FELGA: Unsupervised Fragment Embedding for Fine-Grained Cross-Modal Association

Yaoxin Zhuo, Baoxin Li
Arizona State University
Tempe, AZ, USA
{yzhuo6, baoxin.li}@asu.edu

Abstract

*Vision-and-Language Pre-trained (**VLP**) models have demonstrated their powerful zero-shot ability in multiple downstream tasks. Most of these models are designed to learn joint embeddings of images and their paired sentences, with both modalities considered globally. This does not lead to optimal solutions for applications where what matters more is the local-level cross-modal association, such as the situation where a user may want to retrieve images with query words that link to only small parts of the images. While a VLP model could in principle be retrained to learn a new embedding capturing such fine-grained association, expensive annotation would be needed, making it impractical for big data applications. This paper proposes a novel method named Fragment Embedding by Local and Global Alignment (**FELGA**), which learns fragment-level embeddings that capture fine-grained cross-modal association through utilizing visual entity proposals and semantic concept proposals in an unsupervised manner. Comprehensive experiments conducted on three VLP models and two datasets demonstrate that **FELGA** is not limited to specific VLP models and outperforms the original VLP features. In particular, the learned embeddings support cross-modal fragment association tasks including query-driven object discovery and description assignment.*

1. Introduction

Recent years have witnessed phenomenal growth of multi-modal data on social media platforms. A common task on such platforms is cross-modal retrieval, such as searching for images or videos based on a query textual description. The key to support such cross-modal tasks is the development of some unified representations that can facilitate the association of entities in different modalities. Recent Vision-and-Language Pre-trained (**VLP**) models [4, 14, 24–27, 29, 32, 39, 55, 61–63, 66] have emerged as a powerful way for learning such cross-modal representations. However, current mainstream VLP models are typically designed to learn joint embeddings of images and their paired textual descriptions, with both modalities considered globally. This would inadvertently neglect semantic relations occurring only at the fragment level. Consequently, the learned embeddings may not optimally support tasks that rely on fine-grained cross-modal association.

There are some recent efforts that attempted to address this issue to some extent. For example, RegionCLIP [63], an extension of CLIP [39], focused on learning the region-level visual representations. The limitation is that textual descriptions are still considered globally as only the visual encoder is trained at region level. On the other hand, GLIP [29] and GLIPv2 [62] unify object detection and phrase grounding for object-level visual representations. But they require costly manual annotations for the bounding boxes and thus can hardly scale to large training data.

In an attempt to address the above challenges in supporting fine-grained cross-modal association tasks, we propose the Fragment Embedding by Local and Global Alignment (**FELGA**) to learn the embeddings of visual entities (objects) and semantic concepts (keywords) in an unsupervised manner. **FELGA** leverages an image fragment generator to provide region proposals and a keyword detector to provide semantic concepts. It then learns the fragment embeddings through local and global level alignments, which are guided by the designed pseudo-label matrices.

Our contributions are summarized as follows. (1) We introduce the two fine-grained cross-modal association tasks to validate the performance of learned fragment embeddings. (2) We propose the **FELGA** method that is able to learn the fragment embeddings without requiring expensive manual annotations. (3) Comprehensive experiments and results demonstrate that the fragment embeddings learned by **FELGA** can outperform the reference VLP models, hence establishing a new baseline for unsupervised fragment embedding learning. We emphasize that **FELGA** is not restricted to specific VLP models and can be extended to open real-world image-text data like Tweets.

2. Related Work

Unsupervised Object Discovery Unsupervised object discovery [46] (**UOD**) task refers to identifying objects in images without supervised annotations during training (no annotations of object bounding boxes or class labels). In testing, it requires the method to find the image regions that contain objects and clustering images that contain the same objects, whose evaluation metrics are known as the Correct Localization (CorLoc) and Correct Retrieval (CorRet). Cho *et al.* [5] work on discovering the object instances with distinctive parts by part-based proposals matching method. Vo *et al.* [48] used self-supervised features to construct a pipeline that treats the UOD as a ranking problem and scales it to large-scale datasets. TokenCut [50] is a graph-based method that utilizes unsupervised transformer features to discover objects. Indeed, these UOD methods have primarily focused on learning visual embeddings and may not explicitly incorporate the learning of text embeddings, which limits their capability to perform multi-modal retrieval.

Open-Vocabulary Object Detection Object Detection [68] is traditionally formulated under the closed-vocabulary set settings. Recently, some object detection works aimed at generalizing the limited number of classes to detect novel unseen classes during the testing stage. OVR-CNN [60] is a two-stage training framework. It first constructs the visual-semantic space by image-caption pairs and then learns object detection through base-classes object-level annotation data. Vision Transformer for Open-World Localization (**OWL-ViT**) [34] transferred contrastively trained image-text models to detection by the designed attaching heads with limited object-level data. BARON [51] was proposed to cluster the contextually related regions into a bag. Then it aligned the bag-of-region representations of the object detector and VLP models. Even though these open-vocabulary object detection works support detecting objects in unseen classes, they still require seen classes supervised annotations, *i.e.* object-level bounding boxes and class labels during training. Besides that, they are only able to solve the text-to-image direction query-driven object discovery task at a certain level and not able to work for description assignment task.

Cross-Modal Retrieval There are a lot of works on cross-modal retrieval in recent years. Some of them are focused on learning embeddings with better performance [12, 16, 43, 49, 54]. Others prefer learning hashing codes that have less storage space and faster searching speed [8, 10, 30, 45, 58, 67]. Andrej *et al.* [18] designs a neural network to learn the embedding space based on fragments to further reason the representations of image and text. Lee *et al.* [22] proposed the Stacked Cross Attention Network (**SCAN**) that attends the words in the sentence with respect to image regions to form the attended sentence vector. Wu *et al.* [52] proposed the Self-Attention Embeddings (**SAEM**) that use FasterRCNN and WordPiece models to extract the salient image regions and sentence tokens and then feed them into the self-attention [47] layers to get feature vectors for entire images and sentences. IRRA [15] designed the implicit relation reasoning module to explore the part alignments. Most of these approaches encapsulate the entirety of both images and texts, hence lacking the capacity for supporting inference with local-level association or only supported single-direction (text-to-visual) retrieval.

Vision-Language Unified Representation Learning Since the transformer [47] architecture appears, VLP models became popular in recent years [9]. VisualBERT [28] utilizes the designed masked prediction and object tags as anchor points to pre-train the model to learn the embedding. To avoid the restriction of a fixed number of object categories, CLIP [39] leverages large-scale image-text pairs to learn visual and language representations. ALBEF [27] designs the contrastive loss and momentum distillation to align the visual tokens and word tokens. BLIP [26] focuses on dealing with the noisy image-text pairs from the web by the designed filter. GLIP [29] reformulates object detection as a phrase grounding task by aligning each bounding box to phrases in the text prompt. CoCa [57] is a new design of the pre-training model that combines contrastive loss and captioning loss during training.

Some works also paid attention to the fine-grained fragments during pre-training. GLIPv2 [62] unifies the localization task and understanding task into one framework and utilizes the region-words pairs for pre-training. Ge *et al.* [11] considered the fine-grained local association between nouns/verbs and videos during pre-training. Zhou [66] trains the model with un-parallel image-text data by three granularities: tag alignment, phrase alignment, and entire image-sentence alignment. ALPRO [23] introduces a video-text contrastive loss to align unimodal features by the generated soft entity labels. G-ViLM [53] involves spatiotemporal grounding and temporal grouping during feature learning with the local region-noun alignment design. They require large-scale datasets [1, 3, 35, 36] or even expensive manual annotated datasets [21, 41] for pre-training.

However, the most important thing is these VLP works are still focused on learning the embeddings of the entire image and text. They cannot provide the embeddings of the fragments (visual entities and semantic concepts) directly for the cross-modal fragment association tasks.

3. Method

To facilitate the discussion on the benefit of fragment embeddings, we first introduce the two fine-grained cross-modal association tasks: (text to image) query-driven object discovery (Object Discovery) and the task of assigning the

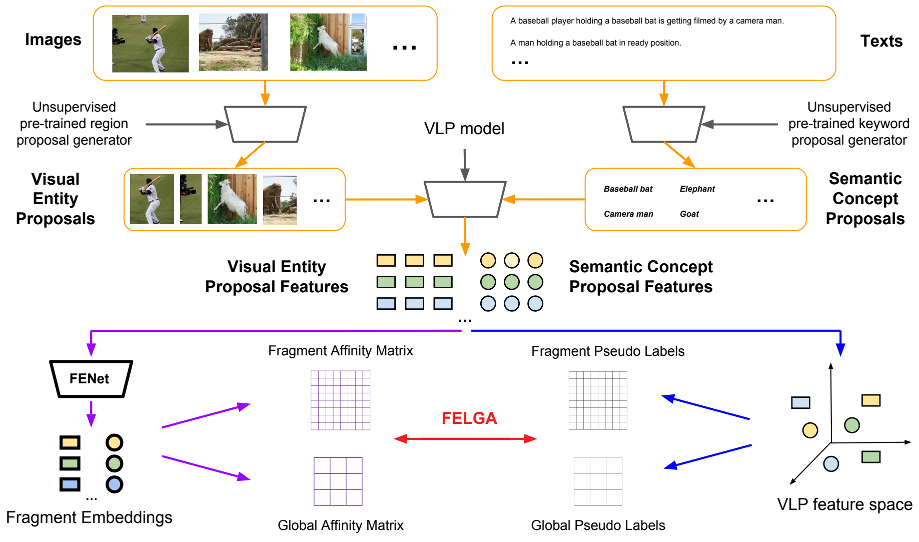
Images
...
Texts
A baseball player holding a baseball bat is getting filmed by a camera man.
A man holding a baseball bat in ready position.
...
Unsupervised pre-trained region proposal generator
Unsupervised pre-trained keyword proposal generator
VLP model
Visual Entity Proposals
Semantic Concept Proposals
Baseball bat
Elephant
Camera man
Goat
Visual Entity Proposal Features
Semantic Concept Proposal Features
FENet
Fragment Affinity Matrix
Fragment Pseudo Labels
FELGA
Fragment Embeddings
Global Affinity Matrix
Global Pseudo Labels
VLP feature space

Figure 1. The framework of the proposed **FELGA** (best viewed in color).

most relevant description (Description Assignment) (image to text) in section 3.1. Then we introduce the proposed unsupervised method **FELGA** in section 3.2.

3.1. Problem Definition

In real applications, a user may want to retrieve images with query words linking to only small parts of the images. While the training set may contain images with paired descriptive sentences, we define the task of query-driven object discovery as finding images that contain visual entities associated with given semantic concepts (but not the entire sentences). In the other direction, we define the task of description assignment as: given visual entities (not the whole image), to find the descriptive sentences that contain the associated semantic concepts. Given: N images $\mathbb{I} = \{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, ..., \mathbf{I}_N\}$ and texts $\mathbb{T} = \{\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, ..., \mathbf{T}_N\}$. The n-th image $\mathbf{I}_n$ has m_n^v visual entities $V_n = \{v_1^n, v_2^n, v_3^n, ..., v_{m_n^v}^n\}$ and the n-th text $\mathbf{T}_n$ has m_n^t semantic concepts $T_n = \{t_1^n, t_2^n, t_3^n, ..., t_{m_n^t}^n\}$. Object Discovery retrieves images that contain visual entities associated with given query semantic concepts. Description Assignment retrieves relevant texts that contain the semantic concepts associated with given query visual entities. The unsupervised setting means that there is no annotation of visual entities and semantic concepts (and their correspondence) provided in the training stage.

3.2. FELGA:Fragment Embedding by Local and Global Alignment

3.2.1 Fragment Proposal Extraction

Given N pairs of images $\mathbb{I} = \{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, ..., \mathbf{I}_N\}$ and texts $\mathbb{T} = \{\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, ..., \mathbf{T}_N\}$, we have no access to the fine-grained annotations of the visual entities $V = \bigcup_{n=1}^{N} V_n$ in the images and the semantic concepts $T = \bigcup_{n=1}^{N} T_n$ in sentences in the unsupervised setting. To overcome this limitation and better deal with the huge amount of naturally paired image-text data like Tweets, we apply the unsupervised pre-trained models to extract the proposals for each image and text. For the n-th image $\mathbf{I}_n$, the unsupervised pre-trained region proposal generator will provide $\bar{m}_n^v$ region proposals as the visual entity proposals $\bar{V}_n = \{\bar{v}_1^n, \bar{v}_2^n, \bar{v}_3^n, ..., \bar{v}_{\bar{m}_n^v}^n\}$. For the n-th text $\mathbf{T}_n$, the unsupervised keyword extraction will extract $\bar{m}_n^t$ keywords from sentences as the semantic concept proposals $\bar{T}_n = \{\bar{t}_1^n, \bar{t}_2^n, \bar{t}_3^n, ..., \bar{t}_{\bar{m}_n^t}^n\}$.

3.2.2 Fragment Pseudo Label Construction

In a supervised setting, we could access the visual entities V and semantic concepts T and assume the knowledge of their association by the association label matrix L^{local}:

$$L_{i,j}^{\text{local}} = \begin{cases} 1, & \text{if } v_i \text{ is associated with } t_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where i is the index of the visual entity in the current batch images, ranging in $[1, \sum_{n=1}^{N} m_n^v]$, and j is the index of semantic concept in the current batch text, ranging in $[1, \sum_{n=1}^{N} m_n^t]$. If the visual entity v_i is associated with the semantic concept t_j, the association label $L_{i,j}^{\text{local}}$ will be 1. Otherwise, it is 0, which means they are not related.

In the unsupervised setting of this paper, we cannot assume knowing the visual entities or the semantic concepts, let alone the association label matrix L^{local}. Following the steps in subsection 3.2.1, we start by building up the fragment (local-level) pseudo label matrix $\bar{L}^{\text{local}}$ for the frag-

Method name	Microsoft COCO 2017		Microsoft COCO 2014	
	Object Discovery	Description Assignment	Object Discovery	Description Assignment
CLIP [39] based on entire image/sentence	50.7	38.6	50.7	39.3
CLIP [39] model based on proposals	50.7	35.6	50.7	36.2
Dense	50.9	39.2	50.5	39.5
Sparse	59.4	40.3	54.7	41.0
Dynamic	57.5	41.0	56.9	41.7
FELGA	**63.1**	**42.2**	**63.6**	**42.8**

Table 1. The mAP results of different methods with CLIP model. **Notes**: the gray means the distance is obtained by the feature of the entire image or sentence.

ment proposals $\bar{V}$ and $\bar{T}$. Firstly, we use a pre-trained VLP model to extract their feature vectors:

$$F_{\bar{V}}, F_{\bar{T}} = \text{VLP model}(\bar{V}, \bar{T}) \tag{2}$$

where the VLP model will take the different modalities' proposals as input and output their features. For visual entity proposals, we crop the corresponding small regions from the original image and feed them into the visual encoder of the VLP model to get their features $F_{\bar{V}} = \{f_{\bar{v}_1^n}, f_{\bar{v}_2^n}, f_{\bar{v}_3^n}, ..., f_{\bar{v}^n_{\bar{m}_n^v}}\}$. The semantic concepts proposals are first tokenized and then fed into the language encoder of the VLP model to get their features $F_{\bar{T}} = \{f_{\bar{t}_1^n}, f_{\bar{t}_2^n}, f_{\bar{t}_3^n}, ..., f_{\bar{t}^n_{\bar{m}_n^v}}\}$. These steps of proposal extraction and corresponding feature extraction are shown as orange arrows in Figure 1.

Then we construct the fragment pseudo label matrix $\bar{L}_{i,j}^{\text{local}}$ for the proposals based on the similarity scores computed in the following ways.

Dense Connection The first way is to assign all fragment pseudo labels to 1 if the proposals are within the same pair as follows:

$$\bar{L}_{i,j}^{\text{local}} = \begin{cases} 1, & \text{if } (\bar{v}_i, \bar{t}_j) \text{ are within same pair} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where i in $[1, \sum_{n=1}^{N} \bar{m}_n^v]$ is the index of the visual entity proposals in the current batch images and j in $[1, \sum_{n=1}^{N} \bar{m}_n^t]$ is the index of semantic concept proposals in the current batch texts. The reason for setting the local-level pseudo labels to 1 for all proposals within a pair is that their similarities should be the highest compared to proposals from different pairs. As a result, setting the pseudo labels to 1 ensures that the proposals within the same pair are aligned, while others are not. Conversely, proposals from different pairs are assigned to the pseudo labels of 0, indicating that they are not aligned.

Sparse Connection Different from "Dense Connection" which assigns all pseudo labels of the fragments within the same pair to 1, "Sparse Connection" selects only the highest relevant visual entity proposal, which has the maximum similarity score with the corresponding semantic concept proposal. Then it assigns the fragment pseudo label to 1:

$$\bar{L}_{i,j}^{\text{local}} = \begin{cases} 1, & \text{if } d(f_{\bar{v}_i}, f_{\bar{t}_j}) = \max\{d(f_{\bar{v}_i}, f_{\bar{t}_j})\}, \\ & \forall i \in [1, \bar{m}_n^v], \\ & \text{and } (\bar{v}_i, \bar{t}_j) \text{ are within same pair} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The main idea is at least one image region should be relevant to the semantic concept in the sentence, implying that they have the highest similarity score with each other. Here we assume that the visual entity proposal with the highest similarity score is associated with the semantic concept proposal.

The above two ways of computing pseudo labels have certain limitations. "Dense Connection" relies on the strong assumption that all proposals within one pair are associated with each other, which may not always be true, while "Sparse Connection" may have a pseudo-label matrix capturing situations where multiple visual entities are associated with the same semantic concept or vice versa, causing difficulties in learning fragment embeddings.

Dynamic Connection To address these limitations, we propose to dynamically construct the fragment pseudo-label matrix. We divide the similarity scores of the same pair proposals into two groups, "Similar Group" and "Dissimilar Group", by a dynamic threshold defined by the local mean$_n$:

$$\text{mean}_n = \frac{\sum_{i=1}^{\bar{m}_n^v} \sum_{j=1}^{\bar{m}_n^t} d(f_{\bar{v}_i}, f_{\bar{t}_j})}{\bar{m}_n^v \times \bar{m}_n^t} \tag{5}$$

where n is the index of image or text in the data batch, and the mean$_n$ is the average of similarities of all visual entity proposals and semantic concept proposals within the n-th image-text pair. The dynamic pseudo-labels are assigned adaptively by the $mean_n$ as follows:

$$\bar{L}_{i,j}^{\text{local}} = \begin{cases} 1, & \text{if } d(f_{\bar{v}_i}, f_{\bar{t}_j}) \geq \text{mean}_n, \\ & \forall n \in [1, N], \\ & \text{and } (\bar{v}_i, \bar{t}_j) \text{ are within same pair} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

This step of constructing the fragment pseudo labels is shown as part of the blue arrows in Figure 1.

Method name	Microsoft COCO 2017		Microsoft COCO 2014	
	Object Discovery	Description Assignment	Object Discovery	Description Assignment
BLIP [26] based on entire image/sentence	45.8	36.3	46.4	36.6
BLIP [26] based on proposals	42.4	39.5	40.1	40.4
Dense	53.3	41.5	54.0	42.0
Sparse	53.7	41.1	50.0	40.8
Dynamic	58.7	44.6	58.6	45.2
FELGA	**66.4**	**46.1**	**66.1**	**46.7**

Table 2. The mAP results of different methods with BLIP model. **Notes**: the gray font means the distance is obtained by the feature of the entire image or sentence.

3.2.3 Global Pseudo Label Construction

Besides the fragment/local-level alignment, we also consider the entire/global-level image-text pair alignment for the fragment embedding learning. We design the global-level pseudo labels as follows:

$$\bar{L}_{k,l}^{global} = \begin{cases} 1, & \text{if } k = l \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where k and l, in $[1, N]$, are respectively indices of image $\mathbf{I}$ and text $\mathbf{T}$. If the image and text are from the same pair, the pseudo label will be 1. Otherwise, the pseudo-label is 0. We designed Equations 7 and 11 (to be introduced in subsection 3.2.4) to make the similarity scores of paired and unpaired data more discriminative. Specifically, We aim to strengthen the maximum similarity within the same pair while weakening the maximum similarity from different pairs. This design is essential for improving the alignment between visual entities and semantic concepts. It ensures that the learned fragment embeddings are more accurate and discriminative. This step of constructing the global pseudo labels is shown as part of the blue arrows in Figure 1.

3.2.4 Models and Loss Functions for Local and Global Alignments

Most VLP models [25–27, 29, 39, 57] adopt dual-encoders for processing image and text separately. Since the proposal features $F_{\bar{V}}$ and $F_{\bar{T}}$ from the pre-trained VLP model are already in the unified representation space, we design a multi-layer perception (MLP) as the **FENet** to learn the fragment embeddings. It takes the features from the VLP model as input to obtain the fragment embedding:

$$\hat{F}_{\bar{V}} = \text{FENet}(\bar{F}_{\bar{V}}), \hat{F}_{\bar{T}} = \text{FENet}(\bar{F}_{\bar{T}}) \tag{8}$$

where $\hat{F}_{\bar{V}} = \{\hat{f}_{\bar{v}_1^n}, \hat{f}_{\bar{v}_2^n}, ..., \hat{f}_{\bar{v}^n_{\bar{m}_n^v}}\}$ are the fragment embeddings of visual entity proposals and $\hat{F}_{\bar{T}} = \{\hat{f}_{\bar{t}_1^n}, \hat{f}_{\bar{t}_2^n}, ..., \hat{f}_{\bar{t}^n_{\bar{m}_n^t}}\}$ are the fragment embeddings of semantic concept proposals. Firstly, the fragment-level affinity matrix of visual entity proposals and semantic concept proposals is constructed as follows:

$$\hat{\mathbf{S}}_{\bar{V}\bar{T}}^{\text{local}} = d(\hat{F}_{\bar{V}}, \hat{F}_{\bar{T}}) = \hat{F}_{\bar{V}} \hat{F}_{\bar{T}}^{\top} \tag{9}$$

where $\hat{\mathbf{S}}_{\bar{V}\bar{T}}^{\text{local}} \in [0, 1]^{\sum_{n=1}^{N} \bar{m}_n^v \times \sum_{n=1}^{N} \bar{m}_n^t}$ represents the similarities of fragment embeddings in current training batch data, and d denotes the cosine similarity of normalized features. We propose to compute the local contrastive loss for all proposal pairs as follows:

$$\mathcal{L}_{\text{con\_l}} = \sum_{i=1}^{\sum_{n=1}^{N} m_n^v} \sum_{j=1}^{\sum_{n=1}^{N} m_n^t} [\log \sum_{\bar{L}_{i,j}^{\text{local}}=1} e^{\lambda_1 - d(\hat{f}_{\bar{v}_i}, \hat{f}_{\bar{t}_j})} + \log \sum_{\bar{L}_{i,j}^{\text{local}}=0} e^{d(\hat{f}_{\bar{v}_i}, \hat{f}_{\bar{t}_j})}] \tag{10}$$

where $\sum_{n=1}^{N} \bar{m}_n^v$ is the total number of visual entity proposals and $\sum_{n=1}^{N} \bar{m}_n^t$ is the total number of semantic concept proposals in the current data batch (N pairs). λ_1 is the hyper-parameter of the threshold. This loss function is inspired by the Lifted Structured Loss [37] and it incorporates all relative relationships of all proposal pairs, regardless of whether they are similar pairs or dissimilar pairs. As this loss function encourages better separation and distinction among clusters in the learned embedding, it facilitates more effective alignment between visual entities and semantic concepts.

Similarity, we construct the global-level affinity matrix $\hat{\mathbf{S}}_{\mathbf{IT}}^{\text{global}}$ by computing the whole image and complete sentence similarity scores as follows:

$$d(\mathbf{I}_n, \mathbf{T}_n) = \max\{d(\hat{f}_{\bar{v}_i}, \hat{f}_{\bar{t}_j})\}, \forall i \in [1, \bar{m}_n^v], \forall j \in [1, \bar{m}_n^t] \tag{11}$$

where $\hat{f}_{\bar{v}_i}$ is the final fragment embedding of the visual entity proposal $\bar{v}_i$ and $\hat{f}_{\bar{t}_j}$ is the final fragment embedding of the semantic concept proposal $\bar{t}_j$. The maximum similarity score $d(f_{\hat{v}_i}, f_{\hat{t}_j})$ was used to represent the entire image-text similarity score due to the focus of the most relevant fragments. Based on that, the global contrastive loss for the

Method name	Microsoft COCO 2017		Microsoft COCO 2014	
	Object Discovery	Description Assignment	Object Discovery	Description Assignment
BLIPv2 [25] based on entire image/sentence	55.3	35.1	55.6	35.5
BLIPv2 [25] based on proposals	52.6	39.1	52.2	39.6
Dense	56.8	42.9	57.1	43.3
Sparse	64.2	42.0	59.3	41.9
Dynamic	62.7	43.7	62.4	44.2
FELGA	**69.7**	**45.5**	**69.9**	**46.3**

Table 3. The mAP results of different methods with BLIPv2 model. **Notes**: the gray means the distance is obtained by the feature of the entire image or sentence.

entire image-text pairs is designed as follows:

$$\mathcal{L}_{\text{con\_g}} = \sum_{k=1}^{N}\sum_{l=1}^{N}[\log \sum_{\bar{L}^{\text{global}}_{k,l}=1} e^{\lambda_2 - d(\mathbf{I}_k,\mathbf{T}_l)} + \log \sum_{\bar{L}^{\text{global}}_{k,l}=0} e^{d(\mathbf{I}_k,\mathbf{T}_l)}] \tag{12}$$

where N is the total number of pairs in the current training batch, and λ_2 is the hyper-parameter of the threshold. Within the same image-text pair, this loss function emphasizes the importance of the proposal pair that has the highest similarity score. For proposal pairs from different image-text pairs, this loss function suppresses the significance of the pair with the highest similarity score.

Finally, **FELGA** train the **FENet** with the following overall loss function for learning the fragment embedding:

$$\mathcal{L} = \mathcal{L}_{\text{con\_l}} + \lambda_3 \mathcal{L}_{\text{con\_g}} \tag{13}$$

where $\mathcal{L}_{\text{con\_l}}$ is guided by Dynamic Connection fragment pseudo labels from Equation 6 and $\mathcal{L}_{\text{con\_g}}$ is guided by global pseudo labels from Equation 7. λ_3 is a hyper-parameter that controls the trade-off between the local-level contrastive loss and global-level contrastive loss.

4. Experiments

4.1. Datasets

Since most traditional cross-modal retrieval datasets like Flickr30k [56], NUS-WIDE [6] and Recipe 1M+ [33, 40] only provide the data in the format of image and sentence pairs, they lack the annotations of the visual entities and semantic concepts. Therefore, we utilize Microsoft COCO [31]. RefCOCO, RefCOCO+ and RefCOCOg [19, 59] datasets were also considered as the training set. However, the referring expressions in these datasets can be viewed as weakly supervised localization annotations. They revealed the relationship information of semantic concepts and visual entities. So we just select the COCO Caption dataset for training. Finally, we use the combination of the Microsoft COCO Captioning dataset and Detection dataset. The Microsoft COCO Captioning dataset serves as the training set, while the Detection dataset is used as the testing set. This combination allows us to evaluate the performance of the proposed method in an unsupervised fashion, where annotations of visual entities and semantic concepts are not available for training. Microsoft COCO 2014 dataset has $82,783$ training images and $40,504$ validation images. Following paper [17], we use the Microsoft COCO 2017 dataset that consists of $118,287$ training images and $5,000$ validation images. For both of them, we use their training images from the Captioning dataset for training and validation images from the Detection dataset for testing. We introduce more details in subsection 4.3.

4.2. Selection of Pre-Trained Models

In fragment proposal extraction (section 3.2.1), **FELGA** uses a region proposal generator and a keyword extractor to provide the visual entity proposals and semantic concept proposals, respectively. For image $\mathbb{I}$, we use DETReg [2] for region proposals, which is pre-trained on ImageNet [7] only. The top-10 proposals are treated as the visual entities proposals $\bar{V}_n$ for n-th image $\mathbf{I}_n$. For the text $\mathbb{T}$, we use the unsupervised KeyBERT [13] model as the keyword proposal generator. The top-10 keywords are treated as the semantic concept proposals $\bar{T}_n$ for n-th text $\mathbf{T}_n$. In fragment pseudo label construction (section 3.2.2), the proposed method requires the VLP model to extract the features of fragment proposals. To verify our method is not limited to any specific VLP models, three VLP models are selected: CLIP [39], BLIP [26] and BLIPv2 [25]. For a fair comparison, all these pre-trained models are not fine-tuned on the Microsoft COCO dataset.

4.3. Training and Testing Settings

In the training stage, the Microsoft COCO Captioning dataset provides the training images $\mathbb{I}$ and texts $\mathbb{T}$. Each image is annotated with five-sentence descriptions, which are concatenated to form the corresponding text $\mathbf{T}$. We do not use any fragment-level annotations, *i.e.*, the bounding boxes of objects, or the object class labels during training.

In the testing stage, to facilitate the query-driven object discovery task, we design a special testing protocol. Rather

Module Name		Microsoft COCO 2017		Microsoft COCO 2014	
Local-level Alignment	Global-level Alignment	Object Discovery	Description Assignment	Object Discovery	Description Assignment
✓		57.5	41.0	56.9	41.7
	✓	59.8	34.6	58.9	35.7
✓	✓	**63.1**	**42.2**	**63.6**	**42.8**

Table 4. Ablation Study of the local-level alignment and global-level alignment design with CLIP model on Microsoft COCO 2017 and 2014 dataset.

Method Name	Microsoft COCO 2017		Microsoft COCO 2014	
	Object Discovery	Description Assignment	Object Discovery	Description Assignment
Dense	50.9	39.2	50.5	39.5
Dense + Global Alignment	59.6	41.7	60.7	41.7
Sparse	59.4	40.3	54.7	41.0
Sparse + Global Alignment	61.0	41.0	59.6	41.6
Dynamic	57.5	41.0	56.9	41.7
Dynamic + Global Alignment (**FELGA**)	**63.1**	**42.2**	**63.6**	**42.8**

Table 5. Ablation Study of adding global-level alignment to different baselines with CLIP model on Microsoft COCO 2017 and 2014 dataset.

than providing the entire text as the query, the 80 labels are treated as semantic concepts and used as the query. The objective is to retrieve the relevant images containing associated visual entity v, *i.e.*, to discover objects within images. The distance between t and $\mathbf{I}$ is computed as the maximum values of $d(\hat{f}_t, \hat{f}_{\bar{v}_i})$, where $\bar{v}_i$ is the visual entities proposals in image $\mathbf{I}$. We do not use $d(\hat{f}_t, \hat{f}_{v_i})$ because in real applications exact object locations are in general not available. For the description assignment task, we crop the bounding boxes of objects as the visual entities and use them as the query. The objective is to retrieve the relevant sentences containing the associated semantic concepts, which are the labels of the objects. Similarly, the distance between v and $\mathbf{T}$ is computed as the maximum values of $d(\hat{f}_v, \hat{f}_{\bar{t}_j})$, where the $\bar{t}_j$ is the semantic concept proposals in $\mathbf{T}$.

For the quantitative evaluation, we select the mean Average Precision (mAP) to evaluate the performance of the learned fragment embedding. In the object discovery experiment, the average precisions (APs) of all semantic concepts are averaged to get mAP. For the description assignment experiment, the average precisions (APs) of all visual entities are averaged to get mAP.

4.4. Baselines

Since the baseline VLP models were designed for generating features of entire images rather than fragments, we also provide their performance based on the entire image (not visual entity proposals) in the object discovery task and the performance based on the entire sentence (not semantic concept proposals) in the description assignment task.

In section 3.2.2, we propose multiple ways to construct the fragment pseudo labels. The "Dense Connection", "Sparse Connection" and "Dynamic Connection" can be applied directly to guide the training of the **FENet** by setting $\lambda_3 = 0$ in Eqn. 13. We name them as "Dense", "Sparse" and "Dynamic". They are viewed as baselines to be compared with **FELGA**.

4.5. Additional Implementation Details

The **FENet** consists of 3 Linear Layers, whose dimensions are $4,096$, $4,096$, and 512 (more advanced layers like transformer layers could be applied but it is not the focus of this work). There is one ReLU layer and one Dropout layer between every two Linear layers. The Dropout layer probability is set to be 0.2. The SGD optimizer is employed with 0.9 momentum and 0.0005 weight decay. The training batch size N is 8. The hyper-parameters are set as $\lambda_1 = 0.5, \lambda_2 = 0.5$ and $\lambda_3 = 0.2$. We implement our method in PyTorch framework [38] with a single NVIDIA V100 GPU. The total training epoch is 50. All the methods are trained with an initial learning rate of 0.1 and it will be decreased by a factor of 0.1 at the 40-th epoch. **FELGA** is an end-to-end framework but we implement the VLP features extraction part and fragment embedding learning part separately due to the hardware constraints.

4.6. Quantitative Results and Analysis

The results of three VLP models CLIP, BLIP, and BLIPv2 on Microsoft COCO 2014 and 2017 datasets are shown in Tables 1, 2 and 3. Our approach **FELGA** achieves the best performance compared with other baselines on both Object Discovery and Description Assignment tasks. **FELGA** also performed better than the original VLP models when the features are based on the whole image or complete sentence.

From the tables, we find that "Dynamic" is better than

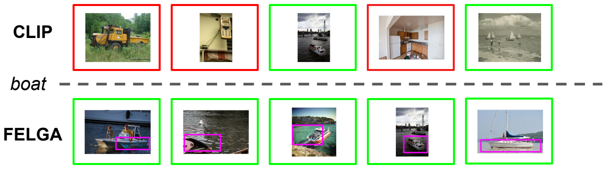

Figure 2. Illustrating results from CLIP and **FELGA** on Microsoft COCO 2017 dataset with the semantic concept query “boat” on the query-driven object discovery task.

“Dense” or “Sparse” in most cases, which means the dynamic pseudo label design is superior in discriminating the relationships between visual entities and semantic concepts. But its performance is still restricted by the original VLP model (BLIPv2 provides more powerful unified features compared with CLIP and BLIP). These also show the limitations of fragment-level only alignment. For all the experiments, the proposed **FELGA** outperforms other baselines, which shows that global-level alignment is important for fragment embedding learning.

Comparing the performance of VLP models and **FELGA**, we also found the performance gain is more in object discovery than in description assignment. The reason is that object discovery has a limited number of queries (80) and each query has a lot of images containing the corresponding objects. But description assignment has a larger number of bounding boxes as the query.

4.7. Ablation Study

To analyze the impacts of the designed local-level alignment and global-level alignment components, we report ablation experiments in Table 4. For the local-level alignment training, we select “Dynamic Connection” (Eqn. 6) to construct the fragment pseudo labels. It is easy to observe that either training the **FENet** by local-level alignment or global-level alignment individually could not achieve the same satisfactory results as in **FELGA**, which considers alignments of different levels during learning.

To show the effectiveness of the global-level alignment, we do experiments on three local-level only baselines with the added global-level alignment ($\mathcal{L}_{global}$) during training. From Table 5, we observe that the global-level alignment is helpful for other fragment-level methods on the object discovery task. But performance on the description assignment task is limited by the design of the fragment pseudo labels, it only has obviously better results when using “Dynamic Connection” local level alignment during training.

4.8. Qualitative Visualization

In Figure 2, we show the illustration of the top-5 retrieved images in the query-driven object discovery task on Microsoft COCO 2017 dataset. The semantic concept query is “boat”. (The text query is a straightforward design and more sophisticated prompt design works [20, 42, 44, 64, 65] may further improve the retrieval results, but it is not the focus part of this work). The green bounding box stands for the correct image that contains the corresponding visual entities, while the red bounding box stands for the incorrect images. The upper part shows the results of the original CLIP model and the lower part shows the results of **FELGA**. The maximum similarity visual proposals are highlighted in purple rectangles (it is worth noting these regions are not ground truth bounding boxes). These proposals are generated from the unsupervised region proposal generator. From the image, we found that **FELGA** can better capture the information of semantic concepts by paying more attention to the related visual entity proposals, which results in better query-driven object discovery performance.

5. Limitations

FELGA requires the unsupervised pre-trained image region proposal generator to provide the visual entity proposals and the keyword extractor to provide the semantic concept proposals. Since our method relies on these pre-trained models, the proposal quality is dependent on their performance, which inevitably affects the learned fragment embeddings. Also, **FELGA**-designed pseudo labels may not always reflect the true association between visual entities and semantic concepts. For example, different images in the same training batch may share the same semantic concepts. In this case, their local pseudo labels should also be set to 1. Further work could explore novel approaches for generating pseudo-labels that better capture the true associations, which can contribute more to the refinement and improvement of **FELGA**.

6. Conclusion

In this paper, we proposed a novel unsupervised method named **FELGA** for fine-grained cross-modal embeddings. To facilitate the evaluation of the learned embeddings, we define the query-driven object discovery task and the description assignment task and report extensive results from comparative experiments. Our results and their analysis suggest that FELGA outperforms existing VLP models used as references in this study, hence establishing a new baseline for the two tasks to facilitate future research. It is also important to note that **FELGA** is not limited to just these two tasks and can be extended to other applications that rely on inference using fine-grained cross-modal correlation.

Acknowledgments

This work was supported in part by National Science Foundation (Grant 2126291). Any opinions expressed in this material are those of the authors only.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2

[2] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14605–14615, 2022. 6

[3] Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. Measuring progress in fine-grained vision-and-language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582, Toronto, Canada, July 2023. Association for Computational Linguistics. 2

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 1

[5] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015. 2

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009. 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[8] Khoa D Doan, Peng Yang, and Ping Li. One loss for quantization: Deep hashing with discrete wasserstein distributional matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9447–9457, 2022. 2

[9] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022. 2

[10] Vijetha Gattupalli, Yaoxin Zhuo, and Baoxin Li. Weakly supervised deep image hashing through tag embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10375–10384, 2019. 2

[11] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 2

[12] Yuying Ge, Yixiao Ge, Xihui Liu, Jinpeng Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. Miles: Visual bert pre-training with injected language semantics for video-text retrieval. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022. 2

[13] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. 6

[14] 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. 1

[15] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023. 2

[16] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482, 2023. 2

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 6

[18] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014. 2

[19] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6

[20] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 8

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2

[22] 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. 2

[23] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 2

[24] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang

Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 1

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 5, 6

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 2, 5, 6

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2, 5

[28] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 2, 5

[30] Tieying Li, Xiaochun Yang, Bin Wang, Chong Xi, Hanzhong Zheng, and Xiangmin Zhou. Bi-cmr: Bidirectional reinforcement guided hashing for effective cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10275–10282, 2022. 2

[31] 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. 6

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[33] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 6

[34] Austin Stone Maxim Neumann Dirk Weissenborn Alexey Dosovitskiy Aravindh Mahendran Anurag Arnab Mostafa Dehghani Zhuoran Shen Xiao Wang Xiaohua Zhai Thomas Kipf Neil Houlsby Matthias Minderer, Alexey Gritsenko. Simple open-vocabulary object detection with vision transformers. *ECCV*, 2022. 2

[35] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 2

[36] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer, 2022. 2

[37] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 5

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[39] 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, 4, 5, 6

[40] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[41] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2

[42] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 8

[43] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multimodal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. 2

[44] Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, and Lu Jiang. Visual prompt tuning for generative transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19840–19851, 2023. 8

[45] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3027–3035, 2019. 2

[46] Tinne Tuytelaars, Christoph H Lampert, Matthew B Blaschko, and Wray Buntine. Unsupervised object discov-

ery: A comparison. *International journal of computer vision*, 88:284–302, 2010. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[48] Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. *Advances in Neural Information Processing Systems*, 34:16764–16778, 2021. 2

[49] Haoran Wang, Dongliang He, Wenhao Wu, Boyang Xia, Min Yang, Fu Li, Yunlong Yu, Zhong Ji, Errui Ding, and Jingdong Wang. Coder: Coupled diversity-sensitive momentum contrastive learning for image-text retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 700–716. Springer, 2022. 2

[50] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 2

[51] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 2

[52] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2088–2096, 2019. 2

[53] Yuanhao Xiong, Long Zhao, Boqing Gong, Ming-Hsuan Yang, Florian Schroff, Ting Liu, Cho-Jui Hsieh, and Liangzhe Yuan. Spatiotemporally discriminative video-language pre-training with text grounding. *arXiv preprint arXiv:2303.16341*, 2023. 2

[54] Hong Xuan and Xi Stephen Chen. Dissecting deep metric learning losses for image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2164–2173, 2023. 2

[55] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. 1

[56] 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. 6

[57] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2, 5

[58] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4626–4634, 2021. 2

[59] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 6

[60] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[61] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR, 17–23 Jul 2022. 1

[62] 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. 1, 2

[63] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, June 2022. 1

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 8

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 8

[66] Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16485–16494, June 2022. 1, 2

[67] Yaoxin Zhuo, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Baoxin Li. Clip4hashing: Unsupervised deep hashing for cross-modal video-text retrieval. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 158–166, 2022. 2

[68] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 2