

CoV-Align: Efficient Fine-grained Cross-Modal Alignment with Cohesive Visual Semantics Priority

Hengqi Liu^{1,2}, Wanting Zhou^{1*}, Longteng Kong³, Fangxiang Feng¹, Lei Ren², Wei Chen², Xiaojie Wang¹

¹Beijing University of Posts and Telecommunications, ²Li Auto, ³Beijing Forestry University

{liuhengqi, wanting.zhou}@bupt.edu.cn {renlei3}@lixiang.com {konglongteng}@bjfu.edu.cn

Abstract

Cross-modal alignment aims to learn semantically consistent latent representations across diverse modalities. Prevaling methods rely on a text-guided aggregation paradigm to achieve fine-grained alignment, while they suffer from redundant patch-word correlations and high computational costs. To address these issues, we propose CoV-Align, an effective and efficient fine-grained cross-modal alignment framework with cohesive visual semantics priority. Through a semantically convergent attention mechanism, it progressively aggregates meaningful visual patches in a text-free manner. We design a coarse visual semantic feature extractor that integrates deformable attention and consistent assign attention to group patches with semantic consistency. A cohesive and discriminative feature optimization is presented to enhance intra-semantic cohesion and inter-semantic discriminability of visual region features, resulting in explicit improvements in cross-modal alignment. Extensive experiments demonstrate that CoV-Align achieves state-of-the-art performance on the Flickr30K and MSCOCO benchmarks. Notably, it delivers a 3–5× computational speedup compared to pioneer approaches, offering compelling advantages for large-scale multi-modal tasks.

1. Introduction

Cross-modal alignment, which aims to bridge the semantic gap between diverse modalities, is a key problem in many multi-modal tasks, including image captioning and visual question answering. Existing solutions can be categorized into two main paradigms: coarse-grained alignment [8, 24] and fine-grained alignment [16, 23]. The former, as depicted in Figure 1 (a), focuses on the global scene and major objects, failing to capture the semantic details [7, 11, 20, 30]. The latter tends to establish correspondences between local visual regions and specific words, which is essential for improving the interpretability and accuracy of

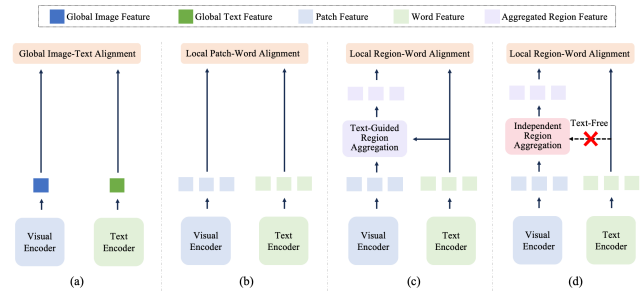


Figure 1. Paradigms of Cross-Modal Alignment. (a) Global Image-Text Alignment; (b) Local Patch-Word Alignment; (c) Text-Guided Region-Word Alignment; (d) Region-Word Alignment with Independent Region Aggregation (Ours).

vision-language models. In fine-grained alignment, object detection based approaches were initially adopted [16, 23], but they are constrained by the fixed category divisions of the detector, making them unsuitable for open-domain downstream tasks. Subsequently, Transformer-based methods [7, 30] selectively associate visual patches with corresponding text through attention mechanisms, as shown in Figure 1 (b). However, individual visual patches usually lack complete semantic information, resulting in ambiguous word matching [10]. To address this issue, recent methods [1, 10] adopt text-guided cross-attention strategy, as illustrated in Figure 1 (c). Resorting to the text embeddings, the visual patches are aggregated at the semantic level, reporting state-of-the-art alignment performance.

However, the text-guided aggregation paradigm computes alignment weights between each text query and all visual patches in the image, leading to two problems. 1) Redundant patch-word alignments. It inevitably introduces noise from semantically irrelevant patches, diluting critical semantic information and compromising retrieval accuracy. 2) High computational cost. For large-scale retrieval with millions of images, this repeated aggregation leads to substantial latency and memory consumption, limiting practical deployment. In fact, the key issue underlying those problems is the involvement of text information dur-

*Wanting Zhou is the corresponding author.

ing patch aggregation. If meaningful visual semantics, such as the “hand”, “pushing”, and “cart” regions in an image of “a person pushing a cart”, could be linked in advance without the text guidance, other irrelevant visual regions would not need to participate in cross-attention computation, as in Figure 1 (d).

In this paper, we propose CoV-Align, an effective and efficient fine-grained cross-modal alignment framework with Cohesive Visual Semantics Priority. Unlike existing methods that introduce text to aggregate visual patches before alignment, CoV-Align originally adopts text-free aggregation paradigm. Precise region generation with a semantically convergent attention improves alignment accuracy, while text-independent aggregation reduces computational complexity. First, we design a coarse visual semantic feature extractor that integrates deformable attention and consistent assign attention with dynamic mask. This enables coarse grouping of semantically consistent patches independently without text guidance. Second, we present a cohesive and discriminative feature optimization method to enhance intra-semantic cohesion and inter-semantic discriminability. A spatial concentration loss is designed to leverage attention maps to tighten local concentration of semantically relevant regions, while a visual contrastive loss is tailored to improve the discriminability of visual semantic features. Finally, the optimized region features are fed into a sparse region-word alignment module for the fine-grained cross-modal alignment. The contributions can be summarized as follows.

- We propose a novel CoV-Align framework, which achieves a win-win scenario: precise fine-grained visual region generation improves alignment accuracy, while text-independent aggregation reduces computational complexity.
- We design a coarse visual semantic feature extractor to group patches to regions with semantic consistency. A cohesive and discriminative feature optimization is presented to enhance intra-semantic cohesion and inter-semantic discriminability of region features.
- CoV-Align achieves state-of-the-art performance on the Flickr30K and MS-COCO benchmarks. Critically, it demonstrates superior efficiency, delivering up to almost a 3–5× speedup and a significant reduction in FLOPs compared to the top counterparts.

2. Related Work

Coarse-grained Alignment Methods. Image-text alignment aims to measure the semantic similarity between visual and textual data, which is broadly divided into coarse-grained and fine-grained approaches. Coarse-grained methods align the global representations of an entire image and a complete text caption. Early works [3, 8] first extract regional features using object detectors [25] and then aggregate them into a single global vector. More recent meth-

ods, notably CLIP [24], employ an end-to-end dual-encoder architecture to directly learn global embeddings. These approaches simplify the pipeline and have shown strong generalization. However, these global methods struggle to capture the fine-grained correspondences between specific image regions and textual phrases, limiting overall performance in complex scenes.

Fine-grained Alignment Methods. Fine-grained methods are developed to address this limitation by focusing on local-level alignment between image regions and words. Initial approaches like SCAN [16] rely on the object detectors like Faster-RCNN [25] for region proposals, but are constrained by the detector’s fixed vocabulary and high computational cost. With the great success of Vision Transformers, detector-free models, e.g., FILIP [30], compute the similarity across all patch-token pairs for alignment, but are sensitive to noisy or redundant patches. To refine this, subsequent works such as LAPS [10] and SPARC [1] introduce cross-modal attention, using textual context to guide the aggregation of visual features. While these approaches show promising results, text-guided aggregation significantly increases computational complexity and inference latency, making it challenging for large-scale retrieval. In view of this, this paper focuses on extracting high-quality, semantically concentrated region features without textual guidance, achieving both high accuracy and efficiency.

3. Methodology

As mentioned in Sec. 1, the proposed CoV-Align is able to generate fine-grained visual semantic tokens independently of text, improving both region-text matching accuracy and efficiency. The overview of CoV-Align is shown in Figure 2. Given an image-text pair, patch and text features are extracted via the visual and the text encoder. The coarse-grained visual semantic feature extractor then identifies salient regions, which are refined by the cohesive and discriminative feature optimization. Finally, a sparse alignment module computes region-word similarity.

3.1. Coarse Visual Semantic Feature Extraction

We adopt a dual-encoder architecture to extract initial features, where a Vision Transformer (ViT) [7] processes an image I into patch features $F_v \in \mathbb{R}^{(N+1) \times d_v}$, and a Transformer-based text encoder [26] generates word features $F_t \in \mathbb{R}^{L \times d_t}$ from text T , where L, d_v, d_t indicate the text length, the dimensions of visual and text features, respectively. Given the initial image tokens, we design a Coarse Visual Semantic Feature Extractor (CVSFE) to generate coarse visual region representations using deformable attention and consistent assign attention.

We employ Deformable Attention (DAT) [34] to attend to a sparse set of semantically relevant spatial locations for

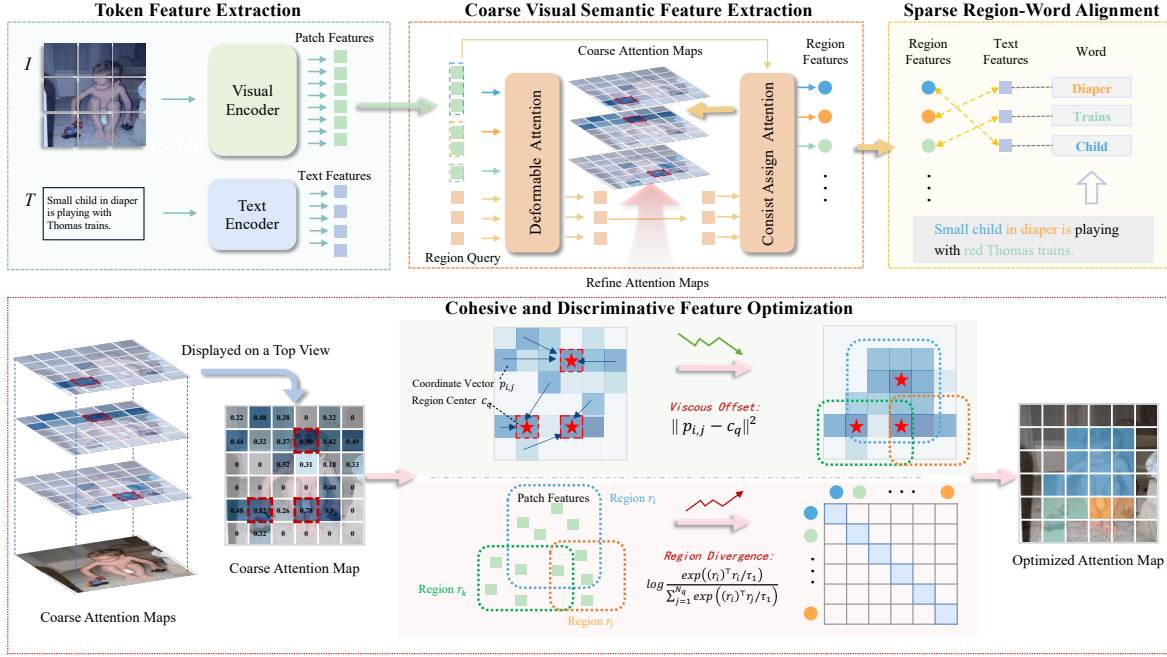


Figure 2. Overview of CoV-Align. Given an image-text pair (I, T) , initial visual and textual features are extracted via a dual-encoder architecture. Then, the CVSFE utilizes Deformable Attention and Consistent Assign Attention to generate initial regions without text guidance. The region representations are further improved by the CDFO, ensuring their intra-semantic cohesion and inter-semantic discriminability (only 3 attention maps referring 3 regions are displayed for simplicity). Finally, the Sparse Region-Word Alignment module computes the alignment scores between these refined region features and word features.

each region query. In addition to the image tokens, learnable queries participate in the deformable attention computation. This process can be formulated as:

$$z_q^{l+1} = \text{MSDeformAttn}(z_q^l, p_q, F_v) = \sum_{m=1}^M W_m^l \left(\sum_{k=1}^K A_{mk}^l \cdot F_v(p_q + \Delta p_{mk}^l) \right), \quad (1)$$

where $z_q \in \mathbb{R}^{N_q \times d}$ denotes the learnable region queries and $p_q \in \mathbb{R}^2$ represents the reference point. K is the number of sampling offsets per reference point, l is the index of the attention layer, and m is the index of the attention head, ranging from 1 to M . A_{mk}^l and Δp_{mk}^l represent the attention weight and the sampling offset for the k -th sampling point in the l -th layer and the m -th attention head. This allows each region query to focus on a sparse set of semantically relevant spatial locations, effectively aggregating semantically visual features while maintaining high computational efficiency.

After the deformable attention, we exploit consistent assign attention to further enforce semantic consistency between a certain region and global patches. To achieve this, a shared projection matrix W is used to constrain the attention and enforce consistency across attention distribu-

tions for generating more semantically coherent coarse-grained features, thereby mitigating the propensity to establish spurious correlations between semantically disparate regions (e.g., irrelevant background or contextual features) [19, 24, 27]. The effectiveness of this shared projection matrix design is demonstrated in Appendix B. The attention scores are computed as Eq. (2).

$$\text{Attn} = \text{Softmax}\left(\frac{z_q W W^\top F_v^\top}{\tau}\right), \quad (2)$$

where τ is a temperature hyperparameter. However, the global attention introduces noise from irrelevant patches that degrades semantic discriminability and consistency of aggregated region features. We apply a masking strategy that prunes visual patches with similarity scores below a predefined threshold σ during feature aggregation, as in Eq. (3) to (4).

$$\bar{\text{Attn}}_{qk} = \begin{cases} \hat{\text{Attn}}_{qk}, & \hat{\text{Attn}}_{qk} \geq \sigma, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

$$\hat{r}_q = \bar{\text{Attn}}_q V_{F_v}^\top, \quad (4)$$

where $\hat{\text{Attn}}_{qk} = \frac{\text{Attn}_{qk} - \min_p \text{Attn}_{qp}}{\max_i \text{Attn}_{qp} - \min_p \text{Attn}_{qp}}$ normalizes the attention scores to $[0, 1]$ for effective threshold comparison.

$(Attn_{qk})_{1 \leq q \leq N_q, 1 \leq k \leq N}$ represents the similarity between the q -th region query and the k -th visual patch, and $\hat{r}_q$ denotes the masked attention-weighted region feature. $\bar{Attn}_q$ is the coarse attention map in terms of the q -th region query, and the coverage area for each region is determined by the threshold σ in Eq. (3). In our experiments, totally 8 maps are determined, each with its own region center.

3.2. Cohesive and Discriminative Feature Optimization

After coarse visual semantic feature extraction, the attention distributions corresponding to region queries often disperse to distant patches, causing the spatial drift issue. Moreover, the partitioned regions tend to exhibit spatial overlaps in the projection view, resulting in ambiguous region representations. To address these limitations, we present a Cohesive and Discriminative Feature Optimization (CDFO) with two complementary constraints: a Spatial Concentration Loss to enforce spatial concentration within each region feature, and a Visual Contrastive Loss to promote discriminability across features of different regions. These constraints collaboratively optimize the models, enabling region representations to achieve both spatial concentration and semantic discriminability.

Spatial Concentration Loss Constraint. In vision tasks, the high semantic correlation among neighboring pixels constitutes a fundamental prior for spatial continuity. In view of this, we design a spatial concentration loss to strengthen the local pattern with definite semantics. Specifically, it explicitly constrains the attention distribution of each region feature to concentrate on visual patches proximal to its computed region center $\mathbf{c}_q$, where q denotes the region index, as formulated in Eq. (5) to (7).

$$a_q = \frac{\bar{Attn}_q}{\sum_{k=1}^N \bar{Attn}_{qk}}, \quad (5)$$

$$\mathbf{c}_q = \sum_{i=1}^H \sum_{j=1}^W (a_q)_{i,j} \mathbf{p}_{i,j}, \quad (6)$$

$$\mathcal{L}_{\text{conc}} = \sum_{i=1}^H \sum_{j=1}^W (a_q)_{i,j} \|\mathbf{p}_{i,j} - \mathbf{c}_q\|^2. \quad (7)$$

Here, N is the number of image patches after threshold masking, and $(a_q)_{i,j}$, $\mathbf{p}_{i,j}$ are the normalized attention weight and 2D coordinate vector at position (i, j) , respectively. By penalizing attention weights deviating from the computed region center, we effectively guide the attention $\bar{Attn}_q$ to form a spatially compact distribution.

Visual Contrastive Loss Constraint. Recall that aggregated regions are inevitably overlapped, leading to ambiguous region representations. In fact, minimizing information redundancy across the extracted region features is

essential for obtaining informative and discriminative representations. To achieve this, we introduce the visual contrastive loss, incorporated with the Spatial Concentration Loss, to optimize the framework collaboratively, ensuring intra-semantic cohesion and inter-semantic discriminability of region representations. Specifically, we transform regions features into the text embedding space through a linear transformation, denoted as r_i . For a certain region feature r_i , we treat the feature itself as the positive sample, while regarding all other region features $r_j (j \neq i)$ within the same image as negative samples. The Visual Contrastive Loss primarily tries to push dissimilar ones to be far apart, encouraging each region feature to be maximally distinctive from other region features within the image, as illustrated in Eq. (8).

$$L_{v2v} = -\frac{1}{N_q} \sum_{i=1}^{N_q} \log \frac{\exp((r_i)^\top r_i / \tau_1)}{\sum_{j=1}^{N_q} \exp((r_i)^\top r_j / \tau_1)}, \quad (8)$$

where N_q is the number of regions and $\tau_1 \in \mathbb{R}$ is a scaling temperature parameter.

3.3. Sparse Region-Word Alignment

It is noteworthy that the procedure of aggregating patches into regions is performed independently of text guidance. At the stage of region-word alignment, the word token features F_t derived from the text encoder are participated in the similarity computation with the region feature. The region-word similarity matrix $\mathbf{s} \in \mathbb{R}^{N_q \times L}$ is achieved by Eq. (9).

$$\mathbf{s}_{ij} = \frac{r_i^T (F_t)_j}{\|r_i\| \|(F_t)_j\|}. \quad (9)$$

To achieve optimal association, we use a max-sum pooling strategy [23]. For each visual region (or textual token), we only remain its maximum similarity score across all text words (or regions).

The scores are then averaged to calculate the accumulated image-text association factor $S(I, T)$.

$$S(I, T) = \frac{1}{N_q} \sum_{i=1}^{N_q} \max_j (\mathbf{s})_{ij} + \frac{1}{L} \sum_{j=1}^L \max_i (\mathbf{s})_{ij}. \quad (10)$$

Similarly, we can obtain $S(T, I)$ using a corresponding formula. Following [15, 24], we use contrastive loss functions, as presented in Eq. (11), (12), to maximize the similarity between embeddings of positive pairs and minimizing it for negative pairs.

$$L_{v2t} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(S(I_i, T_i))}{\sum_{j=1}^N \exp(S(I_i, T_j))}, \quad (11)$$

$$L_{t2v} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(S(T_i, I_i))}{\sum_{j=1}^N \exp(S(T_i, I_j))}, \quad (12)$$

Table 1. Comparison of image-text retrieval performance on Flickr30K and MS-COCO datasets. Results are reported for 1K and 5K test splits, with R@1, R@5, R@10 recall rates and the sum (RSUM) for image-to-text and text-to-image retrieval tasks.

Method	Flickr30K 1K							MS-COCO 1K							MS-COCO 5K						
	Image-to-Text			Text-to-Image			RSUM	Image-to-Text			Text-to-Image			RSUM	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
<i>Faster R-CNN + BERT_{base}</i>																					
TGDT [21]	61.3	86.0	91.4	76.8	93.2	96.4	505.1	65.4	91.8	96.5	78.5	96.4	98.9	527.5	43.3	73.5	83.3	57.5	84.8	91.6	434.0
CHAN [23]	80.6	96.1	97.8	63.9	87.5	92.6	518.5	81.4	96.9	98.9	66.5	92.1	96.7	532.6	59.8	87.2	93.3	44.9	74.5	84.2	443.9
HREM [9]	83.3	96.0	98.1	63.5	87.1	92.4	520.4	81.1	96.6	98.9	66.1	91.6	96.5	530.7	62.3	87.6	93.4	43.9	73.6	83.3	444.1
NUIF [33]	85.6	97.2	98.6	69.8	90.4	94.4	535.9	84.7	97.5	99.1	70.6	93.1	97.2	542.3	67.8	89.8	94.8	49.9	77.9	86.7	466.9
<i>ViT_{base}-224 + BERT_{base}, 14×14 patches</i>																					
VSE++ [8]	71.8	92.8	96.5	59.4	84.7	90.9	496.1	75.0	94.6	98.0	62.7	89.4	94.9	514.6	52.4	80.3	88.8	40.6	70.4	81.1	413.4
SCAN [16]	69.5	90.9	95.6	56.4	83.1	90.0	485.6	76.0	95.4	98.1	64.5	90.8	95.8	520.6	53.9	81.8	90.0	42.9	72.3	82.5	423.5
SGR [6]	69.7	90.8	95.2	59.1	84.1	89.9	488.7	77.2	95.0	98.0	65.1	90.7	95.8	521.8	54.9	82.8	90.5	42.8	72.2	82.5	425.8
CHAN [23]	69.2	91.8	95.0	58.4	84.9	90.6	489.9	77.1	95.1	98.1	65.0	91.0	96.0	522.2	56.3	83.2	90.1	43.0	72.6	82.8	428.0
LAPS [10]	74.0	93.4	97.4	62.5	87.3	92.7	507.3	78.7	95.5	98.3	66.2	91.3	96.2	526.3	57.5	84.0	90.8	44.5	74.0	83.6	434.4
AVSE [22]	76.0	94.6	97.5	62.7	88.4	93.1	512.3	79.8	95.6	98.3	67.0	91.5	96.3	528.5	58.8	84.3	91.0	45.1	74.3	83.9	437.4
CoV-Align	78.5	94.7	96.9	63.3	88.9	93.7	516.0	79.9	96.0	98.6	66.5	91.7	96.6	529.3	60.6	84.7	91.6	44.9	74.5	83.9	440.1
<i>ViT_{base}-384 + BERT_{base}, 24×24 patches</i>																					
VSE++ [8]	77.1	95.7	97.5	65.8	90.2	94.3	520.5	77.0	95.7	98.4	64.6	91.1	96.2	523.0	54.9	82.8	90.4	42.4	72.4	82.8	425.8
SCAN [16]	75.4	94.4	96.9	63.6	88.6	93.5	512.5	76.1	95.5	98.5	65.1	91.6	96.3	523.1	53.3	81.8	90.0	42.6	72.6	82.9	423.1
SGR [6]	76.9	94.9	98.1	64.2	88.4	93.3	515.8	75.8	95.7	98.6	65.6	92.0	96.5	524.2	53.3	81.0	89.6	42.9	73.1	83.7	423.6
CHAN [23]	75.4	94.5	97.6	63.2	88.6	93.1	512.4	78.1	95.8	98.6	66.1	92.1	96.6	527.3	55.6	83.8	91.2	43.4	73.6	83.5	431.1
LAPS [10]	79.0	96.0	98.1	67.3	90.5	94.5	525.4	78.6	96.3	98.9	68.0	92.4	96.8	531.0	57.4	84.9	92.5	46.4	75.8	85.2	442.2
AVSE [22]	80.3	96.4	98.7	67.9	91.2	94.7	529.2	81.1	97.1	99.0	68.3	92.7	97.0	535.2	61.2	86.8	93.2	46.2	75.9	85.0	448.3
CoV-Align	85.0	97.1	98.7	69.7	92.0	95.7	538.2	81.0	97.0	99.0	69.1	92.9	97.3	536.9	63.4	86.9	93.5	47.3	76.6	85.6	453.4
<i>Swin_{base}-224 + BERT_{base}, 7×7 patches</i>																					
VSE++ [8]	82.5	96.5	98.9	70.0	91.4	95.1	534.4	83.3	97.5	99.3	71.0	93.0	96.7	540.9	64.0	88.2	94.2	49.9	78.0	86.6	460.9
SCAN [16]	79.0	95.9	98.2	67.7	90.6	94.9	526.3	80.9	97.0	99.1	69.7	93.1	97.1	536.9	60.7	86.6	93.2	48.1	77.1	86.1	451.8
SGR [6]	80.4	97.0	98.7	66.9	90.2	94.5	527.6	81.2	97.1	99.1	69.9	93.2	97.2	537.7	61.0	86.7	93.2	48.6	77.2	86.3	453.1
CHAN [23]	81.4	97.0	98.6	68.5	90.6	94.5	530.6	81.6	97.2	99.3	70.6	93.7	97.6	539.8	64.1	87.9	93.5	49.1	77.3	86.1	458.0
LAPS [10]	82.4	97.4	99.5	70.0	91.7	95.4	536.3	84.0	97.6	99.3	72.1	93.7	97.3	544.1	64.5	89.2	94.4	51.6	78.9	87.2	465.8
AVSE [22]	83.9	97.4	99.4	70.0	92.4	95.6	538.7	84.9	98.0	99.3	72.1	94.0	97.4	545.7	66.2	89.8	94.7	51.7	79.2	87.3	468.9
CoV-Align	84.8	97.7	99.3	71.4	92.4	95.8	541.3	85.2	97.7	99.2	71.8	93.5	97.4	544.9	67.4	90.1	94.9	51.3	79.2	87.1	470.1
<i>Swin_{base}-384 + BERT_{base}, 12×12 patches</i>																					
VSE++ [8]	83.3	97.5	99.2	71.1	93.2	96.2	540.6	82.9	97.7	99.4	71.3	93.5	97.3	542.1	63.0	88.5	94.3	50.1	78.9	87.4	462.2
SCAN [16]	81.9	96.9	98.9	70.0	92.7	95.8	536.1	81.6	96.8	99.1	69.1	92.7	96.7	536.1	61.1	87.3	93.3	47.8	76.9	85.9	452.4
SGR [6]	80.7	96.8	99.0	69.9	91.7	95.3	533.4	81.9	96.7	99.1	69.3	92.8	96.7	536.6	62.8	87.0	92.9	48.1	77.0	86.0	453.8
CHAN [23]	81.2	96.7	98.8	70.3	92.2	95.9	535.0	83.1	97.3	99.2	70.4	93.1	97.1	540.2	63.4	88.4	94.1	49.2	77.9	86.6	459.5
LAPS [10]	85.1	97.7	99.2	74.0	93.0	96.3	545.3	84.1	97.4	99.2	72.1	93.9	97.4	544.1	67.1	88.6	94.3	53.0	79.5	87.6	470.1
AVSE [22]	87.1	98.3	99.2	73.6	93.5	96.5	548.2	85.1	98.2	99.5	71.6	94.0	97.5	545.9	68.6	90.2	95.6	52.2	79.6	87.8	474.0
CoV-Align	90.2	98.7	99.5	76.7	94.6	97.4	557.0	86.7	98.0	99.5	74.0	94.4	97.8	550.5	70.7	91.5	96.0	54.1	80.9	88.7	481.9

where N is the number of matched image-text pairs. τ_2 denotes a trainable temperature hyperparameter.

The overall training objective is a weighted sum of the global contrastive loss, the spatial concentration loss and the visual contrastive loss, as follows.

$$L_{\text{total}} = \frac{\lambda_g}{2} (L_{t2v} + L_{v2t}) + \lambda_v L_{v2v} + \lambda_c L_{\text{conc}}, \quad (13)$$

where λ_g , λ_v and λ_c are trade-off parameters.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct comprehensive evaluations on the Flickr30K [31] and MS-COCO [4] datasets, where each image comes

with five captions. For Flickr30K, we follow the split from [2, 16], utilizing 29,000 images for training, 1,014 for validation, and 1,000 for testing. Regarding the MS-COCO dataset, we adopt the split from [8] with 113,287 training images, 5,000 validation images, and 5,000 test images. In accordance with [10], MS-COCO results are reported as the average performance over a 5-fold 1K test set and on the complete 5K test set. The evaluation metrics includes Recall@K (R@K, K=1,5,10), the percentage of ground truths in the top-K retrieved results, and the sum of all six recall scores for bidirectional retrieval.

Table 2. Comparison of image-text retrieval performance for Vision-Language Pre-training (VLP) models and CLIP-based variants. Results are evaluated on Flickr30K (1K test set) and MSCOCO (5K test set) datasets with different visual encoder configurations. * represents the results obtained by fine-tuning on the corresponding dataset under the same settings.

Model	Flickr30K (1K Test Set)							MSCOCO (5K Test Set)						
	Image-to-Text			Text-to-Image			RSUM	Image-to-Text			Text-to-Image			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
<i>Pre-Trained Models</i>														
UNITER [5]	87.3	98.0	99.2	75.6	94.1	96.8	551.0	67.5	88.6	93.8	52.9	79.9	88.0	470.7
VILT [14]	83.5	96.7	98.6	64.4	88.7	93.8	525.7	61.5	86.3	92.7	42.7	72.9	83.1	439.2
CLIP [24]	88.0	98.7	99.4	68.7	90.6	95.2	540.6	58.4	81.5	88.1	37.8	62.4	72.2	400.4
FG-CLIP [29]	82.5	-	-	76.4	-	-	-	64.1	-	-	45.4	-	-	-
FineCLIP [13]	88.0	96.4	98.6	67.9	89.1	94.1	534.1	54.5	78.6	85.8	40.2	66.5	76.1	401.7
SOHO [12]	86.5	98.1	99.3	72.5	92.7	96.1	545.2	66.4	88.2	93.8	50.6	78.0	86.7	463.7
ALBEF [17]	95.9	99.8	100.0	85.6	97.5	98.9	577.7	77.6	94.3	97.2	60.7	84.3	90.5	504.6
BLIP [18]	96.6	99.8	100.0	87.2	97.5	98.8	579.9	80.6	95.2	97.6	63.1	85.3	91.1	512.9
<i>CLIP-ViT-Base-224 + CLIP-BERT-Base, 14×14 patches</i>														
CLIP _{ViT-B/16} [24]	88.4	98.7	99.5	76.1	94.6	97.2	554.5	65.2	87.3	93.8	51.2	77.9	86.4	465.3
LG-MGC [32]	92.6	99.5	99.7	78.9	95.5	98.2	564.5	67.6	88.5	93.8	51.2	77.9	86.4	465.3
CoV-Align	93.4	99.4	99.7	79.8	95.5	97.8	565.6	70.6	90.2	94.7	53.7	80.0	87.8	477.0
<i>CLIP-ViT-Large-224 + CLIP-BERT-Large, 16×16 patches</i>														
CLIP _{ViT-L/14} [24]	93.1	99.7	99.8	78.9	95.5	98.2	565.2	68.1	88.6	94.0	51.1	77.8	86.3	466.0
AGREE [28]	92.1	98.7	99.2	82.8	95.9	97.8	566.5	-	-	-	-	-	-	-
LG-MGC [32]	92.6	99.5	99.7	78.9	95.5	98.2	564.5	67.6	88.5	93.8	51.2	77.9	86.4	465.3
CoV-Align	94.7	99.7	99.9	83.8	96.7	98.5	573.3	73.7	92.4	96.4	57.6	82.4	89.5	491.9

4.2. Implementation Details

We employ ViT-B/16, Swin-B/32, and CLIP-ViT-B/L as visual backbones, paired with BERT or CLIP text encoders. All features are subsequently projected to a 512-dimensional shared space. The framework is trained for 30 epochs with the Adam optimizer, using a learning rate of $1e-4$, a weight decay of $1e-4$, and a cosine annealing schedule. The hyperparameters λ_g , λ_v , λ_c , σ and N_q are assigned with 1, 1, 0.5, 0.7 and 8, respectively. The loss coefficients (λ_g , λ_v , λ_c) are determined through extensive ablation studies detailed in Appendix A.

4.3. Comparison with State-of-the-art Methods

We conduct comprehensive comparisons between our method and two categories of methods on the Flickr30K and MS-COCO datasets, including detector-based approaches, e.g., TGD [21], CHAN [23], HREM [9], and NUIF [33], and Transformer-based methods, e.g., VSE++ [8], SCAN [16], SGR [6], CHAN [23], LAPS [10], LG-MGC [32]. The evaluation results for all baseline models are directly sourced from their original publications to maintain result authenticity. Notably, for the Transformer-based models, we adopt the evaluation results reported in [10] for fair comparison under the same evaluation protocols. Table 1 reports the quantitative results on the Flickr30K and MS-COCO datasets. Across various visual backbones, CoV-Align con-

Table 3. Ablation studies for our CoV-Align on Flickr30K dataset. TR denotes Text-to-Image Retrieval, and IR denotes Image-to-Text Retrieval.

Method	TR		IR	
	R@1	R@10	R@1	R@10
Baseline	70.4	96.5	59.4	92.0
+ Coarse Visual Semantic Feature	73.9	96.9	62.1	92.2
+ Visual Contrastive Loss	75.4	97.0	62.6	92.7
+ Spatial Concentration Loss	78.5	96.9	63.3	93.7
Different Threshold Settings of σ				
$\sigma = 0.0$	74.5	96.7	60.1	92.5
$\sigma = 0.3$	73.1	96.6	61.4	92.8
$\sigma = 0.5$	74.3	97.7	62.8	93.9
$\sigma = 0.7$	78.5	96.9	63.3	93.7
$\sigma = 0.9$	76.2	98.0	62.7	93.3
Different Region Settings of N_q				
$N_q = 4$	72.7	96.8	61.5	92.8
$N_q = 8$	78.5	96.9	63.3	93.7
$N_q = 12$	75.7	97.7	62.5	93.6
$N_q = 16$	75.2	97.4	62.0	93.6

sistently outperforms state-of-the-art methods, notably surpassing one of the representative methods, i.e., LAPS[10]. Specifically, in terms of the RSum metric on Flickr30K,

Table 4. Efficiency comparison on Flickr30K dataset. All models use ViT-Base-224 and BERT-Base encoders. FLOPs and GPU Memory are measured with batch size 128 to evaluate feasibility for large-scale retrieval tasks. GPU Memory denotes training memory consumption. * denotes the reproduced results.

Model	#FLOPs	#Params	#GPU Memory
CLIP*	2393.64G	196.66M	27G
CHAN*	2430.11G	196.72M	30G
SCAN*	2451.90G	196.72M	36G
LAPS	2577.61G	196.72M	49G
CoV-Align	2289.22G	217.19M	25G

CoV-Align achieves improvements over LAPS by 8.7 points with ViT-224 (516.0 vs. 507.3), 12.8 points with ViT-384 (538.2 vs. 525.4), 5.0 points with Swin-224 (541.3 vs. 536.3), and 11.7 points with Swin-384 (557.0 vs. 545.3). Similar substantial performance gains are observed on the MS-COCO 1K and 5K test sets, demonstrating the effectiveness of cohesive visual semantics priority for fine-grained cross-modal alignment. It is worth noting that the performance margin over LAPS widens with higher input resolution (384×384 vs. 224×224). This demonstrates the independent region aggregation is able to better utilize the attention maps from high-resolution inputs to improve alignment.

As shown in Table 2, we evaluate CoV-Align by integrating it into CLIP and comparing against both state-of-the-art VLP models, e.g., UNITER [5], ALBEF [17], BLIP [18], and other CLIP fine-tuning methods, e.g., direct fine-tuning, LG-MGC [32], AGREE [28]. Experimental results show that CoV-Align achieves performance improvements over the original CLIP model and other fine-tuned variants in image-text retrieval tasks, as an RSUM score of 491.9 vs. LG-MGC’s 465.3. Moreover, it remains competitive with state-of-the-art vision-language pretraining (VLP) models under the same setting.

Model Efficiency Analysis. Figure 3 illustrates the comparison of evaluation time (in seconds) for different image-text retrieval models on the Flickr30K and MS-COCO datasets, where all methods are evaluated with a consistent ViT-B/224 and BERT-Base architecture for a fair comparison. It can be observed that CoV-Align demonstrates superior computational efficiency without compromising its accuracy superiority. Notably, it maintains comparable efficiency to coarse-grained models like CLIP while delivering significantly accelerated processing over existing fine-grained approaches, including SCAN, CHAN, and LAPS. Specifically, we achieve approximately a $3\times$ speedup (114.0s vs. 442.2s on the COCO dataset) over LAPS, one of the most advanced fine-grained models. Ta-

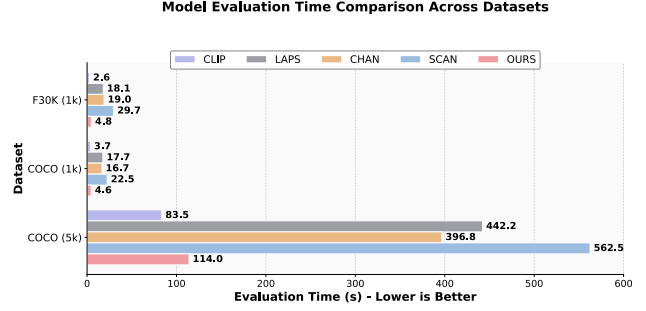


Figure 3. Comparison of evaluation time (in seconds) for different image-text retrieval models on the Flickr30K and MS-COCO datasets.

ble 4 presents detailed computational metrics for large-scale retrieval feasibility. CoV-Align achieves the lowest computational complexity of 2289.22 GFLOPs, 7-15% reduction compared to others. A minimum training memory of 25G is occupied, 49% less than LAPS. The reduced FLOPs and memory consumption validate the scalability of our approach for large-scale retrieval applications, while Cov-Align maintains competitive parameter efficiency.

4.4. Ablation Studies

To evaluate the contribution of each component in our proposed CoV-Align framework, we conducted comprehensive ablation studies on the Flickr30K dataset. In all experiments, we use the same ViT-Base-224 and BERT backbones. The results are summarized in Table 3.

Impact on different semantic feature. Following [18], we create a baseline that integrates a learnable object query-based cross-attention mechanism for text-word alignment. These queries interact with visual patches through stacked attention layers to extract semantic features. As indicated in Table 3, the baseline yields suboptimal performance in cross-modal retrieval tasks. By integrating Coarse Visual Semantic Feature Extraction, the performance is significantly enhanced, with R@1 increasing from 70.4/59.4 to 73.9/62.1. It stems from the module’s ability to generate more semantically coherent region-level features compared to the baseline’s conventional coarse-grained processing approach.

Impact on different optimization methods. Based on the coarse visual semantic feature, we successively introduce two more losses to evaluate the effectiveness of the cohesive and discriminative feature optimization. The experiment results are shown in Table 3. After incorporating the visual contrastive loss, we observe performance gains of R@1 by 1.5 and 0.5 on Text-to-Image and Image-to-Text retrieval tasks, which verifies its ability in improving the inter-semantic discriminability. Building upon this, we further introduce spatial concentration loss, and achieve the high-



Figure 4. Visualizations of our alignment between textual phrases and image patches.

est performance with R@1 scores further increasing by 3.1 and 0.7. The results feature the significance of modeling the intra-semantic cohesion and inter-semantic discriminability of visual region features in fine-grained cross-modal alignment.

Impact on different hyper-parameters. To validate the efficacy of the thresholding design in Consistent Assignment Attention Strategy, we conduct ablation studies on the Flickr30K dataset with different threshold values ($\sigma = 0.0, 0.3, 0.5, 0.7, 0.9$). As shown in Table 3, when $\sigma = 0.0$, the model attends to all visual patches, introducing substantial noise and resulting in poor performance. As σ increases, the model’s ability to filter irrelevant patches improves, with optimal performance achieved at $\sigma = 0.7$. Further increasing to $\sigma = 0.9$ causes slight performance degradation due to over-filtering of semantically meaningful information. These results reveal that appropriate thresholding effectively removes visual noise while preserving relevant features for fine-grained alignment. Therefore, we adopt $\sigma = 0.7$ as the optimal threshold parameter.

To investigate the optimal number of semantic regions, we conduct ablation studies on the Flickr30K dataset with different N_q values. As shown in Table 3, the optimal performance is achieved at $N_q = 8$, which strikes an effective balance between semantic diversity and representation coherence.

4.5. Visualization

In Figure 4, we visualize CoV-Align’s capability of capturing fine-grained cross-modal alignment by highlighting the attended contents. CoV-Align guides each region feature to aggregate spatially concentrated and semantically consistent visual patches, with the correct association between attention maps and text labels demonstrating effective alignment between regional features and relevant words. When

the textual description involves nouns such as “a stop sign” or “a black sweatshirt”, the model accurately focuses on the corresponding object regions, as illustrated in Figure 4 (a) and (b). Conversely, when the description includes verbs like “is walking” or “is pushing”, the model localizes attention to both the action agent and pertinent regions. For instance, the attention for “is walking” in Figure 4 (b) encompasses the individual’s legs and the accompanying dog, while “is pushing” in Figure 4 (d) highlights the girl’s arms and the stroller’s handlebar. This demonstrates that our approach effectively aligns textual phrases with fine-grained visual content, achieving precise visual-semantic association.

5. Conclusions

In this paper, we introduce an effective and efficient fine-grained cross-modal alignment framework with cohesive visual semantics priority, termed CoV-Align, to overcome two major limitations in cross-modal alignment, i.e., redundant patch-word correlations and high computational costs. Specifically, CoV-Align achieve efficient fine-grained alignment by progressively aggregating visual regions without text guidance. Through a Coarse Visual Semantic Feature Extractor as well as a Cohesive and Discriminative Feature Optimization, it enhances intra-semantic cohesion and inter-semantic discriminability of visual region features, contributing to explicit improvements in cross-modal alignment. Extensive experiments on the Flickr30K and MSCOCO datasets demonstrate the superiority of our framework. It has strong potential for downstream tasks in multimodal large models owing to its dual advantages in both accuracy and computational efficiency.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.62376037, 62306042), and Beijing Natural Science Foundation (L247010).

References

- [1] Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, et al. Improving fine-grained understanding in image-text pre-training. *arXiv preprint arXiv:2401.09865*, 2024. 1, 2
- [2] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 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*, pages 12655–12663, 2020. 5
- [3] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798, 2021. 2
- [4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5
- [5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 6, 7
- [6] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1218–1226, 2021. 5, 6
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2
- [8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 1, 2, 5, 6
- [9] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2023. 5, 6
- [10] Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26307–26316, 2024. 1, 2, 5, 6
- [11] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022. 1
- [12] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 6
- [13] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Wei Wei, Huiwen Zhao, Zhiwu Lu, et al. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024. 6
- [14] Wonjae Kim, Bokyoung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning*, pages 5583–5594, 2021. 6
- [15] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*, 2024. 4
- [16] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision*, pages 201–216, 2018. 1, 2, 5, 6
- [17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021. 6, 7
- [18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900, 2022. 6, 7
- [19] Yi Li, Hualiang Wang, Yiqun Duan, Hang Xu, and Xiaomeng Li. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*, 2022. 3
- [20] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, page 111409, 2025. 1
- [21] Chong Liu, Yuqi Zhang, Hongsong Wang, Weihua Chen, Fan Wang, Yan Huang, Yi-Dong Shen, and Liang Wang. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. *IEEE Transactions on Image Processing*, 32:3622–3633, 2023. 5, 6
- [22] Yang Liu, Mengyuan Liu, Shudong Huang, and Jiancheng Lv. Asymmetric visual semantic embedding framework for efficient vision-language alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5676–5684, 2025. 5
- [23] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19275–19284, 2023. [1](#), [4](#), [5](#), [6](#)
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. [1](#), [2](#), [3](#), [4](#), [6](#)
 - [25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. [2](#)
 - [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. [2](#)
 - [27] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *Proceedings of the European Conference on Computer Vision*, pages 315–332, 2024. [3](#)
 - [28] Xiaodan Wang, Lei Li, Zhixu Li, Xuwu Wang, Xiangru Zhu, Chengyu Wang, Jun Huang, and Yanghua Xiao. Agree: Aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2023. [6](#), [7](#)
 - [29] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025. [6](#)
 - [30] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. [1](#), [2](#)
 - [31] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. [5](#)
 - [32] Feifei Zhang, Sijia Qu, Fan Shi, and Changsheng Xu. Overcoming the pitfalls of vision-language model for image-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2350–2359, 2024. [6](#), [7](#)
 - [33] Huatian Zhang, Lei Zhang, Kun Zhang, and Zhendong Mao. Identification of necessary semantic undertakers in the causal view for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7105–7114, 2024. [5](#), [6](#)
 - [34] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. [2](#)