GOAL: Global-local Object Alignment Learning

Hyungyu Choi[1*], Young Kyun Jang[2*], Chanho Eom[1†]

Chung-Ang University[1] Meta AI[2]

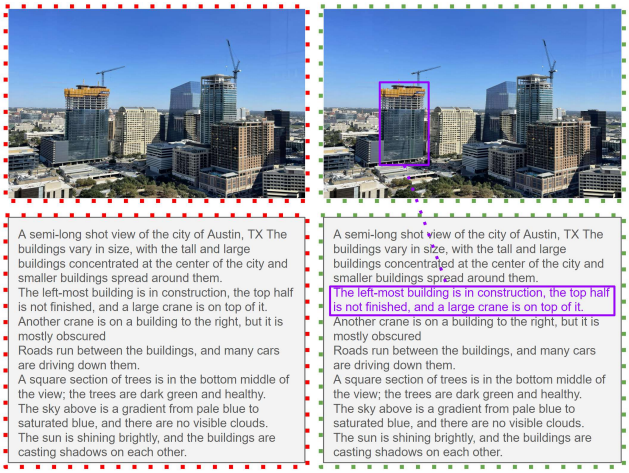
https://github.com/PerceptualAI-Lab/GOAL

Abstract

Vision-language models like CLIP have shown impressive capabilities in aligning images and text, but they often struggle with lengthy and detailed text descriptions because of their training focus on short and concise captions. We present GOAL (Global-local Object Alignment Learning), a novel fine-tuning method that enhances CLIP's ability to handle lengthy text by leveraging both global and local semantic alignments between image and lengthy text. Our approach consists of two key components: Local Image-Sentence Matching (LISM), which identifies corresponding pairs between image segments and descriptive sentences, and Token Similarity-based Learning (TSL), which efficiently propagates local element attention through these matched pairs. Evaluating GOAL on three new benchmarks for image-lengthy text retrieval, we demonstrate significant improvements over baseline CLIP fine-tuning, establishing a simple yet effective approach for adapting CLIP to detailed textual descriptions. Through extensive experiments, we show that our method's focus on local semantic alignment alongside global context leads to more nuanced and representative embeddings, particularly beneficial for tasks requiring fine-grained understanding of lengthy text descriptions.

1. Introduction

After the emergence of CLIP [21], numerous methods [19][36][4][14] have been proposed to bridge the modality gap between images and text showcasing significant advancements. By aligning hundreds of millions of image-caption pairs through contrastive learning, CLIP successfully encodes images and text into a unified embedding space. The resulting distribution of image and text embeddings captures both visual and linguistic semantics, enabling zero-shot transfer to various downstream tasks, such as classification [24][8][25][32] and retrieval [12][29][22], while achieving decent performance.



(a) CLIP (b) GOAL

Figure 1. Comparison of CLIP and our GOAL's capability in handling image-text alignment. (a) CLIP is limited to global image-text matching, treating the entire image and full caption as single units without detailed associations. (b) GOAL can establish precise local alignments between specific regions in the image and their corresponding textual descriptions in the caption (highlighted in purple).

However, fine-tuning a pre-trained CLIP (Fig. 1 (a)) model for specific domains faces limitations, as CLIP is trained on general, short captions (*e.g.*, 77 tokens in the vanilla model) that focus on high-level image concepts. When tasked with longer, more detailed text, CLIP struggles to capture nuanced information, as the unified embedding space is optimized for concise descriptions. This makes adapting CLIP for retrieval tasks requiring lengthy text challenging without architectural adjustments or specialized training techniques.

In this paper, we propose a novel but simple fine-tuning method for image and lengthy text pairs, called **G**lobal-local

*Authors contributed equally. †Corresponding author.

Object **A**lignment **L**earning (GOAL) (Fig. 1 (b)). Here, we refer to *"global"* as the entire image or text and *"local"* as a sub-part, such as a segment of the image or a specific sentence in the text. The idea behind GOAL is to enable the encoder model to focus on the dominant local elements within each image and text sample, thereby enhancing the overall understanding of the sample and producing a more representative embedding.

GOAL has two key components: First, **L**ocal **I**mage-**S**entence **M**atching (LISM), a pipeline that extracts local segments from images and matches them with corresponding descriptive sentences from the entire caption. Second, we introduce **T**oken **S**imilarity-based **L**earning (TSL), a method that effectively propagates attention of local element using the local pairs obtained through the LISM pipeline. To address the challenge of image-lengthy text retrieval, we propose new benchmarks, evaluating GOAL on three diverse datasets (DOCCI [20], DCI [27], and Urban1k [37]) containing image-lengthy caption pairs, and demonstrating substantial fine-tuning performance compared to the original CLIP tuning. The main contributions of our work can be summarized as follows:

- We propose GOAL, a fine-tuning approach that enhances CLIP's understanding of local elements within samples to improve embedding representations.
- GOAL includes two components: Local Image-Sentence Matching (LISM) for generating pseudo local pairs, and Token Similarity-based Learning (TSL) for efficient propagation the attention of local elements.
- Through experiments on newly proposed benchmarks, we show that GOAL significantly improves performance over the original CLIP and baseline models.

2. Related Work

Vision-Language Pre-training. Research on addressing alignment differences between vision and language modalities has brought the Contrastive Language-Image Pre-training (CLIP) [21] model into the spotlight. CLIP, a multi-modal embedding model trained through contrastive learning on over 400 million image-text pairs, effectively aligns visual and textual representations while demonstrating remarkable zero-shot capabilities. Following its success, larger pre-training models emerged, such as ALIGN [10] and Florence [35], trained on image-text pairs from datasets containing 1.8B and 900M samples, respectively. However, these models typically rely on short, broad image descriptions as captions, causing them to miss crucial local-level detailed information. While Long-CLIP [37] addressed this limitation by utilizing synthetic lengthy captions generated by multi-modal LLMs [33][30][7][6], it requires an expensive data preparation process. To overcome this limitation more efficiently, we present a fine-tuning method that enhances CLIP's ability to capture both local-detail and global-semantic information by training it on a dataset containing detailed, multi-sentence captions.

Utilizing Local Elements in Vision-Language Model Training. In terms of vision-language alignment models, using local elements' knowledge to improve the model's general ability has been widely explored across various domains. Visual-Textual Attributes Alignment (ViTAA) [31] learns to align full-person images corresponding to the global-level with text describing the whole person to perform a person re-identification task [26][3][39][40], while also learning to align the image and text for attributes (*e.g.*, hair, pants, shoes) that correspond to the local-level. This approach combines global-local relations, enabling richer visual-language representation learning. CLOC (Contrastive Localized Language-Image Pre-Training) [1] builds 2 billion image-text datasets and uses them for pre-training models by matching local objects and phrase-levels through Open-vocabulary Detector (*e.g.*, OWLv2 [18], GLIPv2 [38]) models to improve localization capabilities while maintaining CLIP's global-level representation, demonstrating superior performance compared to the original pre-trained CLIP model. In contrast, our proposed GOAL method efficiently learns global-local relationships through fine-tuning with significantly fewer datasets and computational resources compared to large-scale pre-training approaches.

3. Method

In this section, we introduce Local Image-Sentence Matching (LISM), a pipline that generates local-level pseudo pairs from a given image-caption pair (Sec. 3.1). We then present the Token Similarity-based Learning (TSL) method, which leverages these pseudo pairs to address global-level biases in CLIP [21] (Sec. 3.2).

3.1. Local Image Sentence Matching

We propose Local Image-Sentence Matching (LISM) Fig. 2, which separates a given caption into individual sentences and identifies corresponding image segments, matching each sentence with its relevant segment. To this end, we first decompose a given caption T_g, which provides detailed descriptions of a given image I_g, into individual sentences, resulting in text segments $\{T_{l,i}\}_{i=1}^{M}$, where M is the number of sentences. We then leverage SAM [11] to segment the image I_g into semantic units, obtaining masks for individual objects along with the background. We expand each mask into a rectangular bounding box that includes the surrounding area, allowing us to leverage contextual information for matching with the caption. As a result, we obtain a set of local images, $\{I_{l,i}\}_{i=1}^{N}$, where N represents the number of local regions. Note that in this process, we filter out

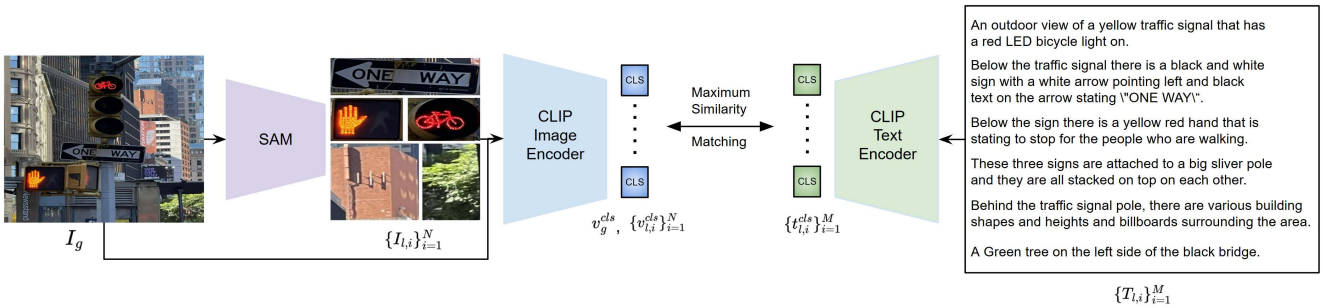


Figure 2. Overview of Local Image-Sentence Matching (LISM) pipeline. Given a global image and its detailed caption, LISM uses SAM to segment the image into local regions and splits the caption into individual sentences. These local pairs are then processed through CLIP encoders to obtain CLS embeddings, which are used for maximum similarity matching to identify the most relevant image-sentence pairs.

segments smaller than 1% of the total image area to exclude very small objects and reduce noise from SAM.

We use CLIP [21] to match the decomposed caption segments with the corresponding image segments. Specifically, we extract the CLS token embeddings for each local text segment $T_{l,j}$ from the text encoder of CLIP, ϕ_t as follows:

$$\{t_{l,i}^{cls}\}_{i=1}^{M} = \phi_t(\{T_{l,i}\}_{i=1}^{M}). \tag{1}$$

Similarly, for both the original image I_g and each image segment $I_{l,i}$, we extract the CLS token embeddings from the visual encoder of CLIP as follows:

$$v_g^{cls} = \phi_v(I_g), \quad \{v_{l,i}^{cls}\}_{i=1}^{N} = \phi_v(I_{l,i}). \tag{2}$$

Next, we compute the cosine similarity between each local text embedding $t_{l,i}^{cls}$ and the global image embedding v_g^{cls} or the local image embeddings $\{v_{l,i}^{cls}\}_{i=1}^{N}$. Among all matched pairs, each local text embedding is matched with its highest similarity image embedding. From all these matched pairs, we select the one pair with the highest similarity score and denote it as (I_l, T_l). If the matched image in this selected pair is the global image I_g, we discard this pair. This matching strategy excludes global image matches from the final selection to ensure high-quality local pair associations.

3.2. Token Similarity based Learning

While CLIP's pretraining with image-text pairs effectively learns global alignment, its training with brief captions limits the model's ability to capture fine-grained local details from lengthy descriptions. To address this, we propose Token Similarity based Learning (TSL) (Fig. 3). Our approach uses local pairs obtained through the LISM pipeline and implements a fine-tuning strategy that effectively propagates local-level information. Specifically, TSL maximizes the similarity between patch tokens of local regions in the global image and their corresponding local image embeddings, while applying the same principle to text by increasing the similarity between sequence tokens of local parts in the global text and their corresponding local text embeddings. To implement this strategy, we need to extract both local and global features from the input pairs. Using CLIP's vision encoder ϕ_v and text encoder ϕ_t, we extract both local and global features as follows: For the local text T_l:

$$t_l^{cls} = \phi_t(T_l) \in \mathbb{R}^d, \tag{3}$$

where t_l^{cls} represents the last layer CLS token embedding. For the global text T_g, the text encoder extracts:

$$S_g = \phi_t(T_g) \in \mathbb{R}^{M \times d}, \tag{4}$$

where M is the sequence length of T_g, and S_g represents the last layer sequence tokens of T_g. To handle text sequences longer than CLIP's standard 77 token limit, we adopt Long-CLIP's [37] positional embedding interpolation method in our text encoder. For the local image I_l, we obtain:

$$v_l^{cls} = \phi_v(I_l) \in \mathbb{R}^d, \tag{5}$$

where v_l^{cls} represents the last layer CLS token embedding. For the global image I_g, the vision encoder extracts:

$$P_g = \phi_v(I_g) \in \mathbb{R}^{N \times d}, \tag{6}$$

where N denotes the number of patch tokens in I_g , d is the embedding dimension and P_g represents the last layer patch tokens of I_g. We process both global and local pairs through shared CLIP encoders to learn both types of features simultaneously. This weight sharing ensures consistent encoding in the shared embedding space. Let $\mathcal{T}$ denote the set of token indices corresponding to the local text segment. We can identify the sequence tokens in S_g that correspond to T_l:

$$S_m = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} S_g[i] \in \mathbb{R}^d, \tag{7}$$

where $|\mathcal{T}|$ denotes the number of selected sequence tokens. The aggregated features are then projected into a shared embedding space, where both text and image representations are aligned:

$$\hat{S}_l = proj(S_m) \in \mathbb{R}^d, \tag{8}$$

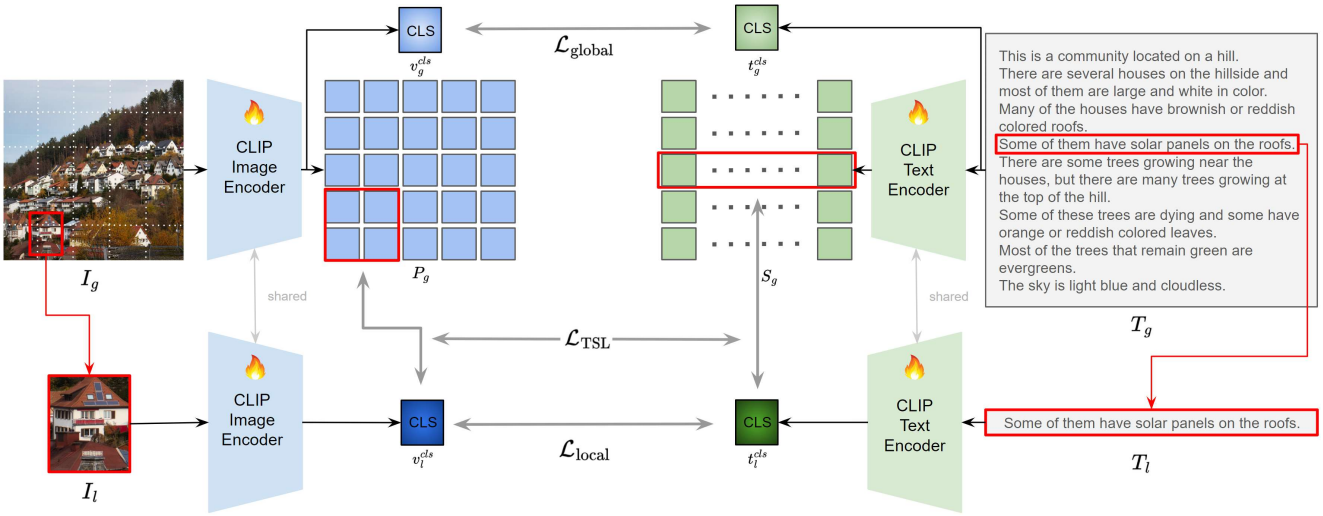


Figure 3. Overview of Token Similarity based Learning (TSL). The framework processes global image-text pairs and their local pairs through shared CLIP encoders, extracting patch and sequence tokens. TSL identifies and projects corresponding token regions to match local CLS embeddings, enabling attention on local element.

where $proj(\cdot)$ represents a learned projection function.

Given that each local image region I_l has its bounding box coordinates (x_1, y_1, x_2, y_2) obtained from LISM in the global image I_g, we can leverage this spatial information to identify specific patch tokens from P_g that correspond to the local image region, filtering out patches from other parts of the global image. Let $\mathcal{B}$ denote the set of indices of patch tokens located inside the bounding box. We aggregate these tokens using average pooling to capture comprehensive information from the selected region:

$$P_m = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} P_g[i] \in \mathbb{R}^d, \tag{9}$$

where $|\mathcal{B}|$ denotes the number of selected patch tokens. The aggregated features are then projected into a shared embedding space where both text and image representations are aligned:

$$\hat{P}_l = proj(P_m) \in \mathbb{R}^d, \tag{10}$$

where $proj(\cdot)$ represents a learned projection function. We train our model with multiple objectives combined into a final loss function:

$$\mathcal{L}_{\text{total}} = \lambda_{global}\mathcal{L}_{\text{global}} + \lambda_{local}\mathcal{L}_{\text{local}} + \lambda_{TSL}\mathcal{L}_{\text{TSL}}, \tag{11}$$

where λ is a hyperparameter controlling the contribution of local alignment. We apply contrastive learning at both global and local levels, adopting the contrastive learning used in CLIP . At the global level:

$$\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{contrast}}(v_g^{cls}, t_g^{cls}), \tag{12}$$

where v_g^{cls} and t_g^{cls} are the CLS token embeddings of the global image I_g and global text T_g, respectively. This global alignment ensures that the model maintains CLIP's original capability to capture global relationships between image-text pairs. Similarly, for local-level contrastive learning:

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{contrast}}(v_l^{cls}, t_l^{cls}), \tag{13}$$

where v_l^{cls} and t_l^{cls} are the CLS token embeddings of the local image I_l and local text T_l, respectively. By applying contrastive learning to local CLS token pairs, we encourage precise alignment between local image regions and their corresponding textual descriptions, enabling the model to learn cross-modal relationships.

The token similarity loss $\mathcal{L}_{\text{TSL}}$ maximizes the similarity between projected tokens and their corresponding local CLS token embeddings for both image and text:

$$\mathcal{L}_{\text{TSL}} = \text{MSE}(sim(\hat{P}_l, v_l^{cls}), \mathbf{1}) + \text{MSE}(sim(\hat{S}_l, t_l^{cls}), \mathbf{1}), \tag{14}$$

where $sim(\cdot)$ denotes a function that computes an $n \times n$ similarity matrix with n being the batch size, and $\mathbf{1}$ is a $n \times n$ matrix with ones on its diagonal entries. By optimizing this loss, the model learns to maximize the similarity between local CLS token embeddings and their corresponding regions in global tokens. This token-level alignment strategy enables the model to attention on local element, enhancing fine-grained understanding capabilities. This fine-tuning method effectively addresses CLIP's inherent limitation in capturing local details from lengthy descriptions, which stems from its pre-training with brief captions. Through the combination of token-level similarity learning and global-local contrastive learning, our approach enables comprehensive understanding of cross-modal relationships with attention on local element from detailed text descriptions.

4. Experiments

In this section, we present our experimental setup in Sec. 4.1. Our ablation study in Sec. 4.2 demonstrates the effectiveness of each component in our framework through experiments. We provide zero-shot experimental results in Sec. 4.3 to show our model's generalization capability across different datasets. Finally, we present qualitative analysis in Sec. 4.4 through visualization of attention maps.

4.1. Experimental setup

Dataset. We conduct experiments on three datasets: DOCCI [20], DCI [27] and Urban1k [37], each containing images with long and detailed captions, designed to enable vision-language models to learn fine-grained visual-textual relationships. The DOCCI dataset consists of 9,647 training samples and a combined test set of 5,100 samples (5,000 from the test set and 100 from the qualification-test set). Since DCI's original test set contains only 100 samples, we instead sampled 2,000 examples from its training set of 7,805 samples to create a larger test set, establishing a train-test ratio similar to DOCCI. For both datasets, we generate pseudo local pairs through our LISM. The datasets and our sampled test sets used in this research are publicly available on GitHub[1].

Training setting. To validate our approach, we conduct experiments using two different CLIP [21] backbone architectures: ViT-B/16, and ViT-L/14 [28][5]. Both models are fine-tuned for 10 epochs with a batch size of 16. We set the balance hyperparameters in the total loss function as $\lambda_{global} = 1$, $\lambda_{TSL} = 1$, and $\lambda_{local} = 0.5$ to maintain strong global and TSL learning while moderating the contribution of local loss. The training was performed on a single NVIDIA RTX 4090 GPU for base models and an NVIDIA A6000 GPU for the ViT-L/14 model, taking approximately 1 and 2 hours respectively.

Test setting. To handle the long text sequences during inference, we adopt the positional embedding interpolation technique from Long-CLIP [37]. We evaluate our method on two different test scenarios: the original test set and our proposed global-local test set. For the original test set, we evaluate Text-to-Image (T2I) and Image-to-Text (I2T) retrieval performance using Recall@k. For the second scenario, we create a pseudo global-local test set by applying our proposed LISM to the original test set. Specifically, we generate local pairs for each image-text pair in the original test set and append the local pair with the highest similarity score to create the pseudo global-local test set. For this extended test set, we using mAP@k as our evaluation metric since we need to evaluate retrieval performance in situations with multiple correct answers in our global-local matching scenario. Both global and local texts are considered correct answers when querying with either global or local images, and similarly, both global and local images are considered correct answers when querying with either type of text.

[1] https://github.com/PerceptualAI-Lab/GOAL/tree/main/datasets

4.2. Ablation study

We conduct ablation studies to validate the effectiveness of our proposed GOAL framework. Table 1 and Table 2 present the results on DOCCI and DCI test sets, respectively. We compare four different settings: (1) global fine-tuning with only $\mathcal{L}_{\text{global}}$, (2) local fine-tuning with only $\mathcal{L}_{\text{local}}$, (3) w/o TSL with both $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{local}}$ without TSL, and (4) our complete GOAL framework with all loss terms.

The results demonstrate the superiority of our framework across all settings. On the DOCCI dataset with ViT-L/14, GOAL achieves 84.37% R@1 for text-to-image retrieval, surpassing the w/o TSL by 12.87% (74.75%), global fine-tuning by 14.01% (74.00%), and local fine-tuning by 25.20% (67.39%). Similar improvements are observed on the DCI dataset, where GOAL with ViT-L/14 achieves 76.89% R@1, outperforming the w/o TSL by 15.83% (66.38%), global fine-tuning by 16.98% (65.73%), and local fine-tuning by 42.70% (53.88%). When combined with our proposed TSL method in the complete GOAL framework, we observe consistent improvements across both datasets, demonstrating the effectiveness of our approach.

We evaluate the methods on a global-local joint test set. Table 3 and Table 4 present mAP@10 scores for both text-to-image (T2I) and image-to-text (I2T) retrieval tasks on DOCCI and DCI datasets, respectively. The results demonstrate our GOAL framework's capability to effectively handle both global and local feature matching simultaneously. Specifically, on the DOCCI dataset with ViT-L/14, GOAL achieves 69.53% mAP@10 for T2I, surpassing the w/o TSL (66.55%) and global fine-tuning (65.79%) for T2I. Similar improvements are observed on the DCI dataset, where GOAL with ViT-L/14 achieves 64.77% and 64.11% for T2I and I2T, respectively, compared to w/o TSL 58.60% and 59.85%. These results show that our approach successfully preserves CLIP's global understanding while incorporating local feature matching capabilities, leading to improved performance on both global and local matching tasks.

4.3. Comparison to the state of the art

We compare our method with Long-CLIP in zero-shot settings across both datasets. For fair comparison, we evaluate fine-tuning methods trained on one dataset and tested on the other (zero-shot), alongside models fine-tuned on the test dataset. In Table 5, our GOAL method fine-tuned on DOCCI outperforms Long-CLIP when tested on the DCI dataset in most metrics, achieving 68.93% vs 67.88% in text-to-image R@1 and 68.43% vs 64.08% in image-

Backbone	Methods	Loss			Text to Image Recall@K				Image to Text Recall@K			
		Global	Local	TSL	R@1	R@5	R@25	R@50	R@1	R@5	R@25	R@50
ViT-B/16	Global fine-tuning	✓			72.41	93.27	99.31	99.76	72.04	93.37	99.35	99.80
	Local fine-tuning		✓		65.82	89.96	98.37	99.39	65.73	90.35	98.35	99.51
	w/o TSL	✓	✓		72.08	93.73	99.24	99.82	71.80	93.57	99.29	99.76
	GOAL	✓	✓	✓	**79.47**	**96.65**	**99.69**	**99.92**	**79.43**	**96.14**	**99.61**	**99.90**
ViT-L/14	Global fine-tuning	✓			74.00	93.84	99.04	99.67	73.55	93.94	99.16	99.78
	Local fine-tuning		✓		67.39	90.67	98.16	99.20	66.33	90.41	98.10	99.43
	w/o TSL	✓	✓		74.75	94.31	99.12	99.71	74.55	94.37	99.27	99.78
	GOAL	✓	✓	✓	**84.37**	**97.55**	**99.76**	**99.98**	**82.57**	**97.37**	**99.82**	**99.98**

Table 1. Original test set results on DOCCI dataset. Comparison of retrieval performance across different fine-tuning approaches using ViT-B/16 and ViT-L/14 models. The evaluation metrics include both text-to-image and image-to-text Recall@K. The best and second-best scores for each method are marked in **bold** and underlined, respectively.

Backbone	Methods	Loss			Text to Image Recall@K				Image to Text Recall@K			
		Global	Local	TSL	R@1	R@5	R@25	R@50	R@1	R@5	R@25	R@50
ViT-B/16	Global fine-tuning	✓			66.43	84.74	93.80	96.10	66.58	84.74	95.10	97.65
	Local fine-tuning		✓		59.38	78.49	90.70	93.85	58.18	78.74	90.05	93.75
	w/o TSL	✓	✓		66.63	84.04	93.75	96.05	66.43	85.29	95.00	97.75
	GOAL	✓	✓	✓	**72.64**	**89.89**	**95.95**	**97.25**	**72.84**	**90.50**	**96.60**	**97.90**
ViT-L/14	Global fine-tuning	✓			65.73	84.24	93.25	96.30	65.73	86.04	94.65	96.25
	Local fine-tuning		✓		53.88	75.54	87.84	91.75	51.63	72.64	87.49	91.10
	w/o TSL	✓	✓		66.38	84.44	93.40	96.30	66.23	86.04	94.75	96.50
	GOAL	✓	✓	✓	**76.89**	**91.05**	**96.55**	**97.75**	**76.59**	**91.20**	**96.55**	**98.25**

Table 2. Original test set results on DCI dataset. Comparison of retrieval performance across different fine-tuning approaches using ViT-B/16 and ViT-L/14 models. The evaluation metrics include both text-to-image and image-to-text Recall@K. The best and second-best scores for each method are marked in **bold** and underlined, respectively.

Backbone	Method	Loss			mAP	
		Global	Local	TSL	T2I	I2T
ViT-B/16	Global fine-tuning	✓			59.03	58.40
	Local fine-tuning		✓		57.62	57.16
	w/o TSL	✓	✓		60.74	59.99
	GOAL	✓	✓	✓	**63.27**	**62.63**
ViT-L/14	Global fine-tuning	✓			65.79	64.97
	Local fine-tuning		✓		62.55	62.87
	w/o TSL	✓	✓		66.55	**66.58**
	GOAL	✓	✓	✓	**69.53**	66.34

Table 3. Comparison of different methods using ViT-B/16 and ViT-L/14 backbones on DOCCI dataset's global and local joint test set. Results show mAP@10 scores for both text-to-image (T2I) and image-to-text (I2T) retrieval tasks. The best and second-best scores for each method are marked in **bold** and underlined, respectively.

Backbone	Method	Loss			mAP	
		Global	Local	TSL	T2I	I2T
ViT-B/16	Global fine-tuning	✓			53.68	54.32
	Local fine-tuning		✓		52.66	53.04
	w/o TSL	✓	✓		56.68	56.35
	GOAL	✓	✓	✓	**57.19**	**57.35**
ViT-L/14	Global fine-tuning	✓			55.36	58.32
	Local fine-tuning		✓		52.69	54.46
	w/o TSL	✓	✓		58.60	59.85
	GOAL	✓	✓	✓	**64.77**	**64.11**

Table 4. Comparison of different methods using ViT-B/16 and ViT-L/14 backbones on DCI dataset's global and local joint test set. Results show mAP@10 scores for both text-to-image (T2I) and image-to-text (I2T) retrieval tasks. The best and second-best scores for each method are marked in **bold** and underlined, respectively.

to-text R@1 with ViT-L/14 backbone. The improvement is more pronounced in the ViT-B/16 setting, where our method achieves 64.13% vs 61.33% in text-to-image R@1 and 65.88% vs 60.03% in image-to-text R@1.

In Table 6, our fine-tuning method on DCI demonstrates strong zero-shot performance compared to Long-CLIP when tested on the DOCCI dataset. With ViT-L/14, GOAL notably outperforms Long-CLIP in higher rank metrics, achieving 95.78% vs 95.25% in R@5, 99.55% vs 99.19% in R@25 for text-to-image retrieval. The improvement is particularly significant in image-to-text retrieval, where GOAL substantially surpasses Long-CLIP across all metrics, achieving 79.16% vs 66.82% in R@1 and 95.96% vs 91.90% in R@5. These results demonstrate that our GOAL fine-tuning method exhibits robust generalization capability and superior performance in zero-shot settings across different datasets, with particularly strong improvements in image-to-text retrieval.

Backbone	Method	Text to Image (Recall@K)				Image to Text (Recall@K)			
		R@1	R@5	R@25	R@50	R@1	R@5	R@25	R@50
ViT-B/16	Long-CLIP	61.33	80.79	91.65	94.35	60.03	81.44	92.80	95.05
	GOAL DOCCI fine-tuning	**64.13**	**82.69**	**92.95**	**95.40**	**65.88**	**83.44**	**92.95**	**95.65**
	GOAL DCI fine-tuning	72.64	89.89	95.95	97.25	72.84	90.50	96.60	97.90
ViT-L/14	Long-CLIP	67.88	83.29	91.80	94.80	64.08	84.84	93.35	95.75
	GOAL DOCCI fine-tuning	**68.93**	**85.74**	**93.95**	**96.00**	**68.43**	**85.99**	**93.90**	**96.25**
	GOAL DCI fine-tuning	76.89	91.05	96.55	97.75	76.59	91.20	96.55	98.25

Table 5. Comparison of different methods using ViT-B/16 and ViT-L/14 backbones on DCI dataset. Results show Text-to-Image and Image-to-Text Recall@K scores in zero-shot setting. The best scores for each method are marked in **bold**.

Backbone	Method	Text to Image (Recall@K)				Image to Text (Recall@K)			
		R@1	R@5	R@25	R@50	R@1	R@5	R@25	R@50
ViT-B/16	Long-CLIP	**71.63**	92.16	**98.90**	**99.73**	63.29	88.80	98.39	99.45
	GOAL DCI fine-tuning	71.22	**92.39**	**98.90**	99.61	**72.18**	**92.88**	**98.88**	**99.55**
	GOAL DOCCI fine-tuning	79.47	96.65	99.69	99.92	79.43	96.14	99.61	99.90
ViT-L/14	Long-CLIP	78.84	95.25	99.19	99.59	66.82	91.90	99.04	99.82
	GOAL DCI fine-tuning	**79.04**	**95.78**	**99.55**	**99.84**	**79.16**	**95.96**	**99.61**	**99.90**
	GOAL DOCCI fine-tuning	84.37	97.55	99.76	99.98	82.57	97.37	99.82	99.98

Table 6. Comparison of different methods using ViT-B/16 and ViT-L/14 backbones on DOCCI dataset. Results show Text-to-Image and Image-to-Text Recall@K scores in zero-shot setting. The best scores for each method are marked in **bold**.

Backbone	Method	Image to Text (Recall@K)			
		R@1	R@5	R@25	R@50
ViT-B/16	CLIP	68.90	88.80	97.90	99.50
	Long-CLIP	79.20	94.80	99.10	**99.70**
	GOAL DOCCI fine-tuning	81.90	95.80	**99.40**	**99.70**
	GOAL DCI fine-tuning	**82.90**	**96.80**	**99.40**	**99.70**
ViT-L/14	CLIP	68.20	88.40	97.00	98.70
	Long-CLIP	82.60	96.70	**99.60**	**100.00**
	GOAL DOCCI fine-tuning	86.30	96.50	99.40	**100.00**
	GOAL DCI fine-tuning	**89.80**	**97.80**	**99.60**	**100.00**

Table 7. Comparison of different methods using ViT-B/16 and ViT-L/14 backbones on Urban1k dataset. Results show Text-to-Image and Image-to-Text Recall@K scores in zero-shot setting. The best scores for each method are marked in **bold**.

Our experiments on the Urban1k dataset Table 7 demonstrate the effectiveness of our approach across fine-tuning methods and pre-trained CLIP. The results show that with the ViT-B/16 backbone, GOAL achieves notable improvements, with GOAL DCI fine-tuning reaching 82.90% in R@1, surpassing Long-CLIP (79.20%) and baseline CLIP (68.90%) by significant margin. The performance gains are even more pronounced with the ViT-L/14 backbone, where GOAL DCI fine-tuning achieves 89.80% in R@1, outperforming Long-CLIP (82.60%) and CLIP (68.20%). Both GOAL variants (DOCCI and DCI fine-tuning) demonstrate competitive performance compared to other fine-tuning methods across recall metrics (R@1, R@5, R@25, R@50), with notable improvements particularly in R@1, which is a crucial metric for retrieval performance. This consistent performance enhancement demonstrates the robustness of our approach in handling image-to-text retrieval tasks, regardless of the backbone architecture used.

Additionally, in the supplementary material Sec. B and Sec. C, we provide further analysis of our method's ability to preserve global understanding through zero-shot classification experiments on standard benchmarks. We also include extended evaluations comparing our method with BLIP2 [15], and present zero-shot performance results on

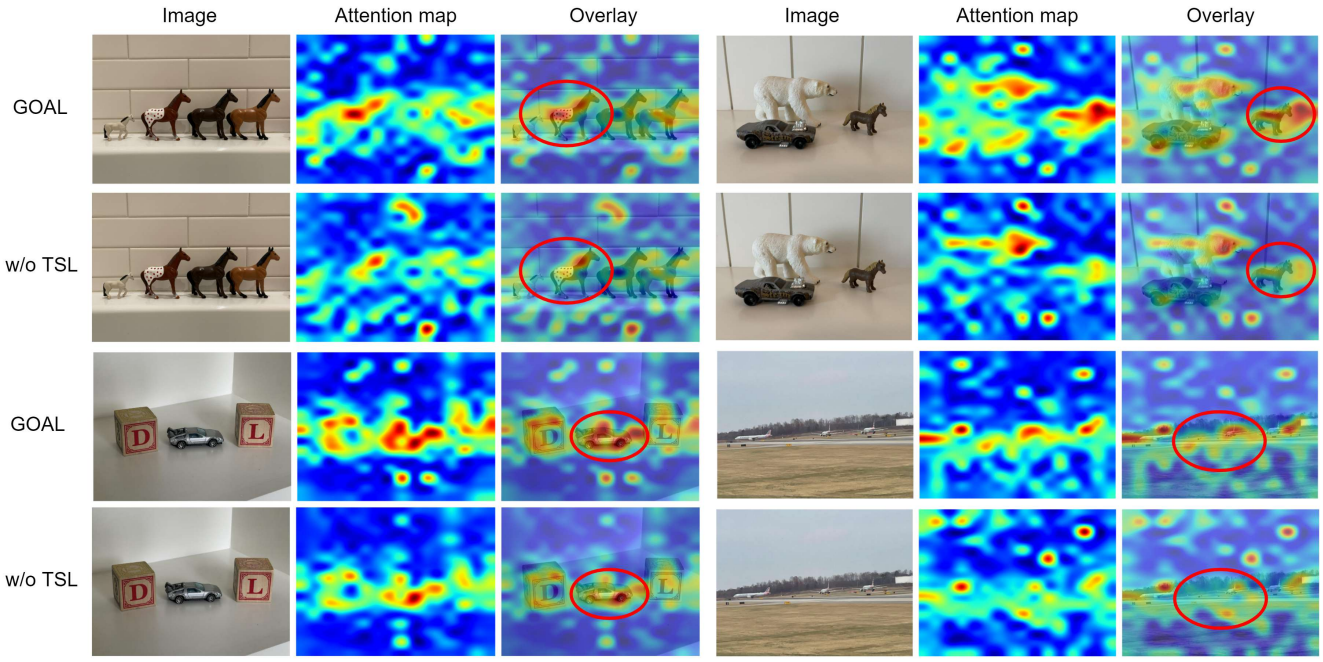


Figure 4. Comparison of attention maps generated by GOAL and w/o TSL methods. For each row pair, we present three components: (1) original input image (left), (2) attention heatmap visualization (middle), and (3) overlay of attention on the original image (right). The examples demonstrate how GOAL achieves more focused attention compared to the baseline w/o TSL method. Red circles in the overlay highlight regions where GOAL shows particularly effective attention localization.

diverse datasets including COCO [16], Flickr30k [34], and ShareGPT4V [2] to further demonstrate the generalization capabilities of our approach.

4.4. Qualitative results

We provide qualitative comparisons of attention maps generated by our GOAL and the w/o TSL approach in Fig. 4. The visualization [41][23] shows that our GOAL framework captures local details more precisely compared to the w/o TSL. The attention maps clearly show that GOAL consistently focuses on specific objects within the images with higher precision. For instance, in the image containing multiple toy animals, GOAL's attention map shows clear activation across each individual animal figure, while the w/o TSL's attention is more dispersed and partially activated on irrelevant background regions. This enhanced attention behavior demonstrates that GOAL successfully maintains CLIP's global understanding, while incorporating local feature learning through our TSL method. These qualitative results further support our quantitative findings, showing that our fine-tuning method effectively preserves global comprehension while significantly improving the model's ability to attention on local element within the scene.

5. Conclusion

In this paper, we have proposed a novel fine-tuning method GOAL that improves CLIP's understanding in image and lengthy text pair datasets. First, Local Image Sentence Matching (LISM) has produced pseudo local pairs through global pairs. Second, Token Similarity based Learning (TSL) has effectively overcome CLIP's limitation of focusing primarily on high-level representations by leveraging attention mechanisms between global and local tokens. Through this research, we have established a foundation for various multi-modal models that perform image-text alignment to effectively learn from lengthy and detailed textual descriptions of images.

6. Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00355008) and the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program (IITP-2024-RS-2024-00418847) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation.

References

[1] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training. arXiv preprint arXiv:2410.02746, 2024. 2

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023. 8, 1

[3] Zhenyu Cui, Jiahuan Zhou, Xun Wang, Manyu Zhu, and Yuxin Peng. Learning continual compatible representation for re-indexing free lifelong person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16614–16623, 2024. 2

[4] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10995–11005, 2023. 1

[5] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 5, 1

[6] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075, 2024. 2

[7] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. Vita: Towards open-source interactive omni multimodal llm. arXiv preprint arXiv:2408.05211, 2024. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1

[9] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15262–15271, 2021. 3

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning, pages 4904–4916. PMLR, 2021. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023. 2

[12] Satwik Kottur, Ramakrishna Vedantam, José MF Moura, and Devi Parikh. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4985–4994, 2016. 1

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[14] 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. 1

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023. 7, 3

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 8, 1

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023. 1

[18] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems, 36, 2024. 2

[19] Sangwoo Mo, Minkyu Kim, Kyungmin Lee, and Jinwoo Shin. S-clip: Semi-supervised vision-language learning using few specialist captions. Advances in Neural Information Processing Systems, 36:61187–61212, 2023. 1

[20] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. arXiv preprint arXiv:2404.19753, 2024. 2, 5, 1

[21] 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. 1, 2, 3, 5

[22] Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. Learning relation alignment for calibrated cross-modal retrieval. arXiv preprint arXiv:2105.13868, 2021. 1

[23] 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. 8

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 1

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE

conference on computer vision and pattern recognition, pages 2818–2826, 2016. 1

[26] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17127–17137, 2024. 2

[27] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26700–26709, 2024. 2, 5, 1

[28] A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017. 5

[29] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6439–6448, 2019. 1

[30] Xiong Wang, Yangze Li, Chaoyou Fu, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. arXiv preprint arXiv:2411.00774, 2024. 2

[31] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pages 402–420. Springer, 2020. 2

[32] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017. 1

[33] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549, 2023. 2

[34] 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. 8, 1

[35] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432, 2021. 2

[36] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11975–11986, 2023. 1

[37] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In European Conference on Computer Vision, pages 310–325. Springer, 2025. 2, 3, 5, 1

[38] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. Advances in Neural Information Processing Systems, 35:36067–36080, 2022. 2

[39] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1367–1376, 2017. 2

[40] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1318–1327, 2017. 2

[41] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016. 8