BiMAC: Bidirectional Multimodal Alignment in Contrastive Learning

Masoumeh Zareapoor[1,3], **Pourya Shamsolmoali**[2,3*], **Yue Lu**[2]

[1] Shanghai Jiaotong University, Shanghai, China
[2] East China Normal University, Shanghai, China
[3] University of York, York, United Kingdom
pshams55@gmail.com

Abstract

Achieving robust performance in vision-language tasks requires strong multimodal alignment, where textual and visual data interact seamlessly. Existing frameworks often combine contrastive learning with image captioning to unify visual and textual representations. However, reliance on global representations and unidirectional information flow from images to text limits their ability to reconstruct visual content accurately from textual descriptions. To address this limitation, we propose `BiMAC`, a novel framework that enables bidirectional interactions between images and text at both global and local levels. BiMAC employs advanced components to simultaneously reconstruct visual content from textual cues and generate textual descriptions guided by visual features. By integrating a text-region alignment mechanism, BiMAC identifies and selects relevant image patches for precise cross-modal interaction, reducing information noise and enhancing mapping accuracy. BiMAC achieves state-of-the-art performance across diverse vision-language tasks, including image-text retrieval, captioning, and classification.

Introduction

Research on the interaction between language and vision has progressed remarkably in recent years, focusing on tasks that require the accurate integration of textual and visual features (You et al. 2024; Zareapoor, Shamsolmoali, and Lu 2024; Chen et al. 2023). For instance, tasks such as image captioning require generating coherent and contextually accurate textual descriptions of visual content. Similarly, image-text retrieval involves matching images with their corresponding textual descriptions or vice versa, without clear boundaries between the two modalities. These tasks demand sophisticated multimodal alignment techniques capable of effectively integrating and interpreting both visual and textual information. Advanced models in this domain utilize a contrastive learning objective to strengthen global representations between modalities. In this approach, models are trained to ensure that similar items from different modalities, such as a specific image and corresponding caption, are positioned closely in the representation space, while dissimilar items are pushed apart. Examples of such models include CLIP (Radford et al. 2021), mPLUG (Xu et al. 2023), and ConVQG (Mi et al. 2024), which have achieved significant improvements in vision-language tasks. BEiT-v3 (Wang et al. 2023a) uses masking techniques to predict missing components in images, treating images as a form of language. However, its reliance on task-specific fine-tuning poses a challenge, as it reduces its flexibility to be directly applied across various tasks without additional training.

One notable model is Contrastive Captioners (CoCa), which combines contrastive learning with image captioning techniques (Yu et al. 2022), producing a pretrained model effective in both retrieval and captioning tasks. However, CoCa only integrates visual cues to generate textual descriptions, it does not utilize the visual context for image reconstruction from textual cues. This one-dimensional approach restricts the model's understanding of multimodal relationships. Recent vision pre-training methods (Xie et al. 2022; He et al. 2022; Bica et al. 2024; Ma et al. 2024) have shown that image reconstruction can lead to strong content representations. By applying this principle to multimodal tasks (where textual information is integrated into the image reconstruction process and local interactions are emphasized), text and image representations can be merged into a single space, thereby enabling more precise and meaningful bidirectional interactions. For instance, Ma et al. (2024) proposed text-guided masked image modeling to construct visual features from textual guidance, addressing the balance between global and local interactions in multimodal tasks.

Building on these, we introduce BiMAC, a simple yet effective framework designed to enhance alignment between image and text data. BiMAC leverages image-to-text generation for summarizing visual data into textual descriptions, and text-to-image reconstruction ensures that these textual summaries retain sufficient information to reconstruct the original image. The contrastive learning objective reinforces alignment by encouraging paired image-text embeddings to be close in a shared latent space, while unrelated pairs are pushed apart. However, achieving effective multimodal alignment requires addressing a critical challenge: while images contain dense, detailed information, text descriptions are often sparse and focus on salient elements (i.e., textual descriptions often omit significant details that are present in images). This creates a mismatch in the level of detail between the two modalities. To bridge this gap, our model identifies the most relevant image patches for align-

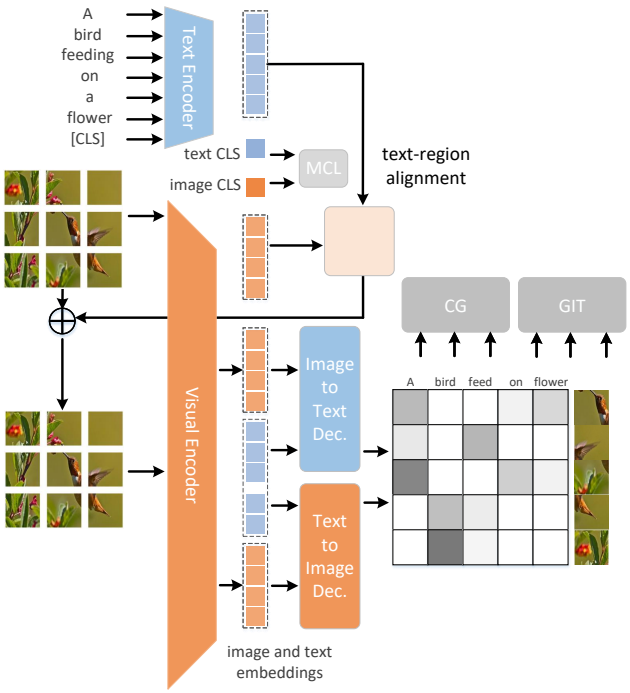

Figure 1: Our multimodal (vision-language) model is structured with four components: encoders for both text and image inputs, along with decoders that handle text-to-image and image-to-text transformations. The model is built upon multiple training objectives: multi-pair contrastive learning (MCL), vision-to-language mapping (caption generation), image reconstruction (GIT), and text-region alignment.

ment with textual descriptions while prioritizing less semantically salient regions (e.g., background details that are often omitted in text) for text-to-image reconstruction. This dual focus ensures that the model captures both the high-level semantics and fine-grained visual details necessary for multimodal alignment. Specifically, we achieve this fine-grained alignment using a cross-modal entropy mechanism, where text tokens attend to image patches based on learned similarity scores. Patches with the highest similarity scores are selected for textual alignment, ensuring that the model prioritizes the most semantically relevant regions.

Unlike global alignment methods that treat images and text as monolithic entities, our fine-grained alignment approach captures local correspondences between tokens and patches, which is particularly important for aligning concise textual descriptions (e.g., a birthday party) with dense visual scenes, where multiple elements such as objects and actions must be jointly considered. Extensive experiments demonstrate the superiority of our model over state-of-the-art models in various vision-language tasks.

Related Work

Multimodal Representation Learning. In recent years, significant advances have been made in multimodal alignment, particularly in integrating vision and language (Cao et al. 2024; Hu et al. 2024; Mi et al. 2024; You et al. 2024). Traditional approaches often relied on separate modules for each modality, such as object detection frameworks to extract visual features and natural language processing to process text. For example, works by Chen et al. (2020) and Zhang et al. (2021) utilized pre-trained object detection models to align visual representations with corresponding textual features, facilitating the fusion of these modalities.

Subsequent research (Kim et al. 2022; Bao et al. 2022b) shifted towards developing multimodal transformers that are trained from scratch to jointly learn from both visual and textual inputs. This evolution led to the introduction of large-scale vision-language models like CLIP and ALIGN (Radford et al. 2021; Jia et al. 2021), which exemplify these advancements through their dual-encoder frameworks with contrastive loss. These models achieve efficient cross-modal alignment and enable tasks like zero-shot classification.

Another technique (Yuan et al. 2021) introduced unified contrastive learning to enhance the interaction in image-text benchmarks. CoCa (Yu et al. 2022) has extended this approach by incorporating image captioning via a decoder, enhancing local interactions between visual and textual input. SyCoCa (Ma et al. 2024) improved upon this by employing an attentive masking strategy for image modeling guided by textual information. In contrast, our work introduces a novel approach that integrates caption generation and image generation tasks to reinforce bidirectional interactions between modalities. This method specifically emphasizes fine-grained alignment by focusing on image patches paired with detailed textual descriptions, addressing the inherent abstractness of image captions. By targeting fine-grained alignment, our model sets itself apart from other bidirectional generation methods that rely on discrete auto-encoders for generating images from text.

Masked Modeling in Language and Vision. Masked modeling has become a cornerstone in both language and vision domains, serving as a powerful pre-training strategy for learning rich representations. In language, masked strategy has been widely adopted, where models are trained to predict masked tokens in a text sequence (Yang et al. 2022; Park and Han 2023). Similarly, in the vision domain, masked image modeling (MIM) has gained significant traction, where models are trained to predict or reconstruct masked portions of an image, often guided by textual descriptions (Wang et al. 2023a; You et al. 2024; Peng et al. 2023; Wang et al. 2023b). This approach has been pivotal in vision transformers (Dosovitskiy et al. 2021; Bao et al. 2022a; Peng et al. 2023), which use strategies like mean color prediction, converting image patches into discrete tokens via a VAE network, and pixel clustering, to enhance representation learning. In the context of multimodal learning, masking strategy (MIM) has been enriched through joint representation learning. MAMO (Zhao et al. 2023), MaskVLM (Kwon et al. 2023), SyCoCa (Ma et al. 2024), BEiT (Wang et al. 2023a; Peng et al. 2023), SPARC (Bica et al. 2024) and EVE (Chen et al. 2023) have integrated MIM into multimodal pre-training, where the models are trained to predict randomly masked image patches or text tokens. These ad-

vancements have significantly contributed to the refinement of cross-modal alignment and representation learning. Unlike existing models that rely heavily on masking strategies, which often introduce challenges for achieving effective bidirectional learning, our model presents a simple yet novel approach to facilitate bidirectional interactions without the need for masking. Indeed, masking, while useful for occluding parts of the input during training, inherently biases the model towards reconstructive tasks, rather than fostering a true bidirectional exchange between modalities (Tou and Sun 2024). Our proposed model eliminates the limitations of masking-based methods, enabling more efficient and precise integration of visual and textual information.

Proposed BiMAC

We propose BiMAC, which uses bidirectional interactions to generate detailed textual and visual representations within a unified latent space. While CoCa (Yu et al. 2022) has made significant progress by combining image captioning with a contrastive objective, it primarily focuses on generating text from images, limiting its capacity to utilize textual information for reconstructing visual content. This unidirectional interaction restricts the potential for comprehensive vision-language alignment. To address this, BiMAC enhances bidirectional alignment through several key objectives (as detailed in Algorithm 1): Multi-Pair Contrastive Learning (MCL) enhances the alignment between visual and textual representations; Caption Generation (CG) focuses on generating textual descriptions based on image content; and Generating Image from Text (GIT) facilitates precise connections between specific visual elements and textual input.

Architecture Overview. Our architecture is shown in Figure 1. The image encoder E_{img} processes the input I and generates patch embeddings $\mathbf{F}_i$ using vision transformer. Each embedding corresponds to a specific image patch and an additional [CLS] token for a global image representation

$$E_{\text{img}}(I) = \{f_1, f_2 \ldots, f_P, f^{\text{cls}}\} \tag{1}$$

with P number of patches. For a caption C, nouns are extracted and embedded using the language encoder E_{txt}

$$E_{\text{text}}(T) = \{\mathbf{w}_1, \mathbf{w}_2 \ldots, \mathbf{w}_W\}, \tag{2}$$

where W is the number of nouns in the caption. To enhance interactions between image and text representations, BiMAC employs a bidirectional multimodal framework with image and text decoders, each utilizing cross-attention transformers, to merge information from both modalities. This enables two complementary modes of interaction: i) Text-to-Image Guidance, where text features guide the visual representation by focusing on relevant regions or features in the image. ii) Image-to-Text Guidance, where image features refine text interpretation by emphasizing salient words or phrases aligned with visual content. Both modes rely on cross-attention mechanisms, where one modality (e.g., image) serves as the query and the other modality (e.g., text) provides the key and value, facilitating interaction between image and text representations. This ensures a comprehensive mutual understanding of the image-text pairs. For example, the image decoder benefits from text features, and the text decoder integrates image features to generate contextually aligned and visually grounded captions.

Algorithm 1: BiMAC: Bidirectional Multimodal Alignment

Require: Paired image-text dataset (I_i, T_i)
Ensure: Multimodal representations, vision-to-language mapping (caption generation), and reconstructed images
Step 1: Encode Image and Text
for each (I_i, T_i) in the batch **do**
  $F_i \leftarrow E_{\text{img}}(I_i)$ ▷ Visual embeddings from image encoder
  $T_i \leftarrow E_{\text{txt}}(T_i)$ ▷ Textual embeddings from text encoder
end for
Step 2: Multi-Pair Contrastive Learning
$L_{\text{MCL}} \leftarrow$ Compute contrastive loss between F_i^{cls} and T_i^{cls}
Step 3: Vision-to-Language Mapping (CG)
$L_{\text{CG}} \leftarrow$ Compute caption generation loss using text decoder D_{txt}
Step 4: Visual Reconstruction Form Text
Reconstruct patches $\hat{r}_i \leftarrow D_{\text{img}}(F_i^{\text{masked}}, T_i)$
$L_{\text{GIT}} \leftarrow$ Compute reconstruction loss
Step 5: Fine-Grained Text-Region Alignment
$S \leftarrow W_i F_i^{\top}$ ▷ Compute alignment scores between text tokens and image patches using bipartite matching
Step 6: Compute Total Loss
$L_{\text{total}} \leftarrow L_{\text{MCL}} + \lambda_{\text{CG}} L_{\text{CG}} + \lambda_{\text{GIT}} L_{\text{GIT}}$
Optimize model parameters using L_{total}

Learning from Image-text Pairs

To train the BiMAC model, we utilize a dataset comprising paired images and their corresponding textual descriptions, denoted as (I_i, T_i). The image encoder E_{img} takes the input image I to generate visual embeddings (as defined in Eq. 1), and the text encoder E_{txt} encodes the textual description T to produce textual embeddings (as defined in Eq. 2). To align these visual and textual embeddings, we employ a multi-pair contrastive learning (MCL), which forms a key component of the model's bidirectional alignment mechanism.

Multi-Pair Contrastive Learning (MCL): The MCL loss consists of two terms: an image-to-text alignment and a text-to-image alignment loss. Together, these components aim to maximize the similarity of positive pairs (a caption and its corresponding image) while minimizing the similarity of negative pairs (unrelated captions and images). For a batch of B image-text pairs, the image-to-text alignment loss is

$$L_{IT} = -\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)}, \tag{3}$$

where s_{ii} is similarity score between the i-th image embedding (f_i) and its paired text embedding w_i, and s_{ij} is similarity score between the i-th image embedding (f_i) and unrelated text embeddings (w_j). This ensures that each image embedding f_i aligns closely with its paired text embedding w_i, while being dissimilar to unpaired text embeddings.

Similarly, the text-to-image alignment loss is defined as

$$L_{TI} = -\log \frac{\exp(s_{jj}/\tau)}{\sum_{i=1}^{B} \exp(s_{ij}/\tau)} \quad (4)$$

This ensures that each text embedding w_j aligns closely with its paired image embedding f_j, while being dissimilar to unpaired image embeddings. The parameter τ is a temperature parameter for sample i with B batch size. By controlling the sharpness of the softmax distribution, we can effectively influence the learning process, leading to improved model performance for interactions between images and text. Then, the final MCL loss is the average of the two alignment losses over the batch B

$$L_{\text{MCL}} = \frac{1}{2B} \sum_{i=1}^{B} (L_{IT} + L_{TI}) \quad (5)$$

This bidirectional formulation ensures that the model's understanding is not one-sided; it reinforces the interaction in both directions (image-to-text and text-to-image), thus improving the overall performance in multimodal tasks.

Vision-to-Language Mapping (caption generation): Caption generation (CG) aims to produce detailed text descriptions T for corresponding images I. Following (Yu et al. 2022; Bica et al. 2024; Ma et al. 2024), our model consists of an image encoder E_{img} and a text decoder D_{txt}. The process involves two main steps: encoding the image into a feature representation and decoding this representation into a sequence of words (caption), as

$$L_{\text{CG}} = -\sum_{i=1}^{B} \sum_{t=1}^{|T_i|} \log P(T_i^t | E(I_i), C_i) \cdot R(T_i, I_i), \quad (6)$$

where $C_i^{<t} = \{T_i^1, T_i^2, \ldots, T_i^{t-1}\}$ represents all previously generated words in the sequence up to step t, B is the number of images in a batch, T_i^t is the t-th word in the generated caption for the i-th image, and $E(I)$ is the image feature representation obtained from the image encoder. In this mapping process, the decoder predicts the next word T_i^t at step t based on the image feature representation and the previously generated words, $P(T_i^t | E(I_i), C_i) = D_{txt}(E(I), C_i)$. The reward function $R(T_i, I_i)$ evaluates the quality of the generated caption, incorporating metrics like CIDEr or BLEU. This function implicitly guides the visual representation by rewarding captions that accurately describe the image regions. Image features from the encoder (E_{img}) directly influence the probability distribution over tokens, ensuring that the generated text aligns with the visual content.

Generating Image Patches from Textual Cues: While visual features are often rich and complex, textual descriptions can be inherently limited in expressiveness. This discrepancy can create challenges in achieving balanced interaction between modalities (He et al. 2022; Ma et al. 2024). To address this, we propose a patch-based reconstruction method where textual cues guide the generation of image patches. Given an image I, its regions are divided into m non-overlapping patches encoded by the image encoder E_{img}, as $R = [r_1, r_2, \ldots, r_m]$. From the associated caption C, all nouns are extracted and embedded using a text encoder E_{txt} as defined by $W = [w_1, w_2, \ldots, w_W]$, where the W is the number of extracted nouns/words. To align textual cues with visual patches, the region (patches)-word assignment task is formulated as a bipartite matching problem (Kuhn 1955). The alignment cost is computed as $S = WR^T$, where S represents the alignment scores between textual tokens and visual patches. This alignment loss is defined as

$$L_{GIT} = -\frac{1}{|T_i|} \sum_{i=1}^{|T_i|} [\log \sigma(s_{ik}^+) + \sum_{j=1}^{W'} \log(1 - \sigma(s_{ij}^-))]. \quad (7)$$

σ is the sigmoid activation, s_{ik} is the alignment score between the i-th word embedding and the k-th region feature, and W' represents a set of negative samples (e.g., nouns from unrelated captions in the batch). Negative samples discourage spurious alignments, further improving the model's grounding capability. This loss ensures that key objects in the image are explicitly reflected in the caption by aligning visual regions with their most semantically relevant textual tokens. Using the grounded text embeddings W_g (resulting from the grounding process), a patch reconstruction decoder D_p reconstructs each patch as: $\hat{r}_j = D_p(w_j)$, where w_j is the grounded text embedding for the j-th patch. The reconstructed patches $\{\hat{r}_1, \cdots, \hat{r}_m\}$ are then assembled into a complete image $\hat{I}$. To ensure accurate reconstruction, we minimize the pixel-wise difference between the original and reconstructed patches using the reconstruction loss $L_{rec} = \frac{1}{m} \sum_{j=1}^{m} \|r_j - \hat{r}_j\|_1$, with $\|.\|_1$ as the L1 norm. This fine-grained interaction enables a more detailed and interpretable connection between visual and textual modalities.

Training Objective: The training objective is defined by integrating the above losses into a unified objective function.

$$L_{total} = \begin{cases} L_{\text{MCL}}, & \text{for image-text pairs} \\ \lambda_{\text{CG}} L_{\text{CG}}, & \text{caption generation from image cues} \\ \lambda_{\text{GIT}} L_{\text{GIT}} & \text{image generation from textual cues} \end{cases} \quad (8)$$

In summary, L_{MCL} ensures that image-text pairs are well-aligned in a shared semantic space. L_{CG} guides the model to generate detailed captions for images, leveraging both visual features and linguistic context. L_{GIT} establishes fine-grained cross-modal interaction, aligning specific image regions with corresponding textual tokens, enabling effective image reconstruction from textual cues. To improve training efficiency for image-text pairs and image captioning, a low-resolution input (256^2) with a batch size of 1024 is adapted.

Experimental Results

To ensure fairness and consistency in our experiments, we adopt the CoCa framework as the base architecture, configuring our model with CoCa-Base settings for the text encoder/decoder and the image encoder. Different from CoCa we also have an image decoder D_p that is similar to the text decoder but focuses on region-specific features rather than single-step pixel prediction. The Adam optimizer with an initial learning rate of 2×10^{-4} combined with a cosine decay schedule. The model was trained over 30 epochs using

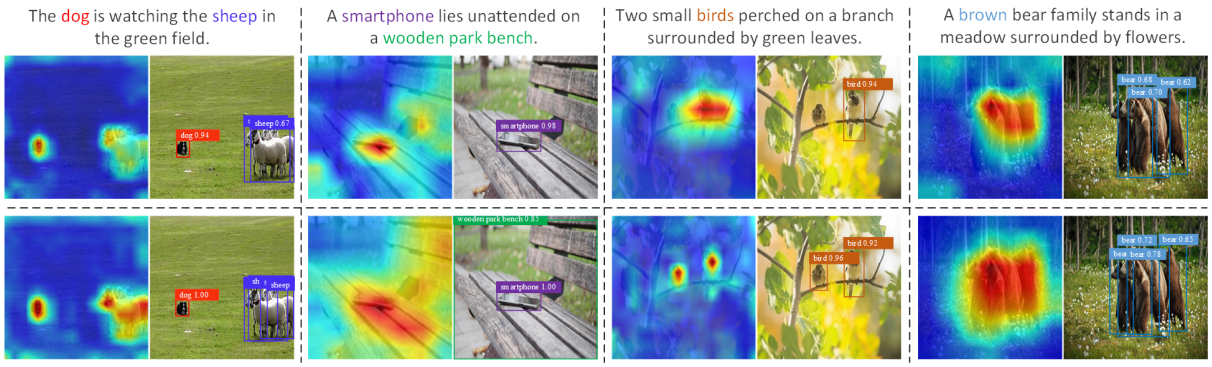

Figure 2: Visual comparison between CoCa (top row) and BiMAC (bottom row) in processing image-caption pairs. BiMAC demonstrates superior alignment with captions by accurately focusing on relevant regions of the images. For example, in the leftmost image, BiMAC distinctly highlights both the dog and the sheep in a green field, aligning precisely with the caption. Similar improvements are evident across other examples, showing BiMAC's capability in fine-grained region-word alignment.

Model	Flickr30K (1K test set)						COCO (5K test set)					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
BEiT-v3 (Peng et al. 2023)	45.3	73.8	83.0	34.2	60.4	72.1	-	-	-	-	-	-
CoBIT (You et al. 2024)	43.9	73.2	-	32.5	60.5	-	16.3	37.5	-	15.9	36.5	-
ALIGN (Jia et al. 2021)	42.3	71.9	81.4	30.9	58.6	71.3	15.7	36.0	48.5	14.6	34.2	46.4
FILIP (Yao et al. 2022)	42.9	72.6	81.8	30.2	58.3	71.2	16.1	36.7	49.1	15.1	34.9	46.9
Florence (Yuan et al. 2021)	43.5	72.4	-	31.4	58.7	-	17.2	37.8	-	15.4	36.7	-
SyCoCa (Ma et al. 2024)	45.9	76.3	83.5	35.9	64.5	75.3	18.5	40.7	53.2	16.9	39.7	51.6
SPARC (Bica et al. 2024)	45.4	76.1	82.9	36.2	64.9	76.1	18.7	41.2	52.8	16.5	39.5	51.8
CoCa (Yu et al. 2022)	43.7	72.9	82.1	32.4	60.7	71.6	16.9	37.6	49.3	16.1	36.2	47.1
BiMAC	46.8	77.2	84.3	36.8	65.1	76.5	19.4	41.6	53.8	17.5	40.8	52.0
Improv. (%)	+7.1	+5.9	+2.7	+13.6	+7.4	+6.9	+14.8	+10.6	+9.1	+8.7	+12.7	+11.0

Table 1: Zero-shot result of image-to-text & text-to-image retrieval on Flickr (Plummer et al. 2015) and COCO (Lin et al. 2014).

8 RTX 3090 GPUs, with a batch size of 1024 and an image resolution of 256×256. The coefficients were set to $\lambda_{\text{CG}} = 1$ in line with CoCa, $\lambda_{\text{GIT}} = 0.7$ and the temperature $\tau = 0.5$.

Main Results

Pretraining Datasets. The pretraining process uses two versions of conceptual captions datasets, i.e., CC3M (Sharma et al. 2018) and CC12M (Changpinyo et al. 2021) (together denoted as CC15M), which, after filtering out invalid URLs, consists of 13M image-text pairs. This dataset is a standard benchmark for vision-language pretraining evaluations and has been employed in several recent studies (Tian et al. 2024; Wang et al. 2023a; Yao et al. 2023). We evaluate BiMAC on several tasks, including image-text retrieval, captioning, zero-shot, and fine-tuned image classification. Details of additional datasets used for these evaluations are provided in the supplementary file. Figure 2 presents a visual comparison between BiMAC (bottom row) and CoCa (top row). Our model, BiMAC, consistently outperforms CoCa by producing more targeted and precise attention maps that focus on the regions of the images most relevant to the captions. For instance, in the third pair (birds perched on a branch), BiMAC sharply focuses on the birds, distinguishing them from the surrounding leaves. In contrast, CoCa's attention is more diffuse, spreading across the branch and the leaves, which dilutes its effectiveness in identifying the specific subjects mentioned in the caption. This shows that BiMAC is more effective at identifying the specific subjects mentioned in the caption, especially in scenes where objects are closely packed. Overall, BiMAC shows a more refined focus on the critical elements described in the captions, where CoCa's attention is less precise and more dispersed.

Zero-shot Image-Text Retrieval: The models are trained to find the most relevant sample corresponding to a specific input across different modalities, using the Flickr30K and MSCOCO datasets. As shown in Table 1, BiMAC outperforms the CoCa model, achieving improvements in Recall@1 ranging from 7% to 15%. This gain highlights the cross-modal alignment in BiMAC, enabling more accurate matching between visual content and textual queries.

Zero-shot Image Captioning: In this experiment, the models generate captions that accurately describe the content of images. Following (Wang et al. 2023b), we fine-tune the models on the COCO Captions dataset (Lin et al. 2014) and

	COCO				NoCaps			
	BLEU@4	METEOR	CIDEr	SPICE	BLEU@4	METEOR	CIDEr	SPICE
BLIP (Li et al. 2022)	21.4	-	70.6	-	12.6	21.5	-	-
VinVL (Zhang et al. 2021)	19.5	20.9	68.2	16.2	-	-	49.4	10.1
SyCoCa (Ma et al. 2024)	22.4	23.6	75.6	16.8	12.9	22.8	54.8	10.5
SimVLM (Wang et al. 2022)	21.7	23.1	72.3	17.1	12.3	-	53.7	10.4
ViLTA (Wang et al. 2023b)	22.6	22.4	74.6	16.2	-	-	54.9	10.7
CoCa (Yu et al. 2022)	21.9	22.7	71.8	16.5	12.8	22.7	54.1	10.4
BiMAC	23.5	24.1	75.9	17.3	13.2	23.4	55.3	10.9
Improv. (%)	+7.3	+6.2	+5.7	+4.8	+3.1	+3.0	+2.3	+4.8

Table 2: Results on image captioning on NoCaps and COCO Caption. We report BLEU@4, METEOR, CIDEr, and SPICE scores on the Karpathy test split. The last row shows the performance improvement of our model over CoCa.

Model	INet-1K	DTD	Flowers	Food-101	CIFAR10	CIFAR100	Caltech	SUN397	Avg.
StableRep (Tian et al. 2024)	29.4	61.5	72.1	72.4	67.2	32.6	66.8	54.3	56.9
CoCa (Yu et al. 2022)	28.2	59.0	69.3	69.7	66.3	31.5	65.1	52.6	55.2
BiMAC	31.6	61.9	70.5	72.9	68.5	33.7	67.5	54.7	58.4
Improv. (%)	+12.0	+4.9	+1.7	+4.5	+3.3	+6.9	+3.6	+3.9	+5.8

Table 3: Zero-shot classification accuracy (%) on various datasets: ImageNet–1K (Deng et al. 2009), DTD (Cimpoi et al. 2014), Flowers (Nilsback and Zisserman 2008), and Food101 (Bossard, Guillaumin, and Van Gool 2014), as well as CIFAR10 and CIFAR100 (Krizhevsky, Hinton et al. 2010), SUN397 (Xiao et al. 2010), and Caltech (Fei-Fei, Fergus, and Perona 2006). The models are pre-trained on the CC15M. BiMAC outperforms CoCa across most datasets with an average improvement of 6.0%.

evaluate their performance using metrics: BLEU@4, METEOR, CIDEr, and SPICE. As shown in Table 2, BiMAC consistently outperforms all baselines, achieving improvements over CoCa ranging from 4% to 7%. We further evaluate the models on the NoCaps dataset (Agrawal et al. 2019) in a zero-shot setting, without any additional fine-tuning. BiMAC again surpasses CoCa, with improvements ranging from 2% to 4%. These results highlight the importance of bidirectional interaction in BiMAC, which enhances modality alignment and leads to more effective caption generation.

Image Classification Results: The zero-shot classification performance of BiMAC and CoCa is evaluated across eight datasets, which include various coarse-grained categories like flowers and foods. As shown in Table 3, BiMAC establishes new state-of-the-art results in zero-shot classification on ImageNet, with an average accuracy improvement of 12% over CoCa. For fine-grained classification tasks, we evaluated BiMAC against CoCa and other vision-language models on three challenging datasets: SPEC (Peng et al. 2024), VALSE (Parcalabescu et al. 2021), and Winoground (Thrush et al. 2022). These benchmarks are designed to test visual fine-grained understanding and are particularly challenging due to the presence of hard negatives and complex multimodal interactions. As detailed in Table 4, BiMAC significantly outperforms CoCa and other baselines across all benchmarks. On the VALSE benchmark, BiMAC improves accuracy from 69.5% (CoCa) to 74.1%, marking a relative improvement of +6.5%. Similarly, on the SPEC dataset, BiMAC achieves notable gains in both image-to-text (I2T) and text-to-image (T2I) retrieval tasks, with improvements of +16% and +11%, respectively. On the Winoground, which emphasizes contextual and semantic alignment, BiMAC shows substantial gains, achieving a +33% improvement in image understanding (I) and +12% in text understanding (T). Notably, VisMin-CLIP (Awal et al. 2024) performs better than most other baselines; however, BiMAC achieves superior results in most cases, demonstrating its effectiveness in handling challenging multimodal tasks.

Method	SPEC		VALSE	Winoground		
	I2T	T2I		T	I	G
CLIP	33.7	30.4	68.2	29.2	10.6	8.4
FLAVA	30.1	29.8	-	25.7	12.8	9.0
BLIP	35.3	32.7	-	29.6	10.2	7.5
VisMin-CLIP	44.2	39.8	72.2	32.8	14.7	9.6
ViLT	36.5	32.9	70.6	33.9	14.0	9.3
CoCa	34.3	32.5	69.5	30.7	11.2	8.7
BiMAC	39.7	36.2	74.1	34.5	14.9	9.8
Improv. (%)	+15.7	+11.4	+6.5	+12.4	+33.0	+12.6

Table 4: Fine-grained classification on the challenging SPEC, VALSE, and Winoground benchmarks. The last row is the percentage improvement of BiMAC over CoCa.

Ablation Study

To evaluate the effectiveness of specific training objectives within BiMAC, we performed comparative experiments. The models are trained on the CC3M+CC12M datasets (re-

	MCL	CG	RF	GIT	Grounding	Flickr30K		MSCOCO		ImageNet-1K	CIFAR10	Caltech	Avg.
						mTR	mIR	mTR	mIR				
	×					60.7	51.4	31.7	30.5	29.1	67.4	64.8	47.9
CoCa	×	×				**65.9**	**54.7**	**34.5**	**33.2**	**28.2**	**66.3**	**65.1**	**49.7**
	×	×	×			67.6	56.8	36.1	35.4	29.7	67.2	66.5	51.3
	×	×	×	×		68.7	58.6	37.5	35.8	30.6	67.9	67.4	52.4
	×	×		×	×	68.9	59.1	37.8	36.3	31.2	68.4	67.5	52.5
Ours	×	×	×	×	×	**69.4**	**59.5**	**38.4**	**36.9**	**31.6**	**68.5**	**67.5**	**53.1**

Table 5: Ablation study shows the impact of different combinations of training objectives on model performance. MCL: multi-pair contrastive learning, CG: caption generation, RF: reward function applying in caption generation, GIT: image generation from textual cues, and grounding is applied on the image generation step. The metrics mIR and mTR represent the mean retrieval accuracy across the top 1, 5, and 10 recall. We evaluate zero-shot retrieval performance on the MSCOCO and Flickr30K datasets, as well as classification accuracy on ImageNet-1K, CIFAR10, and Caltech. Models are pre-trained on the CC15M.

ferred to as CC15M), which, after filtering out invalid URLs, consist of 13 million image-text pairs. This dataset is a standard benchmark for vision-language pretraining evaluations and has been employed in several recent studies (Yao et al. 2023; Tian et al. 2024). The results are summarized in Table 5, presenting the following key findings.

When using only the MCL objective (row-1), the model shows relatively low performance on the retrieval task; for example, the mean retrieval accuracies (mTR/mIR) are 60.7%/51.4% on Flickr30K, which falls short compared to CoCa and BiMAC. This reflects the limitation of relying solely on contrastive learning for cross-modal understanding. Adding the caption generation (CG) objective (row-2) significantly improves performance. For instance, on Flickr30K, the retrieval accuracy increases to 65.9% (mTR) and 54.7% (mIR), and classification tasks also observe consistent gains. This highlights the importance of incorporating caption generation for better alignment between modalities. Integrating the reward function into caption generation (row-3) further enhances performance, particularly in zero-shot retrieval tasks. The introduction of GIT (row-4) boosts performance by adding bidirectional interactions, particularly in tasks requiring textual information to guide visual reconstruction. The average performance increases significantly across both retrieval and classification tasks.

Adding the grounding strategy (row-5) results in notable improvements, especially in zero-shot retrieval. For example, on Flickr30K, the model achieves 68.9% (mTR) and 59.1% (mIR), showing the critical role of grounding in aligning textual tokens with image patches for enhanced fine-grained understanding. With all objectives integrated (MCL, CG, RF, GIT, and Grounding) in BiMAC (row 6), the model achieves state-of-the-art performance across tasks, with an average improvement of 15% over CoCa. While the reward function has a modest impact, it complements caption generation by refining the quality of generated descriptions. Using GIT without adding the grounding (row-5) leads to lower performance. We also explored the effects of the weights for the image generation through the textual cues. Specifically, we adjusted the λ_{GIT} within the range of $[0.1, 1.5]$, and results presented in Table 6 show that changes in λ_{GIT} have a negligible impact on overall performance.

	Flickr30K		MSCOCO		ImageNet	CIFAR10	Caltech
λ_{GIT}	mTR	mIR	mTR	mIR			
0.1	69.1	59.1	37.9	35.8	30.5	65.7	67.4
0.4	**69.5**	59.4	38.2	36.5	31.4	66.9	**67.8**
0.7	69.4	**59.5**	**38.4**	**36.9**	31.6	68.5	67.5
1.0	69.2	**59.5**	38.1	36.7	**31.8**	**68.9**	67.1
1.5	68.7	58.2	37.3	35.4	30.7	67.1	66.5

Table 6: Ablation experiments on different values of λ_{GIT}. For the COCO and Flickr30K, we report mTR (mean Text Retrieval) and mIR (mean Image Retrieval) metrics, while for the ImageNet, CIFAR10, and Caltech datasets, we report classification accuracy. The models are pre-trained on the CC15M dataset. The results indicate that the optimal performance is achieved at $\lambda = 0.7$, however, increasing beyond 1.0 leads to a significant decline in performance.

Consequently, we selected $\lambda = 0.7$ for the final training, as it provided the best balance between the different objectives.

Conclusion

We introduced BiMAC, a novel vision-language pretraining framework that establishes state-of-the-art performance across a diverse range of downstream tasks. Unlike traditional models that rely on masking strategies or unidirectional generation, BiMAC introduces a simple yet effective architecture that incorporates contrastive learning, caption generation, and image reconstruction to foster robust multimodal understanding. Central to our approach is the use of fine-grained text-region alignment, which identifies and aligns the most relevant image patches with their corresponding textual descriptions. This targeted alignment ensures that the model captures meaningful cross-modal connections, enhancing its capability to perform tasks requiring a nuanced understanding of image-text relationships. The effectiveness of BiMAC is validated through comprehensive experiments on different vision-language benchmarks, demonstrating its strong generalization capabilities.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957.

Awal, R.; Ahmadi, S.; Zhang, L.; and Agrawal, A. 2024. VisMin: Visual Minimal-Change Understanding. *Advances in Neural Information Processing Systems*.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022a. Beit: Bert pre-training of image transformers. *International Conference on Learning Representations*.

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022b. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35: 32897–32912.

Bica, I.; Ilić, A.; Bauer, M.; Erdogan, G.; Bošnjak, M.; Kaplanis, C.; Gritsenko, A. A.; Minderer, M.; Blundell, C.; Pascanu, R.; et al. 2024. Improving fine-grained understanding in image-text pre-training. *ICML*.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *European conference Computer vision*, 446–461.

Cao, B.; Guo, J.; Zhu, P.; and Hu, Q. 2024. Bi-directional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 927–935.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3558–3568.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020. Generative pretraining from pixels. In *International conference on machine learning*, 1691–1703.

Chen, Y.; Yuan, J.; Tian, Y.; Geng, S.; Li, X.; Zhou, D.; Metaxas, D. N.; and Yang, H. 2023. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15095–15104.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4): 594–611.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; and Tu, Z. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2256–2264.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916.

Kim, T.; Song, G.; Lee, S.; Kim, S.; Seo, Y.; Lee, S.; Kim, S. H.; Lee, H.; and Bae, K. 2022. L-verse: Bidirectional generation between image and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16526–16536.

Krizhevsky, A.; Hinton, G.; et al. 2010. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7): 1–9.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Kwon, G.; Cai, Z.; Ravichandran, A.; Bas, E.; Bhotika, R.; and Soatto, S. 2023. Masked vision and language modeling for multi-modal representation learning. *International Conference on Learning Representations*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference Computer Vision*, 740–755.

Ma, Z.; Xu, F.; Liu, J.; Yang, M.; and Guo, Q. 2024. SyCoCa: Symmetrizing Contrastive Captioners with Attentive Masking for Multimodal Alignment. *arXiv preprint arXiv:2401.02137*.

Mi, L.; Montariol, S.; Navarro, J. C.; Dai, X.; Bosselut, A.; and Tuia, D. 2024. ConVQG: Contrastive Visual Question Generation with Multimodal Guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4207–4215.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Sixth Indian conference on computer vision, graphics & image processing*, 722–729.

Parcalabescu, L.; Cafagna, M.; Muradjan, L.; Frank, A.; Calixto, I.; and Gatt, A. 2021. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Park, J.; and Han, B. 2023. Multi-modal representation learning with text-driven soft masks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2798–2807.

Peng, W.; Xie, S.; You, Z.; Lan, S.; and Wu, Z. 2024. Synthesize Diagnose and Optimize: Towards Fine-Grained Vision-Language Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13279–13288.

Peng, Z.; Dong, L.; Bao, H.; Ye, Q.; and Wei, F. 2023. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *International Conference on Learning Representations*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2556–2565.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248.

Tian, Y.; Fan, L.; Isola, P.; Chang, H.; and Krishnan, D. 2024. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36.

Tou, K.; and Sun, Z. 2024. Curriculum Masking in Vision-Language Pretraining to Maximize Cross Modal Interaction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3672–3688.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. 2023a. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19175–19186.

Wang, W.; Yang, Z.; Xu, B.; Li, J.; and Sun, Y. 2023b. ViLTA: Enhancing vision-language pre-training through textual augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3158–3169.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2022. Simvlm: Simple visual language model pretraining with weak supervision. *International Conference on Learning Representations*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9653–9663.

Xu, H.; Ye, Q.; Yan, M.; Shi, Y.; Ye, J.; Xu, Y.; Li, C.; Bi, B.; Qian, Q.; Wang, W.; et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, 38728–38748.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15671–15680.

Yao, L.; Han, J.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; and Xu, H. 2023. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23497–23506.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2022. Filip: Fine-grained interactive language-image pre-training. *International Conference on Learning Representations*.

You, H.; Guo, M.; Wang, Z.; Chang, K.-W.; Baldridge, J.; and Yu, J. 2024. Cobit: A contrastive bi-directional image-text generation model. *International Conference on Learning Representations*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zareapoor, M.; Shamsolmoali, P.; and Lu, Y. 2024. Learning Region-Word Alignment with Attentive Masking for Open-Vocabulary Object Detection. In *NeurIPS 2024 Workshop on Open-World Agents*.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5579–5588.

Zhao, Z.; Guo, L.; He, X.; Shao, S.; Yuan, Z.; and Liu, J. 2023. Mamo: Fine-grained vision-language representations learning with masked multimodal modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1528–1538.