mathematics

MDPI

Article

Spatial Position Reasoning of Image Entities Based on Location Words

Xingguo Qin [1,†], Ya Zhou [1,†] and Jun Li [1,2,*]

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; xgqin@guet.edu.cn (X.Q.); ccyzhou@guet.edu.cn (Y.Z.)
[2] Guangxi Key Laboratory of Image and Graphic Intelligence Processing, Guilin University of Electronic Technology, Guilin 541004, China
* Correspondence: lijun5221@126.com
† These authors contributed equally to this work.

Abstract: The endeavor of spatial position reasoning effectively simulates the sensory and comprehension faculties of artificial intelligence, especially within the purview of multimodal modeling that fuses imagery with linguistic data. Recent progress in visual image–language models has marked significant advancements in multimodal reasoning tasks. Notably, contrastive learning models based on the Contrastive Language-Image pre-training (CLIP) framework have attracted substantial interest. Predominantly, current contrastive learning models focus on nominal and verbal elements within image descriptions, while spatial locatives receive comparatively less attention. However, prepositional spatial indicators are pivotal for encapsulating the critical positional data between entities within images, which is essential for the reasoning capabilities of image–language models. This paper introduces a spatial location reasoning model that is founded on spatial locative terms. The model concentrates on spatial prepositions within image descriptions, models the locational interrelations between entities in images through these prepositions, evaluates and corroborates the spatial interconnections of entities within images, and harmonizes the precision with image–textual descriptions. This model represents an enhancement of the CLIP model, delving deeply into the semantic characteristics of spatial prepositions and highlighting their directive role in visual language models. Empirical evidence suggests that the proposed model adeptly captures the correlation of spatial indicators in both image and textual representations across open datasets. The incorporation of spatial position terms into the model was observed to elevate the average predictive accuracy by approximately three percentage points.

Keywords: visual–spatial reasoning; locative preposition; contrastive learning; image–text retrieval

MSC: 68-11

check for updates

Citation: Qin, X.; Zhou, Y.; Li, J. Spatial Position Reasoning of Image Entities Based on Location Words. *Mathematics* **2024**, *12*, 3940. https://doi.org/10.3390/math12243940

Academic Editor: Jonathan Blackledge

Received: 5 November 2024
Revised: 12 December 2024
Accepted: 13 December 2024
Published: 14 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

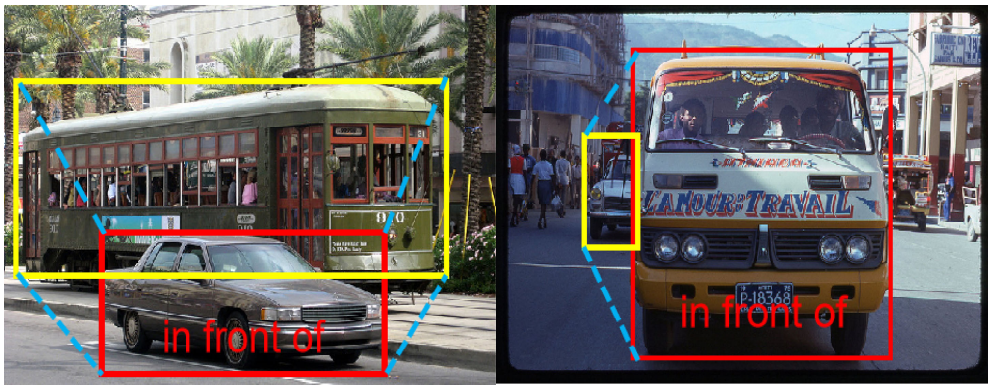
Languages typically have tens of thousands of names for physical objects and less than a hundred spatial prepositions, which greatly increases the difficulty of spatial reasoning based on spatial location words. Spatial reasoning is a spatial cognitive process based on the locations of entities in space, which can well represent the spatial position relationship between entities in images. At present, entity recognition models for images and text representation and understanding models have achieved good results in corresponding fields. In particular, modern language models (LMs), e.g., BERT [1], ALBERT [2], and XLNet [3], have seen great successes in natural language processing (NLP). However, there has been limited investigation into the spatial reasoning capabilities of LMs. The visual language model facilitates the integration of information from disparate modalities through the alignment of image and text features. This methodology emphasizes the correlation

between entity words in textual data and entity objects within visual representations, leveraging the robust learning capabilities of large-scale models to generate accurate predictions and inferences. Nonetheless, the model inherently lacks comprehension of the spatial relationship configurations among various entities depicted in the image. In essence, the model is deficient in the essential modeling and prediction of spatial orientation word features. For spatial locatives, such as prepositions or prepositional phrases, the semantic expression is ambiguous. Especially in the image association process, locative prepositions should have stronger semantic information.

At present, research on multimodal reasoning as it pertains to visual images and text descriptions is a hot issue. The advanced learning method based on CLIP has achieved good results [4]. CLIP aligns different regions in the image through text description and trains a matching model that fuses image and text. The visual language model facilitates the integration of information from disparate modalities through the alignment of image and text features. This methodology emphasizes the correlation between entity words in textual data and entity objects within visual representations, leveraging the robust learning capabilities of large-scale models to generate accurate predictions and inferences. Nonetheless, the model inherently lacks comprehension of the spatial relationship configurations among various entities depicted in the image. In essence, the model is deficient in the essential modeling and prediction of spatial orientation word features. Although the CLIP model has achieved good results in image description and image category classification tasks, it still focuses on entity words and functional words in language description; for example, it focuses on noun entities and action functional verbs between entities [5]. It is undeniable that these entity words and function words usually reflect the main content of the image to a large extent. However, for some images with obvious spatial location relationships between entities, this method easily ignores the role of spatial location words. Usually, spatial location words are composed of prepositions or prepositional phrases, in order to study the positive role of spatial location words in the image description. The task of image description markedly diverges from that of image classification. Image description constitutes a multimodal endeavor, wherein the training data encompass both visual imagery and accompanying textual descriptions. This task may involve generating textual descriptions predicated on visual input or, conversely, synthesizing images based on provided textual descriptions. In contrast, data for image classification tasks may consist solely of visual data, devoid of textual components. In this paper, we investigate spatial location words and attempt to determine whether the model can more accurately express the spatial location words in the process of text describing images, that is, whether the spatial location words can correspond to the spatial positions of entities in images one by one. In Figure 1, we selected two images in the VSR (visual–spatial reasoning) dataset to briefly describe the corresponding relationship between spatial location words and spatial entities in the image. Our goal is to determine whether there is an alignment between spatial location words and spatial locations between entities in the image. In our experimental framework, we meticulously investigated the specific corresponding regions of spatial position words within images. By systematically substituting spatial prepositions, we scrutinized whether such alterations induced any shifts in the predicted regions ascertained by the model.

Extensive research has delved into the fundamental role of language in visual imagery. Shahmohammadi et al. [6] elucidated the impact of lexical and sentential elements on visual perception. Their work extended to examining the intrinsic role of language within images by extracting key entities and functional verbs, correlating them with specific regions within the visual field. Collectively, scholarly inquiry into the nuanced influence of language on visual imagery remains nascent, with a particular deficiency in the domain of spatial locative terms, which to a significant extent constrains the expressive capacity of spatial terminology within imagery. Consequently, this paper, grounded in the paradigm of comparative image–language learning, concentrates on the influence of locative terms on the model's reasoning and judgment faculties. Through the conception of pertinent tasks, including binary judgment tasks, the model is tasked with assessing the accuracy of

provided image descriptions, with a particular emphasis on the precision of spatial locative terms. A further example is the mask-filling task, wherein the model is required to select the most precise locative terms from a given set after obscuring the locative terms within the image description, thereby evaluating its capacity for spatial semantic inference.



(**a**) (**b**)

Figure 1. The spatial position relationship between the prepositional phrase "in front of" and entities in an image. (**a**) The car is in front of the bus. (**b**) The bus is in front of the car.

The novel contributions of this scholarly article are articulated as follows: Firstly, diverging from the predominant multimodal models that concentrate on the correlation and alignment of images and texts, this study delves into the underexplored domain of prepositions, specifically examining the hypothesis that spatial prepositions may correlate with the positioning of entities within images. This exploration extends to the investigation of the integration mechanism by which dual modal information is synthesized within the feature space. Secondly, in contrast to multimodal models that emphasize the role of entity words, this research investigates the role and significance of spatial prepositions in the context of multimodal feature matching. This inquiry is pivotal in understanding the nuanced interplay between linguistic elements and their visual counterparts, thereby enriching the discourse on multimodal information processing and representation.

The architectural framework of this manuscript is outlined as follows: In Section 2, this manuscript presents an exhaustive review of the extant research pertaining to advancements in image and text models, with a pronounced focus on assessing the efficacy of contemporary multimodal models in the domain of feature extraction. This section meticulously delineates the central research query underpinning this study: the potential influence of prepositional representations on the spatial positioning of entities within images. In Section 3, we elucidate the methodology employed in this research, accompanied by a detailed exposition of the specific implementation nuances of the proposed approach. Section 4 further validates the efficacy and feasibility of the proposed method through a series of empirical experiments, while also delving into the correlation between prepositional representation and image location. Emphasis is accorded to the congruence between spatial prepositions and their linguistic correlates, as well as the intricacies of alignment, training, and predictive strategies concerning spatial prepositions and the positioning of entities. Subsequent empirical experiments serve to affirm the effectiveness of the proposed method, with ablation studies employed to reinforce the robustness of the techniques presented. Concluding the discourse, Section 5 deliberates on the inherent limitations of the methodology and posits prospective directions for future enhancement, thereby providing a comprehensive framework for the continued evolution of this research endeavor.

2. Related Work

The landscape of multimodal research on image–text data is burgeoning, with numerous studies on image–text captioning achieving commendable outcomes. The majority of research endeavors in image captioning have concentrated on the correlation between

entity objects within images and their lexical counterparts in textual descriptions [7]. Comparative learning methodologies have enabled the modeling of entity features across two distinct domains. Mohammed [8] introduced a methodology employing deep learning techniques to predict and ascertain the azimuth of spatial orientations, with applications particularly relevant to the field of autonomous driving. Consequently, spatial position reasoning is deemed to possess significant practical utility. Nevertheless, the existing body of research on multimodal matching and alignment methodologies is notably deficient in its exploration of spatial prepositions. Given their inherent attributes related to spatial positioning, prepositions are posited to have a more substantial impact on multimodal text-image models.

CLIP: Radford et al. [4] proposed an efficient and scalable method of learning from natural language supervision for contrastive language-image pre-training, named CLIP, which leverages a much broader source of supervision to learn directly from image raw text. CLIP was a pioneer, and much of the research that followed was based on contrastive learning methods. Rinon et al. [5] presented a text-driven method that allows shifting a generative model to new domains without having to collect even a single image. Crowson et al. [9] proposed a novel methodology that can produce images of high visual quality from text prompts of significant semantic complexity without any training by using a multimodal encoder to guide image generation. To improve the few-shot performance in CLIP, Zhang et al. [10] proposed a training-free adaptation method for CLIP to conduct few-shot classification, named tip-adapter, which constructs the adapter via a key–value cache model from the few-shot training set and updates the prior knowledge encoded in CLIP by feature retrieval. Wang [11] designed a disentangled conditional NeRF architecture that allows disentangling shape and appearance while editing an object [12], and two feedforward code mappers enable fast inference for editing different objects. The CLIP model employs cosine similarity to predict label classifications by assessing the congruence between textual and visual data. This approach is primarily concerned with the alignment of text and image representations. Particularly within the realm of image classification tasks, the model's performance is suboptimal when the output is confined to a single category label, attributed to the sentence-based input during the pre-training phase. The methodology presented in this paper compels the model to prioritize the spatial positional information of the image during the reasoning process, thereby enhancing the model's inferential accuracy. It achieves this by leveraging the descriptive relationships between spatial orientation prepositions and entities to map the spatial positions within the image. Fundamentally, this remains a matching alignment technique; however, this study places a heightened emphasis on the spatial interrelations among entities within the image, offering a more nuanced understanding of their spatial dynamics.

Image Caption: In recent years, significant advancements in the field of image captioning, leveraging vision–language pre-training, have been observed [13]. Mohammed et al. [14] introduced an implicit alignment technique bridging the domains of vision and language, wherein interlingual textual data interacts to enhance pre-training textual word embeddings, thereby significantly bolstering the linguistic articulation of image captions. Shahmohammadi et al. [6] focused on aligning textual embeddings with visual data, preserving the distributional statistics that define word usage in textual corpora, leading to the generation of visually grounded embeddings for previously unseen words. Ramesh [15] delineated a straightforward approach employing a transformer architecture that autoregressively models text and image tokens as a unified stream of data. Sun et al. [16] crafted variants of layerwise relevance propagation (LRP) and gradient-based explanation methods, specifically tailored for image captioning models equipped with attention mechanisms. Li et al. [17] introduced a grounded language–image pre-training (GLIP) model that integrates object detection and phrase grounding tasks, aiming to learn an object-level, linguistically informed, and semantically enriched visual representation. Chen et al. [18] proposed VisualGPT, employing a novel self-resurrecting encoder–decoder attention mechanism designed to swiftly adapt pre-trained language models (PLMs) with

a limited amount of in-domain image–text data [19]. Furthermore, Li et al. [20] presented a distinctive architecture for text-to-image synthesis, termed T2IGAN, which efficiently amalgamates textual and visual features.

Spatial Reasoning: Spatial reasoning constitutes the cornerstone of multimodal data comprehension and inferencing, with the early bAbI project [21] being a seminal example of a spatial reasoning dataset. The dataset's textual content predominantly comprises descriptions of the spatial positional relationships among entities. However, the simplicity of image descriptions within this dataset allows subsequent researchers to achieve high accuracy through rudimentary matching models, which do not necessitate complex reasoning processes [21–23]. There exists a disproportionate focus on the spatial relationships between entities within images, such as the precision of entity recognition and boundary segmentation tasks. These methodologies differ significantly from the spatial locative terms in texts, which are the focus of our interest in this study.

3. Methods

The contrastive learning model described by image language brings great hope to multimodal research. In our method, we bring more abundant spatial information to the model through the perspective of location words to better understand the meaning of specific spatial location words. In addition, because the typical contrast model ignores the different information of locative prepositions or phrases, integrating locative word knowledge into visual image language descriptions can help to obtain better spatially embedded representations.

Our model uses the image–language contrast learning model CLIP in the pre-training stage. To prevent the model from paying too much attention to the entity words or function words in the text, we add the representation T' after random shuffling of the captions in the text representation. The T' disrupts the sequential order of words within the text, compelling the model to focus more intently on the intrinsic features of the text. The introduction of T' is strategically designed to prevent the model from memorizing the positional cues of locational words during training, thereby potentially skewing the model's genuine reasoning capabilities in practical applications. At this time, pre-training data contain three parts,

$$D =< I_embedding, T_original, T_shuffle > \quad (1)$$

where $I_embedding$ indicates the embedded representation of an image after encoding, $T_original$ represents the description text embedded representation of the image, and $T_shuffle$ indicates the embedded representation of the text after the word order is scrambled.

Then, the embedded representation is formed into a comparative learning matrix to calculate the MSE loss of image and text embedding. In the prediction, we need to predict the location words instead of noun entities or functional verbs and finally determine the best words according to the prediction probability of candidate location words. The specific block diagram is shown in Figure 2.

Our method is generally divided into two parts: (a) contrastive pre-training and (b) noun locality prediction. At stage (a), for the input image description text, we use the transformer-based BERT model for encoding and obtain the encoding vector $T_original$ of the original text, which is represented by vector $T_original = [T_1, T_2, \ldots, T_n]$ in Figure 2. Additionally, the original image caption text after shuffle is expressed as $T_shuffle$, which is represented by vector $T_shuffle = [T_1', T_2', \ldots, T_n']$. For image training, we use the Inception-V4 model based on a transformer to train and obtain the encoded representation of the images. In Figure 2, we use the vector $I = [I_1, I_2, \ldots, I_n]$ for representation. The calculation method of element M_{11} in the contrastive matrix is shown in Formula (2).

$$M_{11} = I_1 \cdot T_1 + I_1 \cdot T_1' \quad (2)$$

where, I_1, T_1, and $T_1{}'$ are image, original text, and shuffled text embedding vectors, respectively.

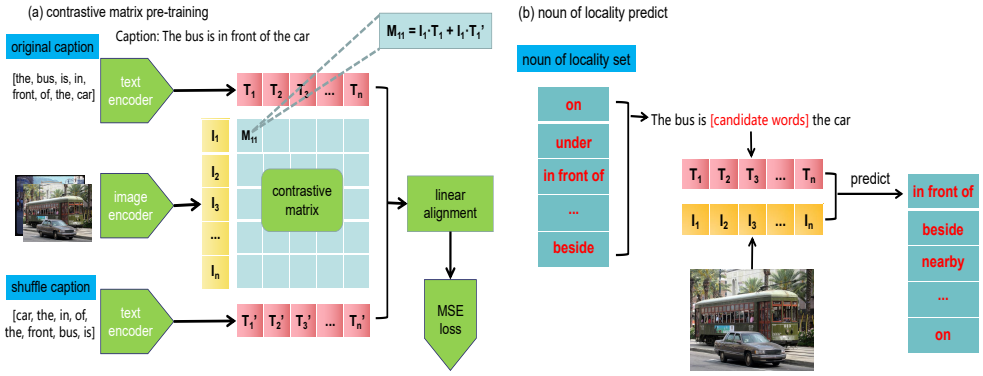


Figure 2. Overall diagram of predictive and inference for locality noun. (**a**) describes contrastive pre-training learning, which constructs a contrastive learning matrix. (**b**) shows the processing of nouns for locality prediction.

The output of the contrastive learning matrix will pass through the linear alignment layer, which integrates the two kinds of text information into the image representation information. Using the parameters of the MSE (Mean Squared Error) loss optimization model, the MSE losses of two texts and the same image are calculated in Formula (3).

$$\begin{aligned} L(\theta_original) &= \frac{1}{2}\sum_{i=1}^{n}(V_I - V_T)^2 \\ L(\theta_shuffle) &= \frac{1}{2}\sum_{i=1}^{n}(V_I - V_T')^2 \end{aligned} \tag{3}$$

where V_I is the embedded vector of the image data in the sample and V_I and $V_T{}'$ are the original text and the text after shuffling, respectively. $\theta_original$ and $\theta_shuffle$ denote the textual training parameters of the model following the processing of text in its original sequence and after the text has been subjected to a permutation, respectively. This step is calculated by MSE after linear alignment, and the final loss is the average of the two, as shown in Formula (4).

$$L(\theta) = \frac{1}{2}[L(\theta_original) + L(\theta_shuffle)] \tag{4}$$

where, $\theta_original$ and $\theta_shuffle$ are the parameters for training the model with the original text and shuffle text, respectively, and θ represents the final optimization parameters of the model. During the pre-training phase, the model employs the gradient descent algorithm to achieve convergence. Concurrently, this phase is instrumental in determining and fixing the parameter θ. In the subsequent prediction phase, the model utilizes the learned parameter θ to compute the output value y. The softmax function is then applied to normalize y, constraining it within the interval [0, 1], thereby yielding the predicted probability.

In phase (b), we provide a set of all location words in the dataset. Additionally, we mask the location words in the image caption text, and this process is very similar to the BERT mask model. The model is capable of generating the corresponding word vector representations for predicting words that are occluded. As it takes into account the contextual information surrounding the occluded words, the resulting word change vectors are imbued with rich semantic features. Specifically, we match the location words in the image captions with those in Table 1 and then randomly mask the matched location words. After text encoding, we measure the similarity with the image encoder vector. Finally, we sort the candidate location words according to their probability values to obtain the

final location words. In this part, we also need the model to optimize the parameters. The optimization strategy comes from the gap between the predictive words and the real words.

Table 1. Comparison of retrieval results (R@K(%)) on the MSCOCO and Flickr30K test sets.

Methods	MSCOCO Dataset						Flickr30K Dataset					
	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [24]	72.7	94.8	98.4	58.8	88.4	94.8	67.4	90.3	95.8	48.6	77.7	85.2
	72.3	94.8	98.3	58.5	87.9	95.0	68.1	89.7	95.2	51.5	77.1	85.3
SGM [25]	73.4	93.8	97.8	57.5	87.3	94.3	71.8	91.7	95.5	53.5	79.6	86.5
VSRN [26]	74.0	94.3	97.8	60.8	88.4	94.1	70.4	89.2	93.7	53.0	77.9	85.7
RDAN [27]	74.6	96.2	98.7	61.6	89.2	94.7	68.1	91.0	95.9	54.1	80.9	87.2
BFAN [28]	74.9	95.2	-	59.4	88.4	-	68.1	91.4	-	50.8	78.4	-
PFAN [29]	76.5	96.3	99.0	61.6	89.6	95.2	70.0	91.8	95.0	50.4	78.7	86.1
MMCA [30]	74.8	95.6	97.7	61.6	89.8	95.2	74.2	92.8	96.4	54.8	81.4	87.8
CAAN [31]	75.5	95.4	98.5	61.3	89.7	95.2	70.1	91.6	97.2	52.8	79.0	87.9
DPRNN [32]	75.3	95.8	98.6	62.5	89.7	95.1	70.2	91.6	95.8	55.5	81.3	88.2
IMRAM [33]	76.7	95.6	98.5	61.7	89.1	95.0	74.1	93.0	96.6	53.9	79.4	87.2
SAF [34]	76.1	95.4	98.3	61.8	89.4	95.3	73.7	93.3	96.3	56.1	81.5	88.0
SGR [34]	78.0	95.8	98.2	61.4	89.3	95.4	75.2	93.3	96.6	56.2	81.0	86.5
SGRAF [34]	79.6	96.2	98.5	63.2	90.7	96.1	77.8	94.1	97.4	58.5	83.0	88.8
Ours	79.8	96.6	99.2	62.4	89.8	95.6	78.2	94.2	97.4	58.6	82.8	89.0

4. Experiment and Analysis

4.1. Experiment Datasets

We used the Cambridge visual–spatial reasoning dataset (VSR) (github.com/cambridgeltl/visual-spatial-reasoning, accessed on 10 May 2024), Microsoft COCO 2017 dataset (https://cocodataset.org/, accessed on 1 March 2024), and Flickr30K (http://shannon.cs.illinois.edu/DenotationGraph/data/index.html, accessed on 1 June 2024) for our experiments. The COCO dataset mainly used recognition and sampling in images. The COCO 2017 dataset contains 12,000+ images and 880,000+ instances. The VSR dataset is mainly used to judge spatial location words or spatial relations. The VSR dataset contains 71 types of locality nouns, and all the locality nouns are listed in Figure 3. The Flickr30K dataset contains 31,783 images, and each image is annotated with five annotated captions. The training dataset encompasses both the original imagery and their corresponding textual descriptions, as well as the textual descriptions that have undergone a shuffling process. Collectively, these elements form the foundation of the pre-training dataset.

In Figure 3, there are seven categories, including 71 types of location words. In fact, due to the limitation of sample data, six types of location words are not reflected in the data, and they are marked as "*" in Figure 3.

In the pre-training stage, we adopted the CLIP comparative learning method. Specifically, we used BERT to train the text data to obtain the word embedded representation of the text description. In fact, we used the word and sentence embedded representation based on the image background proposed by Hassan. This method can best fit the alignment between the image and the text. More importantly, this method also considers the image information for the embedding of nonentity words or functional words. We used a batch size of 256 image–caption pairs. We trained for 20 epochs using the NAdam optimizer [35]

with a learning rate of 0.0001. Additionally, we deleted the punctuation in the caption and selected the 5000 words that most often appeared in the captiona. The rest of the words were ignored. Image vectors were obtained from the pre-training model Inception-V4 [36].

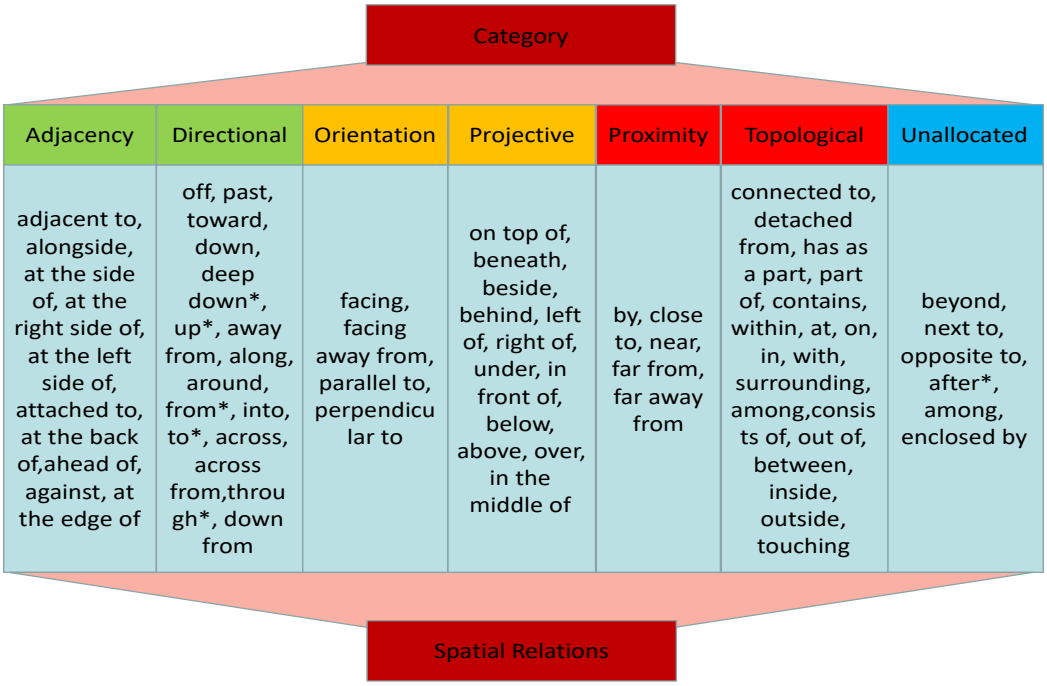


Figure 3. Overall diagram of 7 spatial categories, including 71 types of spatial relations.

In the experiment, we adopted two strategies for the allocation of train/dev/test data. The first was to randomly divide the dataset according to the ratio of train/dev/test data 7:1:2, and the second was to zero-shot the dev/test data. That is, if an entity was paired out of the test data, it would not appear in the dev/test data. The second was use the ratio 5:2:3. Because of training, if the verification and test data did not coincide, the zero-shot split prediction would be more challenging. However, this method helped improve the robustness and overall prediction capability of the model.

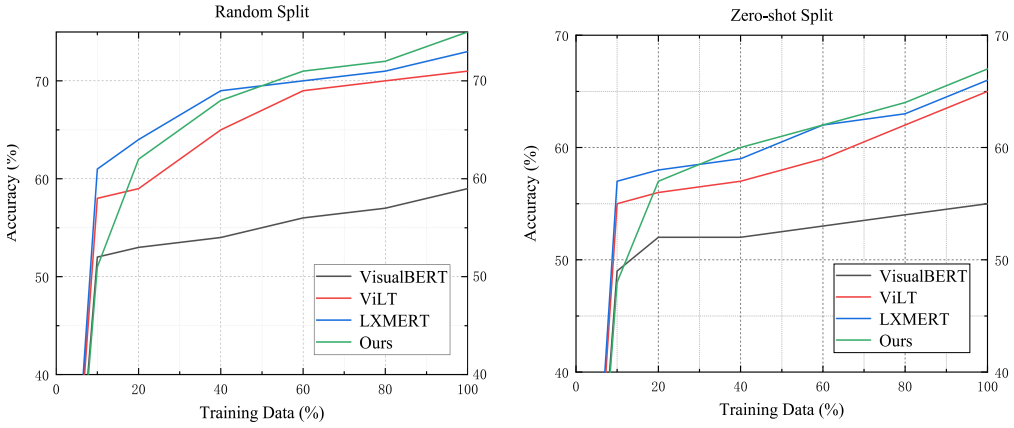
4.2. Baselines and Experiment

For comparison, we chose three mainstream baselines, VLMs: Visual-BERT, LXMERT [37], and ViLT [38]. They all use image and text pairs as input, and they all use a transformer-based mechanism. The greatest difference between the three is the difference in the location and coding mode.

The first group of experiments compares the accuracy of our model with that of the three baselines in the direction word prediction task. The experimental data are randomly split and zero-shot split. The experimental results are shown in Figure 4.

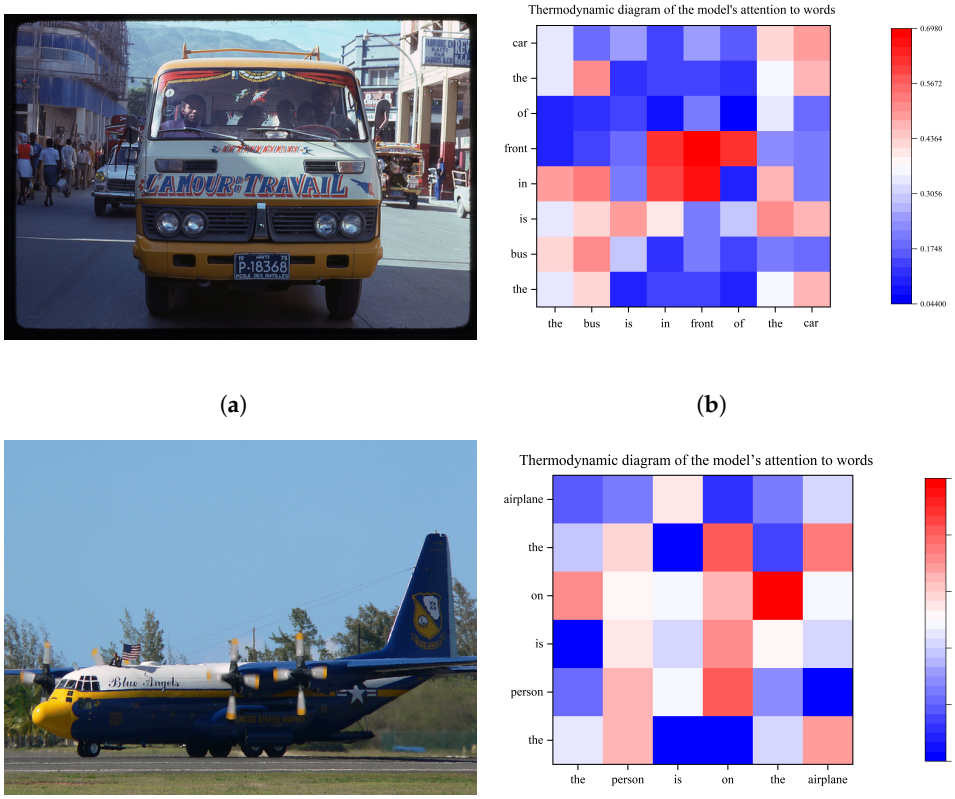
Figure 4 shows the performance of different models. Overall, the random split is better than the zero-shot split, and this result is consistent with our previous prediction. In addition, when the quantity of training data is low, our model performance is not the best. However, with the increase in the quantity of data, the accuracy of the model in the random split and zero-shot split reaches 75% and 67%, respectively, reaching the best value of several models. Even so, in the face of zero-shot split data, the performance of the model still needs to be improved, which also indicates that the model has a strong dependence on data.

To verify that the model pays attention to the location words in the caption, we simulated the thermodynamic diagram of the model's attention to the words in the caption. Figure 5 shows the attention of different words in the caption.



(**a**) (**b**)

Figure 4. Comparison of location word prediction in different datasets: results on both random split (**a**) and zero-shot split (**b**)—model performance under different quantities of training data (10% 100% training set).



(**a**) (**b**)

(**c**) (**d**)

Figure 5. Thermodynamic diagram of the model's attention to words. (**a**) Caption: the bus is **in front of** the car. (**b**) The phrase "in front of" is given more concentrated attention (the red area in the figure). (**c**) Caption: the person is **on** the airplane. (**d**) The word "on" is not the focus (the red area in the figure).

In Figure 5b, we can see a red area in the middle of the diagram, and this part happens at the location of the locality noun "in front of". The red area is relatively concentrated, which reflects that the model pays more attention to the spatial location words. However, in Figure 5d, it is found that the focus of the spatial location word "on" is not high, even divergent. We believe that in Figure 5b, because the aircraft has an oversized tail fin, which is located at the upper right of the image and at the upper space orientation, this acquisition

adds much confusion to the model. Additionally, the physical aircraft is much larger than the physical person, which also weakens the spatial location relationship.

Concurrently, this study introduces an additional experimental paradigm, wherein prepositions or prepositional phrases within the text corpus are obfuscated, thereby necessitating the model to engage in autonomous reasoning to ascertain the appropriate spatial prepositions. Methodologically, the dataset was partitioned into a training subset and a testing subset in an 8:2 ratio, with a random selection of 1000 images from the latter subset for empirical evaluation. The outcomes of this experiment are graphically depicted in Figure 6.

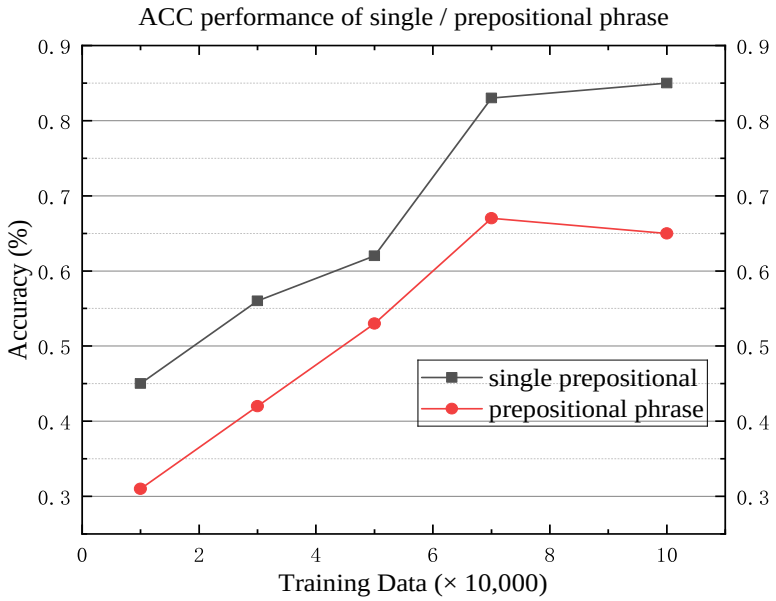


Figure 6. Prediction accuracy of single prepositions and prepositional phrases.

As delineated in Figure 6, the aggregate predictive accuracy of the model concerning spatial prepositional phrases is observed to be inferior to that for individual prepositions. This disparity is posited to stem from the inclusion of lexical elements within prepositional phrases that possess limited semantic content, such as "of" and "the", which may impede the model's comprehensive prediction of prepositional phrases. Optimal performance was achieved when the model was trained on a dataset comprising 70,000 images. Conversely, a decline in performance was noted when the model was exposed to a dataset of 100,000 images. Upon meticulous examination of the dataset, a proliferation of analogous data instances was identified, suggesting the presence of model overfitting.

4.3. Visual Language Reasoning

To rigorously assess the semantic reasoning capabilities of our proposed model, we undertook a visual language reasoning experiment, which integrates both image retrieval and text retrieval tasks. The image retrieval component involves identifying the corresponding image based on its title, whereas text retrieval involves the converse operation. In the context of image–text retrieval, this study employs the recall metric to evaluate performance. Specifically, recall at K (R@K) [34] is utilized to measure the proportion of correct retrievals within the first K predicted outcomes. We employ R@1, R@5, and R@10 as our evaluation metrics, which denote the accuracy of the top 1, 5, and 10 predictions, respectively. A comparative analysis with other high-performing image–text retrieval methods is presented in Table 1, showcasing the results of our experiment.

We briefly describe several baseline methods.

SCAN [24]: presents stacked cross attention to discover the full latent alignments using both image regions and words in a sentence as context and infer image–text similarity.

CAMP [39]: proposes cross-modal adaptive message passing (CAMP), which adaptively controls the information flow for message passing across modalities.

SGM [25]: designs a so-called scene graph matching (SGM) model, which can refine the representation of each node on the graph by aggregating neighbourhood information.

VSRN [26]: a visual–semantic reasoning network (VSRN) is proposed to generate a visual representation that captures both objects and their semantic relationships.

RDAN [27]: proposes a relationwise dual attention network (RDAN) for image–text matching, which encodes the local correlations and the global dependencies between regions and words by training a visual-semantic network.

BFAN [28]: presents a novel bidirectional focal attention network (BFAN), which not only allows attention to relevant fragments but also diverts all the attention into these relevant fragments to concentrate on them.

PFAN [29]: proposes a novel position-focused attention network (PFAN) to investigate the relation between visual and textual views.

MMCA [30]: proposes a novel multimodality cross-attention (MMCA) network for image and sentence matching by jointly modeling the intramodality and intermodality relationships of image regions and sentence words in a unified deep model.

CAAN [31]: proposes a unified context-aware attention network (CAAN), which selectively focuses on critical local fragments (regions and words) by aggregating the global context.

DPRNN [32]: proposes a dual-path recurrent neural network (DP-RNN) that processes images and sentences symmetrically by recurrent neural networks.

IMRAM [33]: proposes an iterative matching with recurrent attention memory (IMRAM) method, which captures the correspondences between images and texts with multiple steps of alignment.

SAF/SGR/SGRAF [34]: proposed similarity attention filtration (SAF), similarity graph reasoning (SGR), and similarity graph reasoning and attention filtration (SGRAF), which selectively attend to significant and representative alignments while casting aside the interferences of nonmeaningful alignments.

In Table 1, we can see that our model performs best in text retrieval tasks for the MSCOCO dataset, and there is still a gap between our model and the best method in image retrieval tasks. For the Flickr30K dataset, the model also performs best in text retrieval tasks, and two recall indicators (R@5, R@10) achieve the best results in image retrieval tasks. Overall, our model met certain expectations, but it did not reach the optimal level in the image retrieval task. We analyzed that the model focused on text training as a whole, especially on the expression of location words in the text, which is the reason why the model performs better in the text retrieval task. In contrast, the regional relevance between locators and images in the text is still insufficient, which leads to the lack of expressiveness of the model in image retrieval tasks.

4.4. Ablation Studies

In this section, we conduct a series of ablation studies to explore the impact of different configurations on the caption shuffle module, random split and zero-shot split. The impact of training configurations on the MSCOCO and Flickr30K test sets is shown in Table 2.

In Table 2, we conduct ablation contrast experiments on four modules. On two datasets, integrating the three modules, original caption, shuffle caption, and random split, has the best performance. This method maximizes the semantics of captions. Additionally, the random split method avoided the negative impact of the zero-shot problem in the test set, and this result was in line with expectations.

Table 2. Impact of training configurations on MSCOCO and Flickr30K test sets.

Dataset	Original Caption	Shuffle Caption	Random Split	Zero-Shot Split	Text Retrieval			Image Retrieval		
					R@1	R@5	R@10	R@1	R@5	R@10
MSCOCO	✓			✓	68.3	88.9	90.1	56.3	79.0	91.2
	✓		✓		68.6	89.5	91.4	57.3	80.3	92.2
	✓	✓		✓	75.4	94.3	96.5	60.3	85.7	94.0
	✓	✓	✓		**79.8**	**96.6**	**99.2**	**62.4**	**89.8**	**95.6**
Flickr30K	✓			✓	70.2	85.4	90.2	50.8	72.9	80.1
	✓		✓		73.3	87.6	91.5	52.7	74.8	82.7
	✓	✓		✓	75.8	89.7	93.2	55.3	76.9	86.4
	✓	✓	✓		**78.2**	**94.2**	**97.4**	**58.6**	**82.8**	**89.0**

5. Conclusions

In this study, we delved into the role of spatial locators in image description tasks, endeavoring to ascertain whether these locators can effectively model the associative ties between images and texts. We employed a comparative learning approach to investigate the spatial relationships between entities within images and their correlation with spatial locators, predominantly prepositions and prepositional phrases. However, the current model's insufficient emphasis on spatial locators hinders the full extraction of the semantic information inherent in these locators, introducing uncertainty into this correlation. Furthermore, the model's performance is significantly influenced by the input data. When the positional disparity between entities in the image is minimal or the location word is perturbed by image entities, the model's focus on the location word is markedly diminished.

In future work, we intend to refine the existing comparative learning framework by integrating a preposition-based orientation word focus module, enhancing the model's embedded representation of orientation words and more effectively capturing the positional relationships between entities within the image. Additionally, we aim to simplify the complexity of image pre-training in the current comparative model, thereby enhancing the model's deployability and operational efficiency.

Author Contributions: The authors confirm their contributions to the paper as follows: study conception and design: X.Q. and J.L.; data collection: J.L. and X.Q.; analysis and interpretation of results: J.L. and Y.Z.; draft manuscript preparation: J.L. and X.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Guangxi Natural Science Foundation (No. 2022GXNSFBA 035510), the Guangxi Key Research and Development Program (No. Guike AB23075178), the Guangxi Key Laboratory of Image and Graphic Intelligent Processing (No. GIIP2207), and the National Natural Science Foundation of China (No. 62267002).

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
2. Roshanak, M.; Hossein Rajaby, F.; Qiang, N.; Parisa, K. ALBERT: A lite bert for self-supervised learning of language representations. In Proceedings of the 2020 International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
3. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
4. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
5. Gal, R.; Patashnik, O.; Maron, H.; Bermano, A.H.; Chechik, G.; Cohen-Or, D. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ACM Trans. Graph.* **2022**, *41*, 1–13. [CrossRef]
6. Shahmohammadi, H.; Heitmeier, M.; Shafaei-Bajestan, E.; Lensch, H.; Baayen, H. Language with vision: A study on grounded word and sentence embeddings. *arXiv* **2022**, arXiv:2206.08823. [CrossRef] [PubMed]
7. Yao, J.; Li, J.; Li, Y.; Zhang, M.; Zuo, C.; Dong, S.; Dai, Z. A Vision–Language Model-Based Traffic Sign Detection Method for High-Resolution Drone Images: A Case Study in Guyuan, China. *Sensors* **2024**, *24*, 5800. [CrossRef] [PubMed]
8. Mohammed, M.S.; Abduljabar, A.M.; Faisal, M.M.; Mahmmod, B.M.; Abdulhussain, S.H.; Khan, W.; Liatsis, P.; Hussain, A. Low-cost autonomous car level 2: Design and implementation for conventional vehicles. *Results Eng.* **2023**, *17*, 100969. [CrossRef]
9. Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; Raff, E. VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance. In Proceedings of the Computer Vision–ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; pp. 88–105.
10. Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; Li, H. Tip-adapter: Training-free adaption of clip for few-shot classification. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 493–510.
11. Wang, C.; Chai, M.; He, M.; Chen, D.; Liao, J. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3835–3844.
12. Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; Chang, S.F. Clip-event: Connecting text and images with event structures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16420–16429.
13. Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; Wang, L. Scaling up vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17980–17989.
14. Mohammed, W.; Shahmohammadi, H.; Lensch, H.; Baayen, R.H. Visual Grounding of Inter-lingual Word-Embeddings. *arXiv* **2022**, arXiv:2209.03714.
15. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8821–8831.
16. Sun, J.; Lapuschkin, S.; Samek, W.; Binder, A. Explain and improve: LRP-inference fine-tuning for image captioning models. *Inf. Fusion* **2022**, *77*, 233–246. [CrossRef]
17. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.N.; et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10965–10975.
18. Chen, J.; Guo, H.; Yi, K.; Li, B.; Elhoseiny, M. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18030–18040.
19. Mu, N.; Kirillov, A.; Wagner, D.; Xie, S. Slip: Self-supervision meets language-image pre-training. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 529–544.
20. Li, W.; Wen, S.; Shi, K.; Yang, Y.; Huang, T. Neural Architecture Search With a Lightweight Transformer for Text-to-Image Synthesis. *IEEE Trans. Netw. Sci. Eng.* **2022**, *9*, 1567–1576. [CrossRef]
21. Weston, J.; Bordes, A.; Chopra, S.; Rush, A.M.; Van Merriënboer, B.; Joulin, A.; Mikolov, T. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv* **2015**, arXiv:1502.05698.
22. Dheeru, D.; Yizhong, W.; Pradeep, D.; Gabriel, S.; Sameer, S.; Matt, G. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 2368–2378.

23. Pradeep, D.; Nelson F., L.; Ana, M.; Noah A., S.; Matt, G. QUOREF: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. 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), Hong Kong, China, 3–7 November 2019; pp. 5925–5932.
24. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 201–216.
25. Wang, S.; Wang, R.; Yao, Z.; Shan, S.; Chen, X. Cross-modal scene graph matching for relationship-aware image-text retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1508–1517.
26. Li, K.; Zhang, Y.; Li, K.; Li, Y.; Fu, Y. Visual semantic reasoning for image-text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4654–4662.
27. Hu, Z.; Luo, Y.; Lin, J.; Yan, Y.; Chen, J. Multi-Level Visual-Semantic Alignments with Relation-Wise Dual Attention Network for Image and Text Matching. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 789–795.
28. Liu, C.; Mao, Z.; Liu, A.A.; Zhang, T.; Wang, B.; Zhang, Y. Focus your attention: A bidirectional focal attention network for image-text matching. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 3–11.
29. Yaxiong, W.; Hao, Y.; Xueming, Q.; Lin, M.; Jing, L.; Biao, L.; Xin, F. Position Focused Attention Network for image-text Matching. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 3792–3798.
30. Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; Wu, F. Multi-modality cross attention network for image and sentence matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10941–10950.
31. Zhang, Q.; Lei, Z.; Zhang, Z.; Li, S.Z. Context-aware attention network for image-text retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3536–3545.
32. Chen, T.; Luo, J. Expressing objects just like words: Recurrent visual embedding for image-text matching. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10583–10590.
33. Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; Han, J. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12655–12663.
34. Diao, H.; Zhang, Y.; Ma, L.; Lu, H. Similarity reasoning and filtration for image-text matching. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 19–21 May 2021; Volume 35, pp. 1218–1226.
35. Zhang, Z. Improved Adam Optimizer for Deep Neural Network. In Proceedings of the IEEE/ACM 26th International Symposium on Quality of Service, Banff, AB, Canada, 4–6 June 2018; pp. 1–2.
36. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
37. Hao, T.; Mohit, B. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 3–7 November 2019.
38. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 5583–5594.
39. Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; Shao, J. Camp: Cross-modal adaptive message passing for text-image retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5764–5773.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.