Linguistic-Aware Patch Slimming Framework for Fine-grained Cross-Modal Alignment

Zheren Fu, Lei Zhang, Hou Xia, Zhendong Mao*
University of Science and Technology of China, Hefei, China
{fzr,overwhelmed}@mail.ustc.edu.cn, {leizh23, zdmao}@ustc.edu.cn

Abstract

Cross-modal alignment aims to build a bridge connecting vision and language. It is an important multi-modal task that efficiently learns the semantic similarities between images and texts. Traditional fine-grained alignment methods heavily rely on pre-trained object detectors to extract region features for subsequent region-word alignment, thereby incurring substantial computational costs for region detection and error propagation issues for two-stage training. In this paper, we focus on the mainstream vision transformer, incorporating patch features for patch-word alignment, while addressing the resultant issue of visual patch redundancy and patch ambiguity for semantic alignment. We propose a novel Linguistic-Aware Patch Slimming (LAPS) framework for fine-grained alignment, which explicitly identifies redundant visual patches with language supervision and rectifies their semantic and spatial information to facilitate more effective and consistent patch-word alignment. Extensive experiments on various evaluation benchmarks and model backbones show LAPS outperforms the state-of-the-art fine-grained alignment methods by 5%-15% rSum. Our code is available at https://github.com/CrossmodalGroup/LAPS.

1. Introduction

Cross-modal alignment aims to bridge the semantic gap between different modalities, such as visual and linguistic ones. It is a fundamental technology for many multi-modal tasks, including image-text retrieval [14], visual question answering [15], image captioning [27]. The critical challenge of cross-modal alignment lies in efficiently measuring the semantic similarities between images and texts to achieve a high-quality alignment.

In general, existing cross-modal alignments can be classified into two paradigms. The first coarse-grained alignment [9, 14, 26] separately encodes the whole images and texts into a unified embedding space, then directly computes the similarity of global embeddings. The second fine-grained alignment [7, 21, 35] applies cross-modal interaction between visual and textual local features, then aggregates local alignments to learn a cumulative similarity. Previous fine-grained methods adhered to the detector-based roadmap, which entails a two-step process: first extracting region features through a pre-trained object detector, e.g., Faster-RCNN [38], then compute region-word alignments

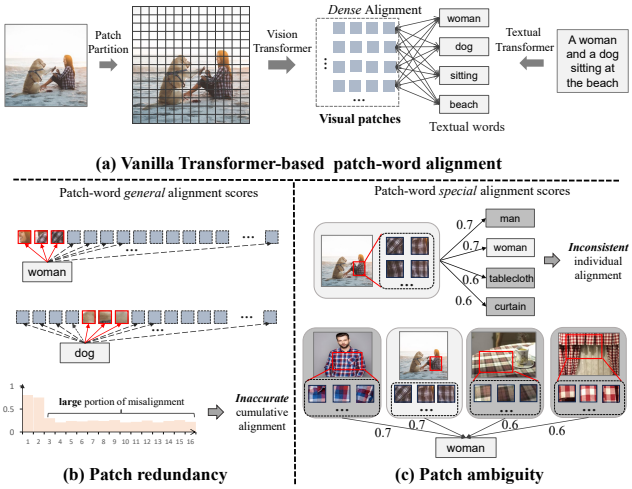

Figure 1. The motivation of our framework. (a) The vanilla transformer-based fine-grained alignment uses vision transformer [8, 31] to divide images into tiny patches and extract patch features, then bridge an interaction between visual patches and textual words to learn a cumulative alignment score. (b) A large part of visual patches are redundant to textual content. These redundant patches will overshadow crucial patches and accumulate inaccurate alignments. (c) Visual patches are disjoint fragments of an image and associated with many patches during training. These patches will obtain the average semantics of various objects and lack structure information, causing patches to have ambiguous semantics and always get moderate alignment scores with distinct texts, hence visual patches easily get the mismatching in local regions and inconsistent alignment with negative textual samples.

*Zhendong Mao is the corresponding author.

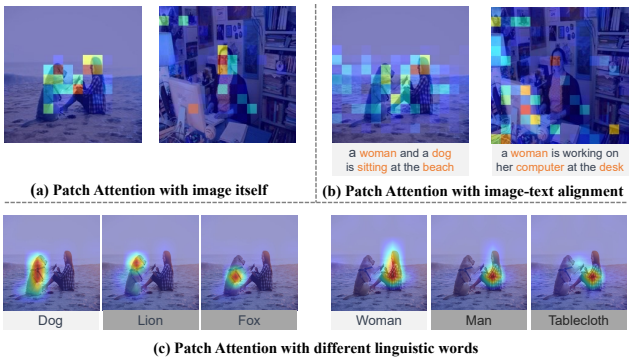

Figure 2. The visualization of patch attention for the vanilla vision transformer, where red colors represent high semantic responses and significant areas in images and blue indicates unimportant. It shows that significant areas in images are highly relevant to the (a) visual saliency (e.g., 'dog' or 'woman' areas are highlighted), or (b) textual contents with image-text alignment (e.g., 'beach' or 'computer' areas are highlighted), while these individual patches lack semantic integrity and spatial structure information. (c) Therefore, the adjacent areas will get high responses with different language words (e.g., the partial visual areas of the dog are highlighted by the textual 'lion' or 'fox', and partial areas of the woman are highlighted by the textual 'man' or 'tablecloth').

between visual regions and textual words. These frameworks heavily rely on the capability of detectors and bring expensive computation [24, 25]. Recent works adopt pure transformer architectures, e.g., vision transformer [8], to divide images into non-overlapping patches and encode patch features to construct patch-word alignments [23, 25]. The transformer-based method is a flexible end-to-end training framework, efficient for feature extraction, owns scalable performances compared to the detector-based, and has become mainstream.

However, the vanilla transformer-based patch-word framework has inherent defects, *i.e.*, the patch redundancy and patch ambiguity problems for semantic alignment. The vision transformer [8] divides images into minute patches (at 224 and 284 image resolutions, it generates $14\times14=196$ and $24\times24=576$ patches) as Fig. 1, a substantial proportion of them proves to be redundant, *e.g.*, non-salient backgrounds or text-irrelevant areas as Fig. 2(a)(b). The massive redundant patches will overshadow crucial visual patches, and accumulate unbearable misalignment during the patch-word interaction, ultimately bringing inaccurate cumulative alignments. More importantly, these fragmented patches are tiny components of an image. Tiny patches lack semantic integrity compared to complete visual regions. It will lead to ambiguous semantic expressions. Visual patches always get moderate alignment scores for distinct language concepts as Fig. 2(c), which brings inconsistent patch-word alignment in local regions with negative image-text pairs.

Consequently, how to ensure the semantic integrity of visual patches to establish accurate alignment, is a core issue for transformer-based cross-modal frameworks. To address these problems, we introduce a **L**inguistic-**A**ware **P**atch **S**limming (LAPS) framework, which effectively eliminates extensive redundant patches through linguistic supervision, and calibrates the semantic and structural information for significant patches to transform an average semantic expression into an optimal semantic for a certain image. To the best of our knowledge, LAPS is the first to explicitly explore visual patch selection and patch calibration with language contexts to facilitate patch-word alignment. As illustrated in Fig. 3, we first effectively estimate the semantic significance of visual patches using the Language-Context Patch Selection (LPS) module to pick out significant patches with differentiable sampling. Next, we adaptively rectify the semantic and structural information for significant patches through the Semantic-Spatial Patch Calibration (SPC) module to obtain distinct semantic expressions. Finally, we employ the Sparse Patch-Word Alignment (SPA) module to facilitate the fine-grained interaction between visual patches and textual words. Therefore, LAPS extends the vanilla transformer-based framework to achieve more accurate and consistent patch-word alignments. The contributions of this paper are as follows:

- We propose a Linguistic-Aware Patch Slimming (LAPS) framework for cross-modal alignment. To the best of our knowledge, this is the first work explicitly studying patch-word alignment with patch selection and calibration.
- We estimate the patch significance scores to identify redundant visual patches through linguistic supervision and select significant ones using differentiable sampling.
- We rectify significant visual patches with semantic and spatial relationships to obtain semantic integrity and structural information comparable to linguistic texts.
- We evaluate LAPS with existing fine-grained methods on diverse model backbones. LAPS outperforms state-of-the-art methods for image-text retrieval on two benchmarks, Flickr30K and MS-COCO, by 5%-15% rSum.

2. Related Work

2.1. Cross-Modal Alignment

According to the implementation of cross-modal interaction, cross-modal alignment methods can be broadly categorized into two types: *coarse-grained* and *fine-grained* alignment. Coarse-grained methods encode images and sentences into a shared embedding space [11–13], and semantic similarity is computed via the cosine similarity of cross-modal embeddings. Previous methods [4, 43] typically boost extracted local features and aggregate them to learn global embeddings, *e.g*, VSE++ [9] utilizes average pooling on region features to learn the unified embeddings. Fine-grained methods explicitly perform cross-modal inter-

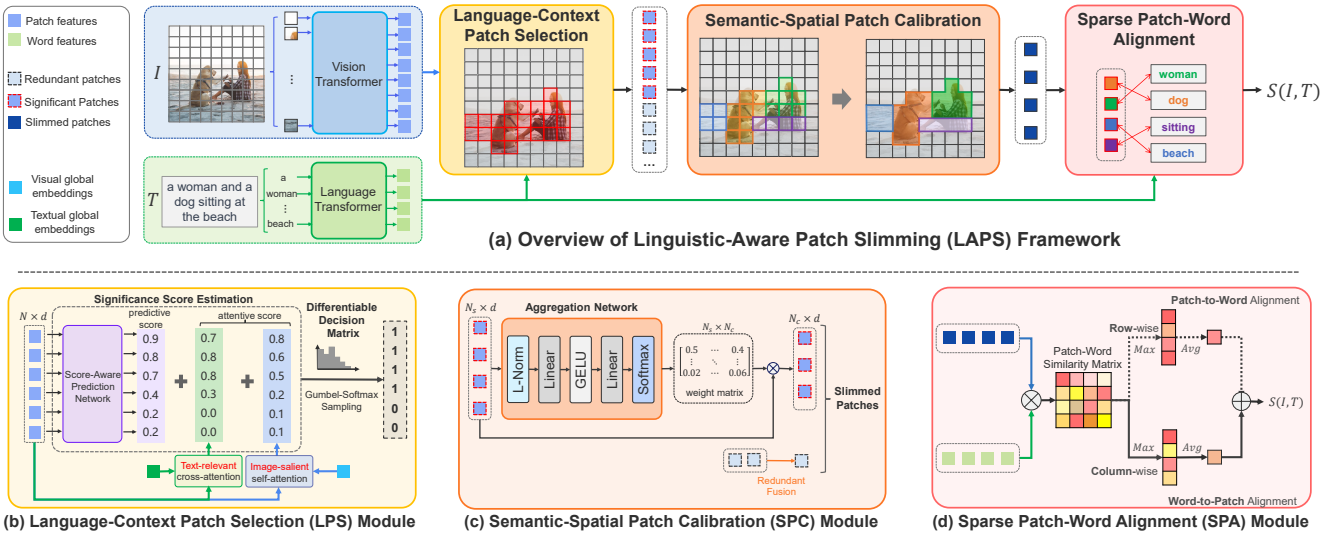

Figure 3. (a) Overview of our Linguistic-Aware Patch Slimming (LAPS) Framework for fine-grained cross-modal alignment. Given an image-text pair (I, T), we first tokenize them and then use pure Transformer encoders to extract visual patch features and textual word features. Then, we propose the Language-Context Patch Selection (LPS) module to identify text-relevant patches by language supervision. Next, we propose the Semantic-Spatial Patch Calibration (SPC) module to rectify the semantic and spatial information for significant patches, then obtain distinct semantic expressions. Finally, we propose the Sparse Patch-Word Alignment (SPA) module to compute the sparse patch-word alignment score $S(I, T)$. (b)(c)(d) The detailed architecture of proposed LPS, SPC, and SPA modules, respectively.

action between local features of two modalities and then calculate a cumulative similarity score. [3, 7, 21, 47]. Previous works typically emphasize the semantic alignment between image regions and text words. For example, SCAN [21] introduces an attention mechanism to focus on salient alignments while suppressing misalignments. SGR [7] extends the SCAN framework by a similarity reasoning network to refine region-word attention. SHAN [35] combines the hard coding assignment with SCAN to achieve efficient region-word alignments.

However, current fine-grained methods [7, 21, 35, 47] heavily rely on object detectors as visual encoders to acquire region features. The exploration of transformer-based architectures on image patches has been relatively limited. We focus on visual patches and tackle the issues of patch redundancy and ambiguity for semantic alignments.

2.2. Efficient Vision Transformer

Vision Transformer (ViT) [8] is the mainstream visual architecture, which divides the whole image into non-overlapping patches based on spatial distribution and feeds patches into a pure Transformer [41] encoder as visual token sequences. Vanilla transformers have high computational and memory costs because the self-attention has quadratic computational complexity concerning the number of tokens [31, 42]. Recently, token pruning approaches [28, 34, 37, 44] are proposed to accelerate ViTs by reducing the number of tokens at the inference stage. Some work [20, 34, 37] introduces pre-defined prediction networks to score visual tokens and drop unimportant tokens according to the scores. Others [10, 28, 44] use the attention of class tokens to evaluate the token importance and aggregate redundant tokens. However, the above methods only focus on vision tasks with the single modality, and are unsuitable for our cross-modal alignment, since they directly prune image tokens without considering the textual contexts.

2.3. Vision-Language Pre-training

The early vision language pre-training (VLP) models adhered to detector-based roadmap [5, 22], which entails a two-step process: First, it extracts visual features through a pre-trained object detector. Then, it integrates textual and visual features into a multi-modal encoder for pre-training. Although this approach has yielded strong performances across various downstream tasks, it also brings the challenges of expensive computation and unstable training on detection. Recently, ViT-based methods [16, 19, 24] employ pure transformer architectures to encode images, eliminating the requirement of object detectors and enabling end-to-end VLP framework. However, they struggle with lengthy visual token sequences and lack fine-grained cross-modal alignment information. These long visual sequences also increase computation costs and introduce noise visual information for cross-modal fusion. Some work [17, 18] proposes patch fusion approaches to learn a concise summary of visual token sequences and enhance cross-modal fusion.

3. Methodology

The overview of our LAPS is illustrated in Fig. 3. We first introduce token feature extraction in Sec. 3.1. Then, we introduce the Language-Context Patch Selection in Sec. 3.2 and Semantic-Spatial Patch Calibration in Sec. 3.3. Lastly, we describe the Sparse Patch-Word Alignment in Sec. 3.4.

3.1. Token Feature Extraction

First, we employ the pure transformer architectures [41] as the feature encoders for image and text inputs, to extract the visual and textual token sequences, respectively.

Visual Patch Tokens. For an image I, we take the vision transformers [8, 31] as the visual encoder. The image is partitioned into N non-overlapping patches based on the spatial distribution. Subsequently, we feed these patches as a visual token sequence into the vision transformer, which consists of multiple self-attention layers. We can obtain a set of visual patch features $V = \{\boldsymbol{v}_{cls}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\} \in \mathbb{R}^{(N+1)\times d}$. $\boldsymbol{v}_{cls}$ is the [CLS] token in the image, and d is the feature dimension.

Textual Word Tokens. For a sentence T, we utilize the standard sequence transformer, BERT [6] as the textual encoder. Similarly, the sentence undergoes tokenization into linguistic words and is fed to the encoder. We get a set of textual word features $T = \{\boldsymbol{t}_1, \cdots, \boldsymbol{t}_M\} \in \mathbb{R}^{M\times d}$, and M is the number of words in the sentence.

3.2. Language-Context Patch Selection

After getting visual patch features and textual word features obtained by Sec. 3.1, we would like to pick out significant visual patches for images, as shown in Fig. 3(b).

3.2.1 Significance Score Estimation

Similar to token pruning [28, 44] of ViT, we treat the patch selection as a discriminative task, that estimates a significance score for each patch and then determines the selection according to the scores. We first introduce spatial information from images into patch features, and use a Score-Aware Prediction Network to learn significant scores. The network consists of a two-layer MLP and a Sigmoid function.

$$a_i^p = \text{Sigmoid}\left(\textbf{MLP}\left(\boldsymbol{v}_i\right)\right),\ i \in \{1, \ldots, N\} \quad (1)$$

where $a_i^p \in [0, 1]$ is the significance score for the i-th patch. A higher value of a_i^p indicates a more significant patch $\boldsymbol{v}_i$. However, relying solely on a score network to predict significant patches without textual supervision is insufficient [34, 37] for cross-modal alignment. Therefore, we calculate attentive scores between visual patches and textual words to introduce linguistic contexts.

We propose two distinct attention scores: First, we compute the *cross-attention* between visual patches and textual representations, resulting in text-relevant attentive scores a^r. Secondly, we compute the *self-attention* within visual patches, yielding image-salient attentive scores a^s.

$$a_i^r = \text{Norm}\left(\boldsymbol{v}_i^T \cdot \boldsymbol{t}_{glo}/d\right), a_i^s = \text{Norm}\left(\boldsymbol{v}_i^T \cdot \boldsymbol{v}_{glo}/d\right) \quad (2)$$

where Norm represents the normalization of attentive scores into a 0-1 range, ensuring consistency with predictive scores a_i^p. And $\boldsymbol{v}_{glo}, \boldsymbol{t}_{glo}$ is the visual/textual global embeddings, the average pooling of patch/word features, We integrate the aforementioned scores to derive the final significance score, with β serving as a weight parameter.

$$a_i = (1-\beta)a_i^p + \frac{\beta}{2}(a_i^s + a_i^r) \quad (3)$$

3.2.2 Differentiable Decision Matrix

The challenge of selecting significant patches lies in transferring significance scores $\boldsymbol{a} = [a_1, a_2, ..., a_N] \in \mathbb{R}^N$ into a binary decision matrix $\{0, 1\}^N$, which determines whether to select each patch or not. The naive sampling, such as selecting the top-K patches based on the values of significance scores, is non-differentiable and thus hinders the feasibility of end-to-end optimization. To overcome this challenge, we employ the Gumbel-Softmax technique [33] to provide a smooth and differentiable sampling process. The Gumbel-Softmax matrix is derived as:

$$\boldsymbol{M}_{i,l} = \frac{\exp\left(\log\left(\boldsymbol{m}_{i,l} + G_{i,l}\right)/\tau\right)}{\sum_{j=1}^{L}\exp\left(\log\left(\boldsymbol{m}_{i,j} + G_{i,j}\right)/\tau\right)} \quad (4)$$

where $\boldsymbol{M} \in \mathbb{R}^{N\times L}$ and L is the total number of categories (L = 2 for the binary decision, $\boldsymbol{m}_{i,1} = a_i$, $\boldsymbol{m}_{i,2} = 1 - a_i$). $G_i = -\log\left(-\log\left(U_i\right)\right)$ is the Gumbel distribution, U_i is the uniform distribution $(0, 1)$, and τ controls the smoothness of $\boldsymbol{M}$. Finally, we sample from the $\boldsymbol{M}$ with arg-max operation to get the differentiable decision matrix $\boldsymbol{D}$.

$$\boldsymbol{D} = \text{Sampling}(\boldsymbol{M})_{*,1} \in \{0, 1\}^N, \quad (5)$$

where $\boldsymbol{D}$ is obtained as the first column of sampled $\boldsymbol{M}$, which is a one-hot matrix by arg-max operation. Hence, $\boldsymbol{D}$ signifies the outcomes of patch selection: '1' indicates a significant patch, and '0' is a redundant patch. At the training stage, the gradients can be back-propagated to the score prediction network via the differentiable decision matrix.

3.3. Semantic-Spatial Patch Calibration

After selecting significant visual patches with language supervision by Sec. 3.2, we would like to enhance the semantic expression of significant patches, as shown in Fig. 3(c).

We mark the selected significant patches as $V_s = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_{N_s}\} \in \mathbb{R}^{N_s\times d}$, N_s is the number of significant

patches. We use an aggregation network [48] to learn the multiple aggregation weights, and aggregate N_s significant patches to generate N_c informative patches.

$$\hat{\boldsymbol{v}}_j = \sum_{i=1}^{N_s} (\boldsymbol{W})_{ij} \cdot \boldsymbol{v}_i, \quad j = [1, ..., N_c] \tag{6}$$

where $(\boldsymbol{W})_{ij}$ is the elements of the normalized weight matrix $\boldsymbol{W} \in \mathbb{R}^{N_s \times N_c}$. N_c is the number of aggregated patches ($N_c < N_s$), and we have $\sum_{i=1}^{N_s} (\boldsymbol{W})_{ij} = 1$. The weight matrix is $\boldsymbol{W}$ learned by an MLP and softmax function with significant patches as input: $\boldsymbol{W} = \mathrm{Softmax}\left(\mathbf{MLP}\left(V_s\right)\right)$. Especially, we would regard the decision matrix $\boldsymbol{D}$ as the mask matrix to select the significant patch features V_s, before computing the softmax function. The aggregation network can adaptively aggregate patches with similar semantics and is differentiable for end-to-end training.

Although the redundant patches can be dropped directly, they may contain valuable visual semantics for cross-modal alignment, hence we fuse redundant patches into one patch.

$$\hat{\boldsymbol{v}}_f = \sum_{i \in \mathcal{N}} \hat{a}_i \cdot \boldsymbol{v}_i, \quad \hat{a}_i = \frac{\exp\left(a_i\right) \boldsymbol{D}_i}{\sum_{i=1}^{N} \exp\left(a_i\right) \boldsymbol{D}_i}, \tag{7}$$

where $\mathcal{N}$ is the index set for redundant patches, $\hat{a}_i$ is the normalized weights based on significance scores a_i. Finally, we obtain the set of slimmed visual patches $\hat{V} = \{\boldsymbol{v}_{cls}, \hat{\boldsymbol{v}}_1, \cdots, \hat{\boldsymbol{v}}_{N_c}, \hat{\boldsymbol{v}}_f\}$, the [CLS] token always is kept.

3.4. Sparse Patch-Word Alignment

As shown in Fig. 3(d), we compute the fine-grained alignment by the set of slimmed visual patches $\hat{V}$ and initial textual words T. For convenience, we approximate that $|\hat{V}| = N_c$, $|T| = M$. We first calculate the token-wise similarity to generate the patch-word similarity matrix $\boldsymbol{A} \in \mathbb{R}^{N_c \times M}$, where $(\boldsymbol{A})_{ij} = \frac{(\hat{\boldsymbol{v}}_i)^T \boldsymbol{t}_j}{\|\hat{\boldsymbol{v}}_i\| \|\boldsymbol{t}_j\|}$ represents the alignment score between the i-th visual patch and the j-th textual word.

Next, we employ a maximum-correspondence interaction to aggregate the alignment: We first pick up the most aligned textual word (or visual patch) for each patch (or each word). We then calculate the average of these aligned scores to represent the overall alignment score between the image I and the sentence T, denoted $S(I, T)$.

$$S(I,T) = \underbrace{\frac{1}{N_c} \sum_{i=1}^{N_c} \max_j (\boldsymbol{A})_{ij}}_{\text{patch-to-word alignment}} + \underbrace{\frac{1}{M} \sum_{j=1}^{M} \max_i (\boldsymbol{A})_{ij}}_{\text{word-to-patch alignment}}, \tag{8}$$

Following previous methods, we use the bi-direction triplet loss with hard negative mining [9].

$$\begin{aligned} \mathcal{L}_{align} = & \sum_{(I,T)} [\alpha - S(I,T) + S(I,\hat{T})]_+ \\ & + [\alpha - S(I,T) + S(\hat{I},T)]_+, \end{aligned} \tag{9}$$

where α represents a margin parameter, $[x]_+ = \max(x, 0)$, and (I, T) is a positive image-text pair in the mini-batch. We represent $\hat{T} = \mathrm{argmax}_{j \neq \boldsymbol{T}} S(I, j)$ and $\hat{I} = \mathrm{argmax}_{i \neq I} S(i, I)$ as the hardest negative text and image examples within a mini-batch, respectively.

Furthermore, we constrain the ratio of the selected patches to a predefined value ρ for stable training [37], using a mean squared error loss to supervise the process. Finally, we combine the cross-modal alignment loss $\mathcal{L}_{align}$ Eq. (9) with ratio constraint loss $\mathcal{L}_{ratio}$.

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{ratio}, \quad \mathcal{L}_{ratio} = (\rho - \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{D}_i)^2, \tag{10}$$

At the inference stage, instead of using the Gumbel-Softmax sampling, we directly select constant N_s patches according to the values of significance scores. N_s is determined by the selection ratio ρ, $N_s = \rho N$. We use the selected N_s patches to execute the following process without the decision matrix to reduce computations. We also pre-define the aggregation ratio λ, $N_c = \lambda N_s = (\lambda \cdot \rho) N$.

4. Experiments

4.1. Datasets & Metrics

Following the previous works [7, 9, 21], we choose the typical Flickr30K [45] and MS-COCO [29] datasets to train the model, where each image is associated with five texts. Flickr30K contains 29,000, 1,000, and 1,014 training, testing, and validation images. MS-COCO contains 82,738, 5,000, and 5,000 training, testing, and validation images, whose results are tested on averaging over 5-fold of 1K test images and on the full 5K test images. The evaluation metrics are the recall **R@K** (the percentage of ground truth in the retrieved top-K lists, K=1,5,10) and **rSum** (sum of multiple R@K in both image-to-text and text-to-image).

4.2. Implementation Details

We use the Vision Transformer (ViT) [8] (a patch is 16×16 pixels), and Swin Transformer (Swin) [31] (a patch is 32×32 pixels) as the visual encoder, then use the BERT [6] as textual encoders. All encoders are the base version. The image resolutions are 224×224 or 384×384, which get 14×14 and 24×24 patches for ViT (7×7 and 12×12 patches for Swin). Besides, we introduce an additional linear layer on the top of encoders to unify the feature dimension as $d = 512$. The whole framework is trained for 30 epochs with AdamW [32] optimizer, and the margin of triplet loss is $\alpha = 0.2$. The weight parameter $\beta = 0.8$, the

Table 1. Comparisons of image-text retrieval performances on Flickr30K and MS-COCO test-set. We list the details of feature encoders, image resolution, and the number of obtained regions/patches by visual encoders (e.g., 'ViT-Base-224' represents the base–version of Vision Transformer [8] with 224×224 image resolution input, regarding 16×16 pixels as one patch, and getting 14×14 visual patches for one image). FG indicates whether it is the fine-grained cross-modal alignment. The best results are marked **bold**.

Method	FG	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	
***Faster R-CNN + BERT–Base*, 36 pre-computed regions**																						
HREM [14]	✗	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
TGDT [30]	✓	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 [35]	✓	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
***ViT-Base-224 + BERT–base*, 14×14 patches**																						
VSE++ [9]	✗	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 [21]	✓	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 [7]	✓	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 [35]	✓	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	✓	**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**
***ViT-Base-384 + BERT–base*, 24×24 patches**																						
VSE++ [9]	✗	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 [21]	✓	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 [7]	✓	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 [35]	✓	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	✓	**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**
***Swin-Base-224 + BERT–base*, 7×7 patches**																						
VSE++ [9]	✗	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 [21]	✓	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 [7]	✓	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 [35]	✓	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	✓	**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**
***Swin-Base-384 + BERT–base*, 12×12 patches**																						
VSE++ [9]	✗	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 [21]	✓	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 [7]	✓	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 [35]	✓	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	✓	**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**

Table 2. The comparisons of image-text retrieval for Vision-Language Pre-training (VLP) Models. FG indicates whether it is the fine-grained alignment. # represents the zero-shot learning.

Method	FG	Flickr30K 1K				MS-COCO 5K			
		Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
UNITER [5]	✓	87.3	98.0	75.6	94.1	65.7	88.6	52.9	79.9
VILT [19]	✓	83.5	96.7	64.4	88.7	61.5	86.3	42.7	72.9
SOHO [16]	✓	86.5	98.1	72.5	92.7	66.4	88.2	50.6	78.0
ALBEF [24]	✓	95.9	99.8	85.6	97.5	77.6	94.3	60.7	84.3
BLIP [25]	✓	96.6	99.8	87.2	97.5	80.6	95.2	63.1	85.3
***CLIP-ViT-Base-224 + CLIP-BERT–Base*, 14×14 patches**									
CLIP# [36]	✗	81.4	96.2	61.1	85.4	52.3	76.2	33.3	58.2
VSE++ [9]	✗	92.2	99.1	80.5	**95.6**	66.8	88.2	53.6	79.7
SCAN [21]	✓	88.2	98.1	75.3	93.1	65.4	88.0	50.7	77.6
LAPS	✓	**92.9**	**99.3**	**80.6**	**95.5**	**69.8**	**90.4**	**54.3**	**80.0**
***CLIP-ViT-Large-224 + CLIP-BERT–Large*, 16×16 patches**									
CLIP# [36]	✗	85.0	97.7	64.3	87.0	55.9	79.1	35.9	60.9
VSE++ [9]	✗	94.0	99.5	83.4	96.4	68.5	89.4	56.7	81.9
SCAN [21]	✓	90.0	98.5	81.0	95.9	68.0	90.4	53.2	80.7
LAPS	✓	**94.6**	**99.9**	**84.9**	**97.3**	**72.9**	**91.7**	**57.1**	**81.3**

Table 3. The zero-shot evaluation on visual grounding task. All models are trained by CLIP backbones of ViT-B/16 in Flickr dataset (*Vanilla* is untrained). Following ReCLIP [40], we apply the Grad-GAM [39] to select the bounding box from proposals.

Models	RefCOCO			RefCOCO+			RefCOCOg	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
Vanilla CLIP [36]	39.3	45.3	34.2	41.2	47.0	36.8	45.0	45.9
VSE++ [9]	40.7	46.3	33.6	43.2	49.0	35.6	44.2	43.9
SCAN [21]	41.8	47.3	44.4	43.2	49.3	36.8	45.2	46.0
SGR [7]	41.4	48.0	34.2	44.1	49.7	36.7	45.5	46.3
LAPS	**44.2**	**49.9**	**38.4**	**46.7**	**52.3**	**41.6**	**51.3**	**51.2**

selection ratio $\rho = 0.5$ and aggregation ratio $\lambda = 0.4$ for ViT backbones ($\rho = 0.8$ and $\lambda = 0.6$ for Swin backbones).

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

Following the standard protocols [9, 47] on two benchmarks, we list the details of feature encoders and cross-modal alignment types for all compared methods. We introduce four typical cross-modal alignment methods [7, 9, 21, 35], and implement them with their official codes:

- **VSE++** [9], the basic coarse-grained alignment method, learns a common embedding space for images and texts, then computes the cosine similarity between embeddings.
- **SCAN** [21], the basic fine-grained alignment method, computes the bi-directional cross-attention between visual regions and textual words to aggregate similarities.
- **SGR** [7], uses a similarity reason module on the SCAN [21] by graph attention network to learn a similarity score, and compute the single text-to-image alignment.
- **CHAN** [35], applies the hard coding on the SCAN [21], which selects the max-alignment attention scores.

As shown in Tab. 1, we present quantitative results on Flickr30K and MS-COCO datasets. LAPS outperforms all state-of-the-art methods by an impressive margin. It's worth noting that for different transformer-based visual encoders, previous fine-grained methods [7, 21, 35] get inferior results compared to simple coarse-grained methods [9], especially on more number of visual patches (e.g., ViT [8] with

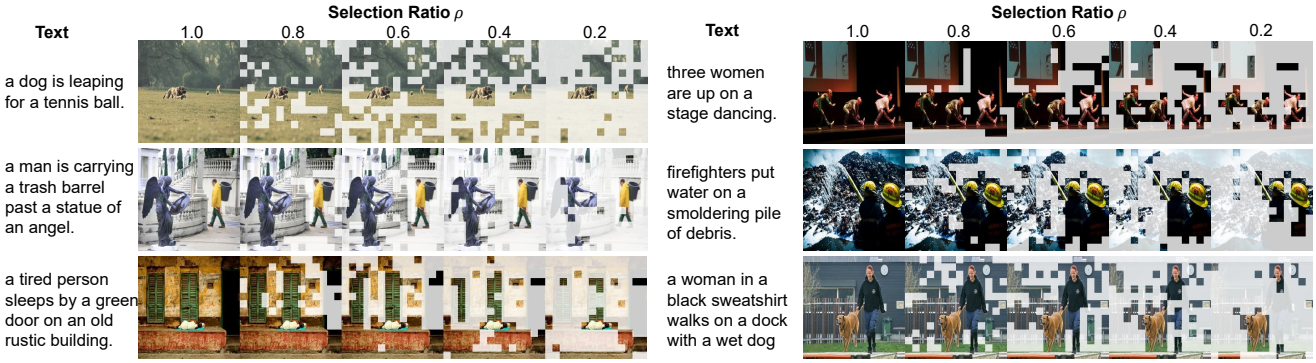

Figure 4. The visualization of selected patches with associated texts under different selection ratios ρ on Flickr30K. Our framework can gradually focus on more salient and text-relevant areas in images as selection ratios decrease and have better interpretability.

Figure 5. The visualization of selected patches with the different language contexts and supervision. The texts below the images describe the semantic contents of images with various perspectives.

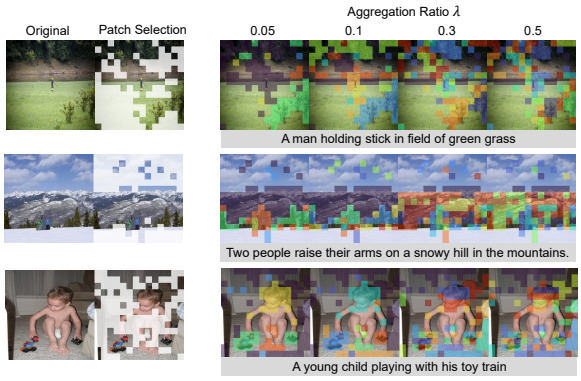

Figure 6. The visualization of aggregated patches with the different aggregation ratios λ. The patches are merged into complete regions with clear semantics and have better interpretability.

384 image resolution) These results diverge from previous observations on detector-based frameworks (e.g., Faster-RCNN [38]). It shows traditional fine-grained alignments are incompatible with patch features, whereas LAPS can address the problem and boost performances.

Besides, we extend our framework to the classical VLP model, CLIP [36] in Tab. 2. We also compare with current state-of-the-art VLP models [5, 19, 24]. The previous fine-grained alignment methods [21] still struggle to achieve satisfactory results even with VLP backbones. LAPS brings large improvements and exhibits competitive performances compared to the mainstream VLP models.

LAPS can be seen as a foundation model to solve the widespread problem of visual redundancy and ambiguity in fine-grained semantic alignment. It works well on cross-modal retrieval but also adapts to more fine-grained recognition tasks. Following previous work [40], we evaluate the cross-modal alignment capability on the visual grounding task in Tab. 3. It shows LAPS significantly outperforms existing methods on all visual grounding benchmarks.

4.4. Ablation Study

We conduct extensive ablation studies and robustness analyses to examine the effectiveness of LAPS. By default, we perform experiments on Vision Transformer with 224 image resolution (*ViT-Base-224 + BERT-base*).

Selection & Aggregation Ratio. We show the impact of selection ratios ρ and aggregation ratios λ on different visual encoders in Fig. 7. Our framework can effectively slim visual patches to enhance cross-modal alignment. The excessive patch selection and aggregation with small ratios also will hurt the performances, especially on Swin [31] encoders, employing local attention within a small window.

Module Gain. To better verify the effectiveness of our LAPS, we provide a comprehensive ablation study in Tab. 4. It shows that the patch selection and patch calibration modules play important roles in semantic alignments. Language supervision is significant for selecting text-relevant visual patches and aggregating redundant patches is help-

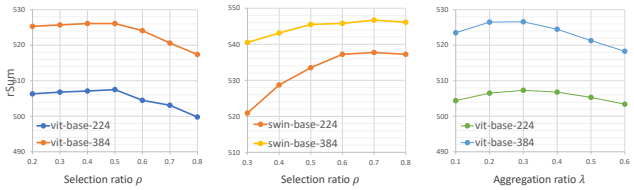

Figure 7. The comparison of different selection ratios ρ and aggregation ratios λ with various visual encoders on Flick30K.

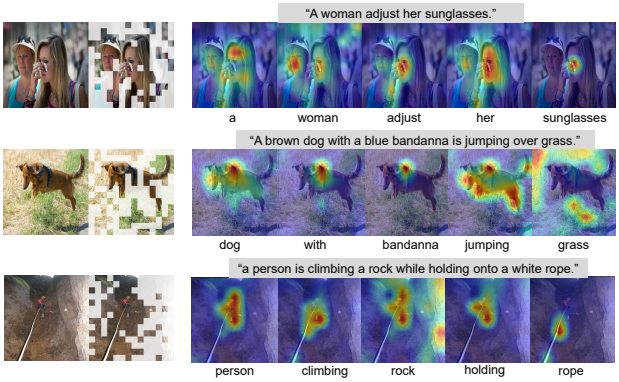

Figure 8. The visualization of fine-grained patch-word alignment with each linguistic word. We show the alignment maps by the gradient-weighted attention [1, 2] on original images.

ful. We find using a clustering algorithm like DPC-KNN [46] to aggregate patches will drop performances, which also has low efficiency. The sparse mechanism is suitable for patch-word alignment, and the bi-directional alignment is optimal. Besides, LAPS is an independent framework for the alignment methods and can be combined with them in a plug-and-play fashion. When replacing the sparse alignment with SCAN [21], the models also get improved performances (69.5/56.4 $\rightarrow$ 71.3/60.8 on R@1 for SCAN).

Word Slimming. Our framework tries to address the problems of visual redundancy and ambiguity. And the challenges are equally important for textual modality. As shown in Tab. 4, we introduce the word slimming process (including the selection and aggregation) to our framework. We find the complementary word slimming will hinder the semantic alignment and drop the performance, since textual tokens usually have high information density [6]. First, texts consist of discrete words created by humans. Compared to pixel images, texts have higher semantic characteristics in nature. Hence textual redundancy is weaker than images. Besides, the text lengths of typical datasets are short (an average of 10 words in COCO/Flickr). The capability of LAPS can not be released in existing datasets.

4.5. Visualization

Patch Selection. We present the visualization results under various selection ratios ρ as Fig. 4. It is evident that our framework can effectively identify significant visual patches based on textual contexts. Besides, we show visualizations under different textual contexts for the same image in Fig. 5. Our patch slimming is highly relevant to language supervision, and can adaptively eliminate redundant patches according to different linguistic descriptions.

Patch Calibration. We present the aggregation visualization of significant patches with various aggregation ratios λ as Fig. 6. For a better view, we give each patch a unique label based on the maximum aggregation weights to represent the aggregated results. Our framework can effectively merge patches with different semantic granularity.

Patch-Word Alignment. We show the visualization of the alignment scores between visual patches and textual words as Fig. 8. It shows our framework exhibits a reasonable and interpretable patch-word semantic alignment. The distinct word-to-patch alignments are centralized and precise on the corresponding visual areas in images.

Table 4. Comparison of different module ablations for our framework on Flickr30K. We also show the results of the word slimming (selection + aggregation) of textual modality for our framework.

Modules	Different Settings	IMG → TEXT		TEXT → IMG	
		R@1	R@5	R@1	R@5
LPS	without patch selection process	69.2	91.9	58.5	84.9
	without language-context	71.1	92.2	59.4	85.5
	only attentive scores	73.5	93.1	61.9	86.8
SPC	without patch calibration process	70.4	91.3	58.9	85.3
	without redundant fusion	73.5	93.2	61.1	87.2
	use the clustering algorithm [46]	68.4	88.5	57.0	82.6
SPA	replace with SCAN alignment [21]	71.3	91.4	60.8	85.6
	only patch-to-word alignment	70.9	90.8	58.9	85.1
	only word-to-patch alignment	72.7	92.5	60.3	86.4
	introduce word selection	70.1	90.3	57.5	82.7
	introduce word aggregation	71.3	91.6	58.8	84.3
	introduce word selection & aggregation	67.7	88.2	55.1	80.5
	Complete LAPS	**74.0**	**93.4**	**62.5**	**87.3**

5. Conclusion

In this paper, we introduce a novel Linguistic-Aware Patch Slimming framework (LAPS) for cross-modal alignment, which is the first work that explicitly focuses on patch-word alignment on pure transformer-based architectures to solve patch redundancy and ambiguity problems. LAPS identifies significant visual patches with language supervision and then rectifies the semantic and structural information to construct more accurate and consistent alignment. Extensive experiments on various benchmarks and visual encoders demonstrate the superiority of our framework.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62222212, 62336001.

References

[1] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, 2021. 8

[2] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021. 8

[3] 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. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12652–12660, 2020. 3

[4] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Chang Lian Wang. Learning the best pooling strategy for visual semantic embedding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15784–15793, 2021. 2

[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*, abs/1909.11740, 2019. 3, 6, 7

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 4, 5, 8

[7] Haiwen Diao, Ying Zhang, Lingyun Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. *ArXiv*, abs/2101.01368, 2021. 1, 3, 5, 6

[8] 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, 3, 4, 5, 6

[9] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 1, 2, 5, 6

[10] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, pages 396–414. Springer, 2022. 3

[11] Zheren Fu, Yan Li, Zhendong Mao, Quan Wang, and Yongdong Zhang. Deep metric learning with self-supervised ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1370–1378, 2021. 2

[12] Zheren Fu, Zhendong Mao, Chenggang Yan, An-An Liu, Hongtao Xie, and Yongdong Zhang. Self-supervised synthesis ranking for deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4736–4750, 2021.

[13] Zheren Fu, Zhendong Mao, Bo Hu, An-An Liu, and Yongdong Zhang. Intra-class adaptive augmentation with neighbor correction for deep metric learning. *IEEE Transactions on Multimedia*, 2022. 2

[14] 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. 1, 6

[15] Dalu Guo, Chang Xu, and Dacheng Tao. Image-question-answer synergistic network for visual dialog. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10426–10435, 2019. 1

[16] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 6

[17] Chaoya Jiang, Haiyang Xu, Wei Ye, Qinghao Ye, Chenliang Li, Ming Yan, Bin Bi, Shikun Zhang, Fei Huang, and Songfang Huang. Bus: Efficient and effective vision-language pre-training with bottom-up patch summarization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2900–2910, 2023. 3

[18] Chaoya Jiang, Haiyang Xu, Wei Ye, Qinghao Ye, Chenliang Li, Ming Yan, Bin Bi, Shikun Zhang, Ji Zhang, and Fei Huang. Copa: Efficient vision-language pre-training through collaborative object-and patch-text alignment. *arXiv preprint arXiv:2308.03475*, 2023. 3

[19] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3, 6, 7

[20] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Mengshu Sun, Wei Niu, Xuan Shen, Geng Yuan, Bin Ren, Minghai Qin, et al. Spvit: Enabling faster vision transformers via soft token pruning. *arXiv preprint arXiv:2112.13890*, 2021. 3

[21] 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 (ECCV)*, pages 201–216, 2018. 1, 3, 5, 6, 7, 8

[22] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 3

[23] Haoxuan Li, Yi Bin, Junrong Liao, Yang Yang, and Heng Tao Shen. Your negative may not be true negative: Boosting image-text matching with false negative elimination. In *ACM Multimedia*, 2023. 2

[24] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2, 3, 6, 7

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 6

[26] Kunpeng Li, Yulun Zhang, K. Li, Yuanyuan Li, and Yun Raymond Fu. Visual semantic reasoning for image-

text matching. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4653–4661, 2019. 1

[27] Sheng Li, Zhiqiang Tao, Kang Li, and Yun Raymond Fu. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3:297–312, 2019. 1

[28] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 3, 4

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[30] 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. 6

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 3, 4, 5, 7

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[33] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 4

[34] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022. 3, 4

[35] 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, 3, 6

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6, 7

[37] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 3, 4, 5

[38] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 1, 7

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[40] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 6, 7

[41] 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. 3, 4

[42] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 3

[43] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. Learning dual semantic relations with graph attention for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:2866–2879, 2021. 2

[44] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2964–2972, 2022. 3, 4

[45] 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

[46] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 8

[47] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15661–15670, 2022. 3, 6

[48] Zhuofan Zong, Kunchang Li, Guanglu Song, Yali Wang, Yu Qiao, Biao Leng, and Yu Liu. Self-slimmed vision transformer. In *European Conference on Computer Vision*, pages 432–448. Springer, 2022. 5