Fuxi AI Lab, Netease Inc.
[3]College of Computer Science and Technology, Zhejiang University
{huangyufeng, zhuo.chen, zhaozhou, zhang.wen}@zju.edu.cn
{tangjiji01, zhangrongsheng, zhangxinfeng01, chenweijie05, hzlvtangjie, zphu}@corp.netease.com

Abstract

Large-scale vision-language pre-training has achieved significant performance in multi-modal understanding and generation tasks. However, existing methods often perform poorly on image-text matching tasks that require structured representations, i.e., representations of objects, attributes, and relations. The models cannot make a distinction between "An astronaut rides a horse" and "A horse rides an astronaut". This is because they fail to fully leverage structured knowledge when learning multi-modal representations. In this paper, we present an end-to-end framework Structure-CLIP, which integrates *Scene Graph Knowledge* (SGK) to enhance multi-modal structured representations. Firstly, we use scene graphs to guide the construction of *semantic negative* examples, which results in an increased emphasis on learning structured representations. Moreover, a *Knowledge-Enhance Encoder* (KEE) is proposed to leverage SGK as input to further enhance structured representations. To verify the effectiveness of the proposed framework, we pre-train our model with the aforementioned approaches and conduct experiments on downstream tasks. Experimental results demonstrate that Structure-CLIP achieves *state-of-the-art* (SOTA) performance on VG-Attribution and VG-Relation datasets, with 12.5% and 4.1% ahead of the multi-modal SOTA model respectively. Meanwhile, the results on MSCOCO indicate that Structure-CLIP significantly enhances the structured representations while maintaining the ability of general representations. Our code is available at https://github.com/zjukg/Structure-CLIP.

Introduction

Vision-language models (VLMs) have demonstrated significant performance in various multi-modal understanding and generation tasks (Radford et al. 2021; Li et al. 2022; Singh et al. 2022; Li et al. 2019). Despite the impressive performance of multi-modal models in various tasks, the question of whether these models can effectively capture structured knowledge(i.e., the ability to comprehend object properties and the relationships between objects) remains unresolved.

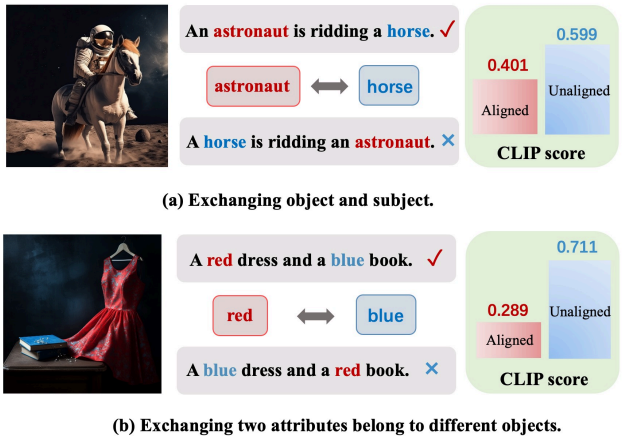


Figure 1: CLIP scores (after normalizing among two results) between the image and aligned/unaligned captions. The results show that the CLIP model does not have the ability to distinguish sentences with structured semantic differences.

For example, as shown in Fig. 1 (a), the CLIP score (i.e., semantic similarity) between the image and the correctly matched caption ("An astronaut is riding a horse"), exhibits a lower value in contrast to the score between the image and a non-matching caption ("A horse is riding an astronaut"). Subsequently, Fig. 1 (b) illustrates that exchanging attributes between two objects can also pose challenges for the model to accurately distinguish their semantics. These findings suggest that the generic representations yielded by the CLIP model are unable to differentiate between text segments that encompass identical words but diverge in terms of structured knowledge. In other words, the CLIP model exhibits a tendency similar to that of a bag-of-words approach, which does not understand fine-grained semantics in sentences (Lin et al. 2023).

Winoground (Thrush et al. 2022) is the first work focusing on this problem and performing a broad-based examination. They intentionally created a dataset consisting of 400 instances, where each instance consists of two sentences with identical word compositions but different semantic mean-

*These authors contributed equally.
†Corresponding Author.

ings. They evaluated various well-performing VLMs (e.g., VinVL (Zhang et al. 2021), UNITER (Chen et al. 2020), ViLBERT (Lu et al. 2019), and CLIP (Radford et al. 2021)), intending to assess the structured representations about objects, attributes, and relations. Unfortunately, their findings indicate that the outcomes are on par with a random selection, despite these models demonstrating human-level proficiency in other tasks. The results of these tasks demonstrate that general representations are insufficient for semantic comprehension. It is thus inferred that an increased emphasis should be placed on structured representations.

NegCLIP (Yüksekgönül et al. 2022) enhances structured representations by integrating task-specific negative samples, which are generated by randomly exchanging any two words in a sentence. Thus, while general representations maintain consistency in positive and negative samples, structured representations exhibit divergence. Employing the contrastive learning approach compels the model to acquire structured representations rather than general representations. Moreover, NegCLIP also provides a large-scale test bed to evaluate the capabilities of VLMs in terms of structured representations. Nevertheless, NegCLIP suffers from a lack of understanding and modeling of the semantic knowledge during negative sample construction, which results in a notable deterioration in the quality of negative examples. For example, when the attributes "white" and "black" are interchanged in the original caption "Black and white cows", the underlying semantic meaning of the sentence remains invariant. Such low-quality negative examples further lead to performance degradation.

In this paper, we propose Structure-CLIP, a novel approach that leverages *Scene Graph Knowledge* (SGK) to enhance multi-modal structured representations. Firstly, in contrast to the random swap method in NegCLIP, we utilize SGK to construct word swaps that better match the underlying intent. Secondly, we propose a *Knowledge-Enhanced Encoder* (KEE), leveraging SGK to extract essential structure information. By incorporating structured knowledge at the input level, the proposed KEE can further enhance the ability of structured representations. Results on Visual Genome Relation and Visual Genome Attribution show the *state-of-the-art* (SOTA) performance of Structure-CLIP and the effectiveness of its components. Additionally, we perform cross-modal retrieval evaluations on MSCOCO, which shows that Structure-CLIP still retains sufficient general representation ability.

Overall, our contributions are three-fold:

- To the best of our knowledge, Structure-CLIP is the first method to enhance detailed structured representations by constructing *Semantic Negative* samples.
- A *Knowledge-Enhanced Encoder* is introduced in Structure-CLIP to leverage the structured knowledge as the input to enhance structured representations.
- We conduct comprehensive experiments demonstrating that Structure-CLIP is able to achieve SOTA performance on structured representations downstream tasks and yield significant improvements on structured representations.

Related Work

Vision Language Pretraining

Vision-Language Models (VLMs) aim to learn universal cross-modal representations, which are beneficial for achieving strong performance in downstream multi-modal tasks. Depending on the multi-modal downstream task, different model architectures have been developed, including the dual-encoder architecture (Radford et al. 2021; Jia et al. 2021), the fusion-encoder architecture (Tan and Bansal 2019; Li et al. 2021a), the encoder-decoder architecture (Cho et al. 2021; Wang et al. 2022c; Chen et al. 2022), and more recently, the unified transformer architecture (Li et al. 2022; Wang et al. 2022a).

The pre-training tasks have a great impact on what VLMs can learn from the data. There are mainly 4 types of tasks: *(i)* Cross-Modal Masked Language Modeling (MLM) (Kim, Son, and Kim 2021; Lin et al. 2020; Li et al. 2021a; Yu et al. 2022); *(ii)* Cross-Modal Masked Region Prediction (MRP) (Lu et al. 2019; Chen et al. 2020; Huang et al. 2021); *(iii)* Image-Text Matching (ITM) (Li et al. 2020; Lu et al. 2019; Chen et al. 2020; Huang et al. 2021); *(iv)* Cross-Modal Contrastive Learning (CMCL) (Radford et al. 2021; Jia et al. 2021; Li et al. 2021a; Huo et al. 2021; Li et al. 2021b).

Recent research mainly focuses on the study of CMCL. Taking the CLIP model (Radford et al. 2021) as an example, the model learned sufficient general representations by comparing positive examples with negative examples from all other samples in the dataset.

Structured Representation Learning

Structured Representations denote the ability to match images and texts that have identical word compositions. Winoground (Thrush et al. 2022) first presented a novel task and dataset for evaluating the ability of VLMs. The dataset comprises primarily 400 hand-crafted instances, where each instance includes two sentences with similar word compositions but distinct semantics, along with corresponding images. The evaluation results of Winoground identified the dataset's main challenges through a suite of experiments on related tasks (i.e., probing task, image retrieval task), suggesting that the main challenge in vision-language models may lie in fusing visual and textual representations, rather than in the understanding of compositional language.

Due to the limited quantity of Winoground test data, it is challenging to draw dependable experimental results on the ability of structure representations. Recently, NegCLIP (Yüksekgönül et al. 2022) provided a large-scale test bed to evaluate structured representations of VLMs. Additionally, NegCLIP also proposes a negative sampling method to enhance structured representations.

Scene Graph Generation

A scene graph is a type of structured knowledge, which describes the most essential parts of a multi-modal sample, through modeling objects, attributes of objects, and relations between objects and subjects. Generally, *Scene Graph Generation*(SGG) models consist of three main modules: proposal generation localizing the bounding box of objects, ob-

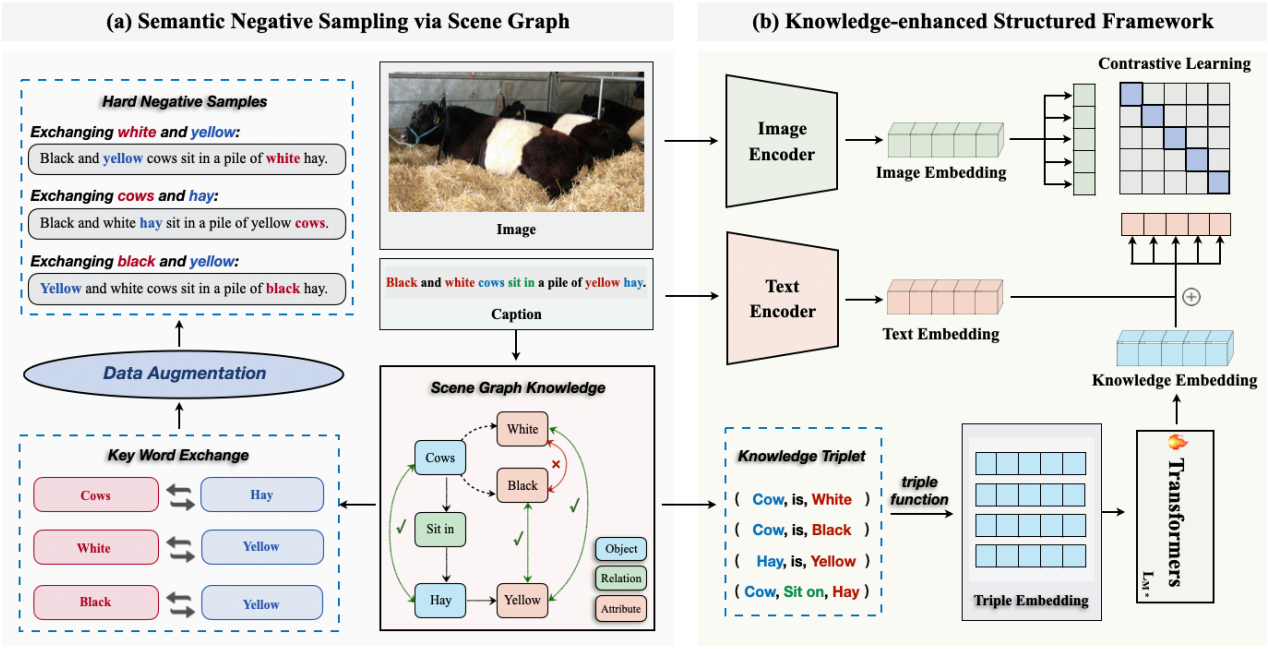


Figure 2: Overview of Structure-CLIP. (a) *Semantic negative sampling via scene graph*: we extract a scene graph from the caption to help construct high-quality negative samples(left part). (b)*Knowledge-Enhanced Encoder*: Knowledge embedding module and multiple Transformers layers are used to model structured knowledge at the input level(right part).

ject classification labeling the detected objects, and relationship prediction predicting the relations between pairwise objects. Some existing works (Xu et al. 2017; Yang et al. 2018; Zellers et al. 2018) applied RNNs and GCNs to propagate image contexts in order to achieve better utilizing the contexts for object and relationship prediction. VCTree (Tang et al. 2019) captured local and global visual contexts by exploiting dynamic tree structures. Gu et al. (2019) and Chen et al. (2019) integrated external knowledge into SGG models to address the bias of noisy annotations.

As a beneficial prior knowledge describing the detailed semantics of images and captions, scene graphs have helped achieve excellent performance in several vision-language tasks. Such as image captioning (Yang et al. 2019), image retrieval (Wu et al. 2019a), visual question answering (Zhang, Chao, and Xuan 2019; Wang et al. 2022b), multimodal sentiment classifications (Huang et al. 2022), image generation (Johnson, Gupta, and Fei-Fei 2018) and vision-language pretraining (Yu et al. 2021).

Methodology

The overview of Structure-CLIP is illustrated in Fig. 2. Firstly, our approach leverages the scene graph to enhance fine-grained structured representations by generating semantic negative samples with identical word compositions but differing detailed semantics(*left part of Fig. 2*). Secondly, we propose a Knowledge-Enhanced Encoder that utilizes the scene graph as an input to integrate structured knowledge into the structured representations (*right part of Fig. 2*). We will introduce semantic negative sampling via the scene graph in Section 3.1 and present the Knowledge-Enhanced Encoder in Section 3.2.

Semantic Negative Sampling via Scene Graph

Faghri et al. (2018) proposed a negative sampling method that involves constructing negative examples to enhance the representations by comparing them with positive samples. Our objective is to construct samples with similar general representations but differing detailed semantics, thereby encouraging the model to focus on learning structured representations.

Scene Graph Generation. Detailed semantics, including objects, attributes of objects, and relationships between objects, are essential to the understanding of visual scenes. And they are critical to cross-modal learning, which aims to enhance the joint representation of vision and language. In our framework, the Scene Graph Parser provided by (Wu et al. 2019b) is adopted to parse texts to scene graphs. Given the text sentence $\mathbf{w}$, we parse it into a scene graph (Johnson et al. 2015), which denotes as $G(\mathbf{w}) =< O(\mathbf{w}), E(\mathbf{w}), K(\mathbf{w}) >$, where $O(\mathbf{w})$ is the set of objects mentioned in $\mathbf{w}$, $R(\mathbf{w})$ is the set of relationship nodes, and $E(\mathbf{w}) \subseteq O(\mathbf{w}) \times R(\mathbf{w}) \times O(\mathbf{w})$ is the set of hyper-edges representing actual relationships between objects. $K(\mathbf{w}) \subseteq O(\mathbf{w}) \times A(\mathbf{w})$ is the set of attribute pairs, where $A(\mathbf{w})$ is the set of attribute nodes associated with objects.

As shown in Figure 2, we generate the scene graph based on the original caption. Using the caption "Black and white cows sit in a pile of yellow hay" in Fig. 2 as an example, in the generated scene graph, the objects, such as "cows"

and "hay" are the fundamental elements. The associated attributes, such as "white" and "yellow" characterize the color or other attributes of objects. Relations such as "sit in" represent the spatial connections between objects.

Choice of Semantic Negative Samples. Contrastive learning aims to learn effective representations by pulling semantically close neighbors together and pushing apart non-neighbors. Our objective is to construct semantic negative samples with similar composition but different detailed semantics. Therefore, the quality of negative samples plays a vital role in structured representation learning.

A multi-modal dataset usually consists of N image-text pairs, where image and text are denoted as I and W with subscripts, respectively. Given an image-text pair (I_i, W_i) and a related scene graph $G(W_i)$ generated from W_i, a high-quality semantic negative sample W_i^- is generated via

$$W_i^- = F(W_i, G(W_i)), \tag{1}$$

where F is the proposed sampling function, W_i^- denotes the high-quality semantic negative sample. Specifically, for triples $(object, relation, subject)$ in the scene graph, W_i^- is generated via

$$W_i^- = Swap((O_1, R, O_2)) = (O_2, R, O_1), \tag{2}$$

where $Swap$ is the function to exchange object and subject in the sentence, O_1, R, O_2 denote the object, relation and subject. For attribute pairs $(A1, O1)$ and $(A2, O2)$ in the scene graph, W_i^- is generated via

$$Swap((A_1O_1), (A_2O_2)) = \begin{cases} (A_2O_1), (A_1O_2) & if\ O_1 \neq O_2, \\ pass & if\ O_1 = O_2, \end{cases} \tag{3}$$

Overall, we leverage scene graph guidance to construct high-quality semantic negative samples, instead of randomly swapping word positions. Our semantic negatives maintain the same sentence composition while altering detailed semantics. As a result, our model can more effectively learn structured representations of detailed semantics.

Contrastive Learning Objective. Our contrastive learning objective is to learn sufficient representations by pulling image I_i and origin caption W_i together and pushing apart image I_i and negative sample W_i^-. Specifically, we introduce a multi-modal contrastive learning module with the loss function:

$$\mathcal{L}_{hinge} = max(0, \gamma - d + d'), \tag{4}$$

where γ is the margin hyper-parameter, d denotes the distance between image I_i and origin caption W_i and d' denotes the distance between image I_i and origin caption W_i^-. The contrastive learning objective is introduced to improve the performance of structured representations. Meanwhile, in order to maintain the general representation ability of the model, we combine the original mini-batch image-text contrastive learning loss and the proposed loss for joint training.

The original image-text contrastive learning loss $\mathcal{L}_{ITCL}$ contains an image-to-text constrastive loss $\mathcal{L}_{i2t}$ and a text-to-image contrastive loss $\mathcal{L}_{t2i}$ that

$$\mathcal{L}_{ITCL} = (\mathcal{L}_{i2t} + \mathcal{L}_{t2i})/2, \tag{5}$$

The image-to-text contrastive loss $\mathcal{L}_{i2t}$ is formulated as

$$\mathcal{L}_{i2t} = -\log \frac{\exp((\tilde{v}_i, e_{text_i})/\tau)}{\sum_{k=1}^{N} \exp((\tilde{v}_i, e_{text_k})/\tau)}, \tag{6}$$

where τ is the temperature hyper-parameter. Similarly, the text-to-image contrastive loss $\mathcal{L}_{t2i}$ is

$$\mathcal{L}_{t2i} = -\log \frac{\exp((e_{text_i}, \tilde{v}_i)/\tau)}{\sum_{k=1}^{N} \exp((e_{text_i}, \tilde{v}_k)/\tau)}, \tag{7}$$

Thus the final loss, which combines the hinge loss and InfoNCE loss, is

$$\mathcal{L}_{final} = \mathcal{L}_{hinge} + \mathcal{L}_{ITCL}. \tag{8}$$

Knowledge-Enhanced Encoder

In this section, we propose a Knowledge-Enhanced Encoder, which utilizes scene graphs as the textual input to enhance the structured representations. To begin with, we use the following function to encode image I_i and text W_i:

$$\tilde{v} = CLIP_{vis}(I_i), \tag{9}$$

$$\tilde{z} = CLIP_{text}(W_i), \tag{10}$$

where $CLIP_{vis}$ and $CLIP_{text}$ denote the visual encoder and text encoder of the CLIP model, respectively.

However, the CLIP model processes text input in a word-bag manner, which ignores the detailed semantics of the text. In contrast, incorporating a scene graph captures crucial structural information from the sentence, thereby enabling the model to gain deeper insights into the fine-grained semantics of the text.

Therefore, the Knowledge-Enhanced Encoder explicitly models the detailed knowledge as model input, i.e., objects, attributes of objects, and relations between paired objects. Specifically, we make a unified input specification for two structured knowledge: pairs and triples. We add the relationship conjunction "is" to the pair to unify the representations. For example, The pair $(white, cow)$ will be treated as the triple $(cow, is, white)$ in this manner. In this manner, a set of triples $\mathcal{T}_{in} = \{(h_i, r_i, t_i) | i \in [1, k]\}$ are obtained, where (h_i, r_i, t_i) represent the head entity, relation entity and tail entity respectively. For each triple (h_i, r_i, t_i) in $\mathcal{T}_{in}$, we use Tokenizer and Word Vocabulary Embeddings from BERT (Devlin et al. 2019) to obtain each entity embedding w_h, w_r, w_t:

$$w_x = WordEmb(x), x \in [h, r, t], \tag{11}$$

In order to get the triple embedding with each entity embedding, we use the following encoding function:

$$e_{triple_i} = ENC_{triple}(h_i, r_i, t_i) = w_{h,i} + w_{r,i} - w_{t,i}, \tag{12}$$

where $ENC_{triple}(.)$ is the triple encoding function. With this triple encoder, our method can better solve the problem that the order of the head and tail entities is reversed, a detailed analysis is illustrated in Sec. 4.4.3.

In this way, K triples can be processed into K semantic embeddings. Then we input e_{triple} to multiple Transformer layers to get the final representations.

$$e_{KE} = TRMs([e_{triple_1}, ..., e_{triple_K}]), \tag{13}$$

Domains	Models	Params	Visual Gnome		MSCOCO	
			Attribute	Relation	IR-R@1	TR-R@1
-	Random Chance	-	50.00	50.0	0.02	0.1
Multi-modal Models	VILT (VIT-B/32)	87 M	20.3	39.5	37.3	53.4
	FLAVA	241 M	58.1	28.0	38.5	43.5
	CLIP-Base (ViT-B/32)	151 M	60.1	59.8	30.4	50.1
	CLIP-Large (ViT-L/14)	427M	61.1	61.5	36.5	56.3
	Neg-CLIP	151 M	71.0	81.0	41.0	56.0
Large Language Models	BART	300 M	73.6	81.1	-	-
	FLAN-T5	11 B	76.5	84.4	-	-
	OPT	175 B	79.8	84.7	-	-
Ours	Sturcture-CLIP-Base	220 M	82.3	84.7	41.2	55.6
	Structure-CLIP-Large	496 M	**83.5**	**85.1**	**48.9**	**58.2**

Table 1: Results (%) comparison between our method and other baselines on the VG-Relation, VG-Attribution and MSCOCO datasets. The matching scores are obtained via semantics similarities between image embeddings and text embeddings in multi-modal models and Maximum Likelihood Probability in large language models, respectively.

The Knowledge-Enhanced Encoder enables us to extract sufficient structured knowledge from all input triples, which can be utilized as effective structured knowledge to improve the performance of structured representations.

Thus, the Knowledge-Enhanced Encoder can be utilized to obtain text knowledge embeddings. However, relying solely on structured knowledge may result in a loss of representing general semantics. Therefore, we integrate both text embeddings and structured knowledge embeddings :

$$\begin{aligned} e_{text} &= \tilde{z} + \lambda e_{KE} \\ &= CLIP_{text}(W_i) + \lambda \cdot TRMs([e_{triple_*}]), \end{aligned} \tag{14}$$

where λ is a hyper-parameter, $\tilde{z}$ and e_{KE} denote the origin text embedding and the structural knowledge embedding.

Our textual representations contain both the word information carried by the whole sentence and the structured knowledge composed of the detailed semantics in the sentence. Similarly, we used the same loss strategy illustrated in Eq. 5 in the training process.

Experiments

Datasets

Pretraining Datasets. High-quality image-text alignment data is a critical aspect of training models. We adopt the widely-used cross-modal text-image retrieval dataset, MSCOCO (Lin et al. 2014). Consistent with prior work (Li et al. 2022), we utilize the Karpathy (Karpathy and Fei-Fei 2017) split for training and evaluation. In our experiment, pre-training is conducted by filtering approximately 100k image-text pairs that involve multiple objects, attributes, and relationships. Subsequently, the models are evaluated on test splits, encompassing 5k images. We report Recall@1 on image-to-text retrieval(**IR**) and text-to-image retrieval(**TR**) to measure the ability of general representations.

Downstream Datasets. Two novel datasets (Yüksekgönül et al. 2022) are used to evaluate the structured representation performance of different models, where each test case consists of an image with matched captions and swapped mismatched captions. The model is tasked with distinguishing between aligned and unaligned captions based on the corresponding image.

- **Visual Genome Relation (VG-Relation).** As given an image and a caption containing a relationship triple, we evaluate the model's ability to select the caption where the relation is aligned with the image. Specifically, we expect the model to distinguish between "X relation Y" and "Y relation X" with a certain image (e.g., "an **astronaut** is riding a **horse**" v.s. "a **horse** is riding an **astronaut**" with the image in Fig. 1(a)).
- **Visual Genome Attribution (VG-Attribution).** Given the form "A_1 O_1 and A_2 O_2" and "A_2 O_1 and A_1 O_2", we evaluate the model's ability to accurately attribute the properties of objects. As shown in Fig. 1(b), we expect the model to distinguish between the caption "the **red** dress and the **blue** book" and the caption "the **blue** dress and the **red** book" according to the image.

Experimental Settings

All of our experiments are performed on a single NVIDIA A100 GPU with the Pytorch framework. We utilize a pre-trained Scene Graph Generator (Wu et al. 2019b) to extract the Scene Graph Knowledge. The structured Knowledge-Enhanced Encoder is implemented using a 6-layer Transformer architecture initialized with BERT-base (Devlin et al. 2019). During the training stage, we initialize the model with a pre-trained CLIP model and train it on our dataset for 10 epochs using a batch size of 128. We use a mini-batch AdamW optimizer with a weight decay of 0.1. The learning rate is initialized as 2e-6. The knowledge weight λ is 0.2.

Overall Results

Structured Representation Tasks. We compare our method with *8* representative or SOTA methods, including multi-modal models and large language models. As shown

Methods	Finetune	Negatives	KEE	VG-Attribution	VG-Relation
CLIP	✗	✗	✗	60.1	59.8
CLIP (fine-tune)	MSCOCO (*ours*)	✗	✗	64.0	66.5
Neg-CLIP	MSCOCO (*full*)	Random	✗	71.0	81.0
w/ {Random Change}	MSCOCO (*ours*)	Random	✗	73.9	77.7
w/ {Semantic Negative}	MSCOCO (*ours*)	Semantic	✗	77.8	79.0
w/ {Transformer}	MSCOCO (*ours*)	✗	✓	65.7	68.8
Structure-CLIP (Ours)	MSCOCO (*ours*)	Semantic	✓	**82.3** (↑ 11.3)	**84.7** (↑ 3.7)

Table 2: Results (%) of ablation study on VG-Relation and VG-Attribution datasets to analyze different components. Results show that each component greatly improves the ability of structured representation.

Types	KEE Layers	Fusion Weight (λ)	Embedding Fusion (ENC_{triple})	VG-Attribution	VG-Relation
Layers	1 layer	0.2	head + relation - tail	82.1	82.9
	2 layers	0.2	head + relation - tail	82.2	83.3
	6 layers	0.2	head + relation - tail	**82.3**	**84.7**
	12 layers	0.2	head + relation - tail	81.9	83.2
Weight	6 layers	0.0	head + relation - tail	77.8	79.0
	6 layers	0.01	head + relation - tail	**82.7**	83.5
	6 layers	0.2	head + relation - tail	82.3	**84.7**
	6 layers	1.0	head + relation - tail	82.3	83.8
Embedding	6 layers	0.2	Concat	81.1	83.3
	6 layers	0.2	head + relation + tail	81.9	83.3
	6 layers	0.2	head + relation - tail	**82.3**	**84.7**

Table 3: Ablation study of different hyperparameters and embedding methods.

in Table 1, we note that our Structure-CLIP has achieved the SOTA performance over all baselines across VG-Relation and VG-Attribute datasets.

Firstly, it is evident that NegCLIP outperforms the CLIP model in terms of structured representations, demonstrating that the aforementioned negative example sampling method can significantly enhance structured representations. Furthermore, by leveraging the guidance of Scene Graph Knowledge to improve the quality of constructing negative examples, Structure-CLIP achieves even further enhancement of structured representations. As a result, Structure-CLIP outperforms the existing multi-modal SOTA model (NegCLIP) by 12.5% on VG-Attribution and 4.1% on VG-Relation, respectively.

We also compare Structure-CLIP with existing Large Language models (LLMs) which use Maximum Likelihood Probability for an image and a text as the matching score. Our results demonstrate that as the model parameters of LLMs increase significantly, the structured representations also improve accordingly. However, Structure-CLIP still outperforms the OPT model by 3.7% and 0.4% respectively even though its parameters are less than 1% of it. Our results indicate that increasing model parameters to improve structured representations is resource-intensive and yields suboptimal performance, as the model primarily learns general representations rather than structured representations during the training stage. In contrast, our proposed Structure-CLIP method can significantly enhance structured representations with only a minimal increase in model parameters and a small amount of training.

General Representation Tasks. We evaluate the performances of Structure-CLIP on general representation tasks. Under the base model manner, Structure-CLIP achieves comparable performances with NegCLIP on the MSCOCO dataset. In other words, while greatly improving the performance of structured representations, Structure-CLIP retains the ability of general representations. Furthermore, our results demonstrate that both adequate general representations and structured representations can be obtained simultaneously using Structure-CLIP, whereas previous models generate insufficient structured representations. Under large model settings, our proposed method for domain fine-tuning significantly enhances both structured and general representations compared to the out-of-domain model.

Ablation Studies

Component Analysis. We perform an ablation study to assess multiple enhanced versions of the CLIP-base model on the VG-Relation and VG-Attribution datasets. The results of each variant are presented in Table 2.

Firstly, our experimental results demonstrate a significant performance improvement when applying *semantic negative* rather than *random negative* sampling strategy (*Line 4 vs Line 5*). A notable increase of 3.9%, 1.3% on VG-Attribution and VG-Relation datasets indicates that the proposed approach generates higher quality negative examples, resulting in superior structured representations.

Incorporating structured knowledge as input via the proposed Knowledge-Enhanced Encoder yields only a slight improvement (*Line 2 vs Line 6*). These findings imply that

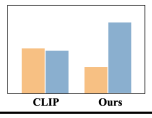
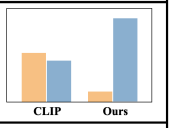
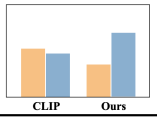
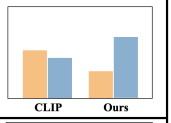
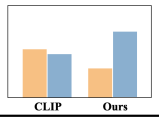
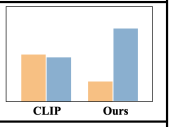
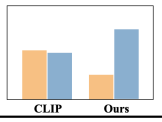
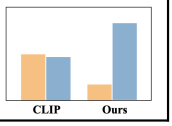
VG-Attribution Dataset			VG-Relation Dataset		
Image Case	**Caption**	**CLIP vs Ours**	**Image Case**	**Caption**	**CLIP vs Ours**
	**Aligned caption:** The **blue** sky and the **white** truck. **Unaligned caption:** The **white** sky and the **blue** truck.			**Aligned caption:** The **fans** is watching the **game**. **Unaligned caption:** The **game** is watching the **fans**.	
	**Aligned caption:** **Brown** dog and **grouped** buildings. **Unaligned caption:** **Grouped** dog and **brown** buildings.			**Aligned caption:** The **cup** is to the right of the **plates**. **Unaligned caption:** The **plates** is to the right of the **cups**.	
	**Aligned caption:** **White** wall and **wood** coffee table. **Unaligned caption:** **Wood** wall and **white** coffee table.			**Aligned caption:** The **bus** is on the **road**. **Unaligned caption:** The **road** is on the **bus**.	
	**Aligned caption:** The **large** bus and the **green** grass. **Unaligned caption:** The **green** bus and the **large** grass.			**Aligned caption:** The **dog** is on the **motorcycle**. **Unaligned caption:** The **motorcycle** is on the **dog**.	

Figure 3: Predictions of different approaches. The words in red and blue are two exchanged words. We compare our structure-CLIP with CLIP to calculate CLIP scores (i.e., semantic similarity) between the image and captions.

in order to achieve adequate structured representations, the incorporation of negative example sampling is necessary. Therefore, the Knowledge-Enhanced Encoder achieves a significant enhancement after combining it with semantic negative sampling (*Line 5 vs Line 7*).

Hyperparameter Analysis. Based on the results of Structure-CLIP presented in Table 3, we can conclude: *(i)* As the number of knowledge transformer layers increases, the model's capacity to represent multi-modal structured representations improves. However, it is important to note that beyond a certain threshold, the available data may become insufficient to support the model's increased capacity, leading to potential over-fitting. *(ii)* The experimental results demonstrate that without structured knowledge integration, the performance of the model is unsatisfactory (*Line 5*). Conversely, when structured knowledge is integrated, the variance in performance across different weights is minimal, indicating the effectiveness and intuitiveness of our method in enhancing structure representations.

Triple embeddings. We explored three different methods of triple embedding. The *concat* approach considers the order of input triple elements but fails to take into account the combination of head entities, relation entities, and tail entities. The *head + relation + tail* methods incorporate the combination relationship among triples. However, they lack the ability to distinguish the order of triples. For example, the final embeddings of two triples, $(cow, is, white)$ and $(white, is, cow)$ are identical, which can not help the model to make a distinction. Compared with these approaches, our triple embedding method takes into account both location and composition information. In this way, our Structure-CLIP model is better able to leverage structured knowledge within sentences to capture fine-grained semantic information and enhance multi-modal structured representations.

Case Study

The prediction results of the cases are presented in Fig. 3, which illustrates that Structure-CLIP can successfully distinguish between aligned and unaligned captions as given an image in a very large margin. However, the CLIP model encounters challenges in accurately determining the semantic similarities between these captions and the given image. In particular, the CLIP model exhibits near-uniform semantic similarities when two attributes or objects are swapped, indicating a lack of capacity to capture structured semantics. In contrast to the CLIP model, Structure-CLIP exhibits sensitivity to modifications in fine-grained semantics, indicating its ability to represent structured knowledge. As an example, the caption "the blue sky and the white truck" is used to evaluate the ability of Structure-CLIP to distinguish between aligned and unaligned captions when two attributes (i.e., blue and white) are exchanged. The results show that Structure-CLIP can make a distinction between aligned and unaligned captions with a margin of 25.16%, which further verifies the effectiveness of the proposed method in enhancing multi-modal structured representations.

Conclusion

In this paper, we propose Structure-CLIP aiming to integrate Scene Graph Knowledge to enhance multi-modal structured representations. Firstly, we use scene graphs to guide the construction of semantic negative examples. Additionally, we introduce a Knowledge-Enhanced Encoder that leverages Scene Graph Knowledge as input, thereby further enhancing the structured representations. Our proposed Structure-CLIP outperforms all recent methods on pre-training tasks and downstream tasks, which illustrates that Structure-CLIP can effectively and robustly understand the detailed semantics in multi-modal scenarios.

Acknowledgments

This work is supported by the Key Research and Development Program of Zhejiang Province (No. 2022C01011). This work is supported by the Fundamental Research Funds for the Central Universities (226-2023-00138). This work is funded by the National Natural Science Foundation of China (No. 62306276), Zhejiang Provincial Natural Science Foundation of China (No. LQ23F020017), Yongjiang Talent Introduction Programme (2022A-238-G), and Ningbo Natural Science Foundation (2023J291).

References

Chen, T.; Yu, W.; Chen, R.; and Lin, L. 2019. Knowledge-Embedded Routing Network for Scene Graph Generation. In *CVPR*, 6163–6171. Computer Vision Foundation / IEEE.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A. J.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; Kolesnikov, A.; Puigcerver, J.; Ding, N.; Rong, K.; Akbari, H.; Mishra, G.; Xue, L.; Thapliyal, A.; Bradbury, J.; Kuo, W.; Seyedhosseini, M.; Jia, C.; Ayan, B. K.; Riquelme, C.; Steiner, A.; Angelova, A.; Zhai, X.; Houlsby, N.; and Soricut, R. 2022. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *CoRR*, abs/2209.06794.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 1931–1942. PMLR.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*, 12. BMVA Press.

Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; and Ling, M. 2019. Scene Graph Generation With External Knowledge and Image Reconstruction. In *CVPR*, 1969–1978. Computer Vision Foundation / IEEE.

Huang, Y.; Chen, Z.; Zhang, W.; Chen, J.; Pan, J. Z.; Yao, Z.; Xie, Y.; and Chen, H. 2022. Aspect-based Sentiment Classification with Sequential Cross-modal Semantic Graph. *CoRR*, abs/2208.09417.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing Out of the Box: End-to-End Pre-Training for Vision-Language Representation Learning. In *CVPR*, 12976–12985. Computer Vision Foundation / IEEE.

Huo, Y.; Zhang, M.; Liu, G.; Lu, H.; Gao, Y.; Yang, G.; Wen, J.; Zhang, H.; Xu, B.; Zheng, W.; Xi, Z.; Yang, Y.; Hu, A.; Zhao, J.; Li, R.; Zhao, Y.; Zhang, L.; Song, Y.; Hong, X.; Cui, W.; Hou, D. Y.; Li, Y.; Li, J.; Liu, P.; Gong, Z.; Jin, C.; Sun, Y.; Chen, S.; Lu, Z.; Dou, Z.; Jin, Q.; Lan, Y.; Zhao, W. X.; Song, R.; and Wen, J. 2021. WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training. *CoRR*, abs/2103.06561.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 4904–4916. PMLR.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image Generation From Scene Graphs. In *CVPR*, 1219–1228. Computer Vision Foundation / IEEE Computer Society.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *CVPR*, 3668–3678. IEEE Computer Society.

Karpathy, A.; and Fei-Fei, L. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4): 664–676.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 5583–5594. PMLR.

Li, G.; Duan, N.; Fang, Y.; Gong, M.; and Jiang, D. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *AAAI*, 11336–11344. AAAI Press.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Li, J.; Selvaraju, R. R.; Gotmare, A.; Joty, S. R.; Xiong, C.; and Hoi, S. C. 2021a. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, 9694–9705.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.; and Chang, K. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *CoRR*, abs/1908.03557.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021b. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *ACL/IJCNLP (1)*, 2592–2607. Association for Computational Linguistics.

Lin, J.; Yang, A.; Zhang, Y.; Liu, J.; Zhou, J.; and Yang, H. 2020. InterBERT: Vision-and-Language Interaction for Multi-modal Pretraining. *CoRR*, abs/2003.13198.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. Springer.

Lin, Z.; Chen, X.; Pathak, D.; Zhang, P.; and Ramanan, D. 2023. VisualGPTScore: Visio-Linguistic Reasoning with Multimodal Generative Pre-Training Scores. *arXiv preprint arXiv:2306.01879*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 13–23.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. FLAVA: A Foundational Language And Vision Alignment Model. In *CVPR*, 15617–15629. IEEE.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP/IJCNLP (1)*, 5099–5110. Association for Computational Linguistics.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *CVPR*, 6619–6628. Computer Vision Foundation / IEEE.

Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; and Ross, C. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *CVPR*, 5228–5238. IEEE.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; and Wei, F. 2022a. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. *CoRR*, abs/2208.10442.

Wang, Y.; Yasunaga, M.; Ren, H.; Wada, S.; and Leskovec, J. 2022b. VQA-GNN: Reasoning with Multimodal Semantic Graph for Visual Question Answering. *CoRR*, abs/2205.11501.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2022c. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *ICLR*. OpenReview.net.

Wu, H.; Mao, J.; Zhang, Y.; Jiang, Y.; Li, L.; Sun, W.; and Ma, W. 2019a. Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations. In *CVPR*, 6609–6618. Computer Vision Foundation / IEEE.

Wu, H.; Mao, J.; Zhang, Y.; Jiang, Y.; Li, L.; Sun, W.; and Ma, W.-Y. 2019b. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6609–6618.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene Graph Generation by Iterative Message Passing. In *CVPR*, 3097–3106. IEEE Computer Society.

Yang, J.; Lu, J.; Lee, S.; Batra, D.; and Parikh, D. 2018. Graph R-CNN for Scene Graph Generation. In *ECCV (1)*, volume 11205 of *Lecture Notes in Computer Science*, 690–706. Springer.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*, 10685–10694. Computer Vision Foundation / IEEE.

Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3208–3216.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *CoRR*, abs/2205.01917.

Yüksekgönül, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *CoRR*, abs/2210.01936.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural Motifs: Scene Graph Parsing With Global Context. In *CVPR*, 5831–5840. Computer Vision Foundation / IEEE Computer Society.

Zhang, C.; Chao, W.; and Xuan, D. 2019. An Empirical Study on Leveraging Scene Graphs for Visual Question Answering. In *BMVC*, 288. BMVA Press.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In *CVPR*, 5579–5588. Computer Vision Foundation / IEEE.