Fine-Grained Semantically Aligned Vision-Language Pre-Training

Juncheng Li 1,2† Xin He 2† Longqiang Li 2† Long Qian 1 Linchao Zhu 1
Lingxi Xie 2 Yueting Zhuang 1† Qi Tian 2† Siliang Tang 1‡

1 Zhejiang University, 2 Huawei Cloud
{junchengli, qianlong0926, zhulinchao, yzhuang, siliang}@zju.edu.cn
{hexin80, weilonghui1, tian.qi1}@huawei.com, 198808xc@gmail.com

Abstract

Large-scale vision-language pre-training has shown impressive advances in a wide range of downstream tasks. Existing methods mainly model the cross-modal alignment by the similarity of the global representations of images and texts, or advanced cross-modal attention upon image and text features. However, they fail to explicitly learn the fine-grained semantic alignment between visual regions and textual phrases, as only global image-text alignment information is available. In this paper, we introduce LOUPE, a fine-grained semantically aligned vision-language pre-training framework, which learns fine-grained semantic alignment from the novel perspective of game-theoretic interactions. To efficiently compute the game-theoretic interactions, we further propose an uncertainty-aware neural Shapley interaction learning module. Experiments show that LOUPE achieves state-of-the-art performance on a variety of vision-language tasks. Furthermore, without any object-level human annotations and fine-tuning, LOUPE achieves competitive performance on object detection and visual grounding. More importantly, LOUPE opens a new promising direction of learning fine-grained semantics from large-scale raw image-text pairs. The repository of this work is at https://github.com/YYJMJC/LOUPE.

1 Introduction

Learning transferable cross-modal representations from large-scale vision-language pre-training has exhibited remarkable performance on a wide variety of downstream tasks. Most existing works can be classified into two categories: dual-encoder and fusion-encoder. The dual-encoder methods [17, 26, 33, 49] adopt two separate encoders to embed images and texts, and model the cross-modal alignment by the cosine similarity between the global features of images and texts. While such architecture is efficient for large-scale image-text retrieval by pre-computing image and text representations offline, they fail to model fine-grained semantic alignment between visual regions and textual phrases. On the other hand, the fusion-encoder methods [7, 19, 20, 25, 30, 34, 43, 42] attempt to use a single multi-modal encoder to jointly model the concatenated sequence of images and texts. These methods simulate soft alignment via advanced cross-modal attention [45]. However, they can only learn implicit alignment by end-to-end training, lacking explicit supervision to encourage semantic alignment.

†Work done when interning at Huawei Cloud.
††Equal Contribution.
‡‡Corresponding Authors.

alignment between visual regions and textual phrases. And the learned cross-modal attention matrices are often scattering and uninterpretable. Further, they are inefficient for retrieval since it requires jointly encoding every image-text pair during inference.

Learning fine-grained semantic alignment from image-text pre-training is crucial to many cross-modal reasoning tasks (e.g., visual grounding [51], image captioning [48]), but it is particularly challenging as the alignment information between visual regions and textual phrases is not available, posing fine-grained semantic alignment learning a weakly-supervised learning problem. In this paper, we address this problem while simultaneously maintaining high retrieval efficiency by proposing LOUPE, a fine-grained semantically aligned visio-lang pre-training framework, from the novel perspective of game theory. We formulate input patch and word tokens as multiple players into a cooperative game and quantify game-theoretic interactions (i.e., Shapley Interaction [12, 40]) among them to investigate the semantic alignment information. LOUPE learns fine-grained semantic alignment from two stages: token-level Shapley interaction modeling and semantics-level Shapley interaction modeling, where we first learn to identify semantic regions of images that correspond to some semantically meaningful entities, and then align these regions with phrases in the paired text.

Specifically, token-level Shapley interaction modeling aims to group patch tokens of images into semantic regions that semantically correspond to some visual instances. From the game-theoretic view, we take patch tokens as players and the similarity score between images and texts as the game function. Intuitively, supposing a set of patch tokens correspond to a visual instance in the image, then they tend to have strong interaction to form the complete semantics of the corresponding instance, which contributes to the better similarity judgment with the paired text. Based on this insight, we take the token-level Shapley interaction as soft supervision labels to encourage the model to capture semantic regions from images. Then, semantics-level Shapley interaction modeling infers the fine-grained semantic alignment between semantic regions and phrases. We consider every region and phrase as players and define a fine-grained similarity score as the game function. If a region and a phrase have strong correspondence, they tend to interact with each other and contribute to the fine-grained similarity score. By measuring the Shapley interaction between each region-phrase pair, we obtain the alignment information to guide the pre-training model.

As computing the exact Shapley interaction is an NP-hard problem [32], existing methods mainly employ sampling-based method [6] to obtain unbiased estimation. However, as the number of players grows, they require thousands of model evaluations. To reduce the computational cost, we further propose an efficient hybrid Shapley interaction learning strategy, where an uncertainty-aware neural Shapley interaction learning module cooperates with the sampling-based method. Experimental results show that our hybrid strategy significantly saves the computational cost while maintaining the estimation accuracy. More analysis is shown in Section 4.5.

Our framework serves as a proxy training objective that explicitly establishes the fine-grained semantic alignment between local region and phrase representations. This proxy objective can be directly removed for downstream tasks, rendering an efficient and semantics-sensitive dual-encoder model. Experiments show that LOUPE achieves new state-of-the-art on image-text retrieval benchmarks. For text-to-image retrieval on MSCOCO, LOUPE surpasses its strongest competitor by 4.2% on recall@1. Further, without any fine-tuning, LOUPE successfully transfers to object detection and visual grounding tasks in a zero-shot manner. For object detection, it achieves 12.1% mAP on COCO and 19.5% mAP on PASCAL VOC. For visual grounding, it achieves 26.8% accuracy on RefCOCO and 23.6% accuracy on RefCOCO+. Our contributions are summarized as follows:

- We propose LOUPE that explicitly learns fine-grained semantic alignment between visual regions and textual phrases while preserving the high retrieval efficiency of dual-encoder.
- We introduce an efficient and effective hybrid Shapley interaction learning strategy, based on an uncertainty-aware neural Shapley interaction learning module and a sampling-based method.
- Pre-trained on image-text data, LOUPE achieves new state-of-the-art on image-text retrieval and successfully transfers to the tasks that require more fine-grained object-level visual understanding (i.e., object detection and visual grounding) without any fine-tuning.
- As manual annotations for masses of object categories is time-consuming and unscalable, our work demonstrates a promising alternative, that is, learning fine-grained semantics from raw texts about images, which are easily available and contain a broader set of visual concepts.
2 Related Work

Vision-Language Pre-Training. The great success of pre-train-and-fine-tune paradigm in natural language processing [5, 9] and computer vision [10, 14, 47] has been expanded to the joint domain of vision and language [2, 3, 22]. Dominant vision-language pre-training models can be categorized into two groups: dual-encoder and fusion-encoder. The dual-encoder methods [17, 26, 35, 49] adopt two individual encoders to embed images and texts separately, and model the cross-modal interaction by cosine similarity. Such architecture is efficient for large-scale image-text retrieval as image and text representations can be pre-computed offline. However, simply measuring the cosine similarity between global representations is shallow to capture fine-grained semantic relationships between regions and phrases. The fusion-encoder methods [7, 15, 16, 20, 23, 25, 30, 32, 42, 43, 50, 55] adopt a single multi-modal encoder to jointly model the concatenated sequence of images and texts, which achieves deeper cross-modality interaction. However, these methods are less efficient as images and texts are intertwined to compute the cross-modal attention and can not be pre-computed offline. Further, there are no explicit supervision signals to encourage the alignment between regions and phrases. Some works [7, 24, 25, 30, 43, 50, 55, 56] attempt to leverage an off-the-shelf object detector to extract object features for pre-training. However, the detector is usually pre-trained on limited object categories. Furthermore, considering the excessive demand on memory and computation, existing methods usually fix the parameters of detection models and regard region detection as a pre-processing step, disconnected with vision-language pre-training. Thus, the performance is also restricted by the quality of detection models. FILIP [49] uses a token-wise maximum similarity to enhance the cross-modal interaction of dual-encoder methods. To learn explicit fine-grained semantic alignment, GLIP [23] and X-VLM [53] utilize human-annotated datasets, where regions with bounding-box annotations are aligned with text descriptions. Such a manner is time-consuming and hard to scale to larger raw image-text data from the Internet. In contrast, our proposed framework explicitly learns the fine-grained semantic alignment from raw image-text data and at the same time maintains the high efficiency of dual-encoder. Detailed discussions can be found in Appendix K.

Shapley Values. The Shapley value [40] was initially introduced in game theory. It has been theoretically proven to be the unique metric to fairly estimate the contribution of each player in a cooperative game such that certain desirable axioms are satisfied [46]. With solid theoretical foundations, Shapley value has recently been studied as post-hoc explanation methods for Deep Neural Networks (DNN) [8, 31, 54]. Lundberg et al. [31] propose a unified attribution method based on Shapley value to interpret the predictions of DNN. Ren et al. [48] propose to explain adversarial attacks by Shapley value. In this paper, we propose to model fine-grained semantic alignment by game-theoretic interactions, along with an efficient Shapley interaction learning strategy.

3 Method

In this section, we first introduce the problem formulation of fine-grained semantically aligned vision-language pre-training in Section 3.1. Then, we propose the corresponding LOUPE framework for fine-grained semantic alignment learning in Section 3.2 and an efficient approach for Shapley interaction learning in Section 3.3.

3.1 Problem Formulation and Model Overview

Generally, vision-language pre-training aims to learn an image encoder \( f_I \) and a text encoder \( f_T \) by cross-modal contrastive learning, where the matched image-text pairs are optimized to get closer and the mismatched pairs are optimized to get further. Let \( f_I(I_i) \) and \( f_T(T_i) \) denote the global representations of the image and text. Then the cross-modal contrastive loss can be formulated as:

\[
\mathcal{L}_{CMC} = - \log \frac{\exp(f_I(I_i)^\top f_T(T_i)/\tau)}{\sum_B \exp(f_I(I_i)^\top f_T(T_j)/\tau)} - \log \frac{\exp(f_I(I_j)^\top f_T(T_i)/\tau)}{\sum_B \exp(f_I(I_j)^\top f_T(T_i)/\tau)} \quad (1)
\]

where \( B \) is the batch size and \( \tau \) is the temperature hyper-parameter.

While intuitive, such a manner can only learn coarse alignment between images and texts but fails to explicitly capture the fine-grained semantic alignment between visual regions and textual phrases. To learn fine-grained semantic alignment while simultaneously maintaining high retrieval efficiency, we
propose LOUPE, a fine-grained semantically aligned vision-language pre-training framework that germinates from cooperative game theory.

As illustrated in Figure 1, LOUPE learns fine-grained semantic alignment from two stages: **token-level Shapley interaction modeling** and **semantics-level Shapley interaction modeling**. For token-level Shapley interaction modeling, we learn to aggregate patch tokens of images into semantic regions that semantically correspond to some visual concepts, under the guidance of token-based semantic aggregation loss $L_{TSA}$. As for semantics-level Shapley interaction modeling, the semantic alignment between the aggregated regions and textual phrases is learned, supervised by the fine-grained semantic alignment loss $L_{FSA}$. Combined with the two newly proposed losses, the full objective of fine-grained semantically aligned vision-language pre-training can be formulated as:

$$L = L_{CMC} + L_{TSA} + L_{FSA}$$  \hspace{1cm} (2)

Such a new pre-training objective enforces the image encoder to capture semantic regions and establishes fine-grained semantic alignment between visual regions and textual phrases. During inference, it can be directly removed, rendering an efficient and semantics-sensitive dual-encoder.

### 3.2 Interpreting Fine-Grained Semantic Alignment as Game-Theoretic Interaction

#### 3.2.1 Preliminaries

**Shapley Values.** The Shapley value [40] is a classic game theory solution for the unbiased estimation of the importance or contribution of each player in a cooperative game. Considering a game with $N = \{1, ..., n\}$ players, $S \subset N$ denotes a potential subset of players. A game $v(\cdot)$ is implemented as a function that maps each subset $S$ of players to a score, modeling the outcome of a game when players in $S$ participate in. Specifically, $v(N) - v(\emptyset)$ denotes the contribution obtained by all players in the game. The Shapley value $\phi(i|N)$ for player $i$ is defined as the average marginal contribution of player $i$ to all possible coalitions $S$ that are formed without $i$:

$$\phi(i|N) = \sum_{S \subseteq N \setminus \{i\}} p(S)[v(S \cup \{i\}) - v(S)], \quad p(S) = \frac{|S|!(|N| - |S| - 1)!}{|N|!}$$  \hspace{1cm} (3)

where $p(S)$ is the likelihood of $S$ being sampled. The Shapley value has been proved to be the unique metric that satisfies the following axioms: **Linearity**, **Symmetry**, **Dummy**, and **Efficiency** [46]. We summarize these axioms in Appendix B.

**Shapley Interaction.** In the game theory, some players tend to form a coalition and always participate in the game together. The players in the coalition might interact or cooperate with each other, which...
We also prepend a learnable special token to the input to form a visual instance, which contributes to the better similarity judgment with the words in the text. From the game-theoretic view, we take patch tokens and word tokens as players in the game. A textual phrase usually refers to a specific semantic region in an image, which is composed of multiple patches and represents a visual instance. Thus, we first introduce token-level Shapley interaction modeling to aggregate patches into semantic regions.

### Understanding Semantic Region via Shapley Interaction

Supposing a set of patches represent a visual instance in an image, then they tend to have a strong Shapley interaction because they work jointly to form a visual instance, which contributes to the better similarity judgment with the text. From the game-theoretic view, we take patch tokens and word tokens as players in the game, and the global similarity between images and texts as the game score. To compute the similarity between words and pixels (patches), we keep tokens in and mask input tokens in to zero. Thus, the global similarity only considers the tokens in and the global similarity only considers the tokens in to zero.

#### Semantic Region Generation

Inspired by YOLOv3 [57], we design a lightweight region generation module. It takes each patch token representation as input and generates a bounding box prediction centered on , which corresponds to a visual region with patch tokens. The region generation module also predicts a confidence score for each region. We select the top-M predictions as the semantic regions. Then, the Shapley interaction of can be defined as:

\[
\mathcal{I}([R_i]) = \phi([R_i]|X \setminus R_i \cup \{[R_i]\}) - \sum_{x_{i,k}^I \in R_i} \phi(x_{i,k}^I|X \setminus R_i \cup \{x_{i,k}^I\})
\]

According to the Equation, we can reformulate the Shapley value into the form of expectation:

\[
\phi([R_i]|X \setminus R_i \cup \{[R_i]\}) = \mathbb{E}_c \left\{ \mathbb{E}_{S \subseteq X \setminus R_i \cup \{R_i\} \mid |S| = c} \left[ v_1(S \cup R_i) - v_1(S) \right] \right\}
\]

where is the coalition size. The Shapley interaction of can be reformulated as (we provide the proof in Appendix C):

\[
\mathcal{I}([R_i]) = \mathbb{E}_c \left\{ \mathbb{E}_{S \subseteq X \setminus R_i \cup \{R_i\} \mid |S| = c} \left[ v_1(S \cup R_i) - \sum_{x_{i,k}^I \in R_i} v_1(S \cup \{x_{i,k}^I\}) + (K - 1)v_1(S) \right] \right\}
\]
Taking normalized $\mathcal{Y}^i([R_i])$ as the soft supervision label, the token-based semantic aggregation loss is defined as cross-entropy loss:

$$L_{TSA} = -\frac{1}{M} \sum_{i=1}^{M} [\mathcal{Y}^i([R_i])] \log(s(R_i)) + (1 - \mathcal{Y}^i([R_i])) \log(1 - s(R_i))]$$  \hspace{1cm} (8)$$
which propagates gradients to the region generation module and image encoder to adjust bounding box predictions such that more accurate semantic regions can be captured.

### 3.2.3 Semantics-Level Shapley Interaction Modeling

After obtaining the inferred semantic regions, we propose semantics-level Shapley interaction modeling to explicitly model the fine-grained semantic alignment between regions and phrases. We first define the fine-grained similarity score and then explain semantic alignment based on game theory.

We adopt Avg-Pooling over learned patch representations in each $R_i$ to obtain region representation $h^i_I \in \mathbb{R}^d$. We employ an off-the-shelf constituency parser to extract phrases from text and obtain phrase representation $h^i_T \in \mathbb{R}^d$ by Avg-Pooling. Totally, we obtain $M$ regions $H^I = \{h^i_I\}_{i=1}^M$ and $N$ phrases $H^T = \{h^j_T\}_{j=1}^N$. And the alignment matrix can be defined as: $A = \{a_{ij}\}_{M \times N}$, where $a_{ij} = h^i_I h^j_T$ represents the alignment score between $i$-th region and $j$-th phrase. Next, we apply softmax-normalization over each row of $A$, obtaining $\hat{A}$. For the $i$-th region, we calculate its maximum alignment score as $\max_j \hat{a}_{ij}$. Then, we use the average maximum alignment score over all regions as the fine-grained image-to-text similarity $p_1$. Similarly, we can obtain the fine-grained text-to-image similarity $p_2$, and the total fine-grained similarity score can be defined: $p = (p_1 + p_2)/2$.

**Understanding Semantic Alignment via Shapley Interaction.** If a region and a phrase have strong semantic correspondence, then they tend to cooperate with each other and contribute to the fine-grained similarity score. Thus, we can consider $\mathcal{H} = H^I \cup H^T$ as the players and the fine-grained similarity score $p$ as the game score $v_2(\cdot)$. The Shapley interaction of them can be formulated as:

$$\mathcal{Y}([H_{ij}]) = \phi([H_{ij}] \cup [\mathcal{H}_{ij} \cup \{H_{ij}\}]) = \phi(h^i_I | \mathcal{H} \setminus H_{ij} \cup \{h^i_I\}) = \phi(h^j_T | \mathcal{H} \setminus H_{ij} \cup \{h^j_T\})$$  \hspace{1cm} (9)$$

$$= \mathbb{E}_{c \subseteq N \setminus \mathcal{H}_{ij}} \left[ v_2(S \cup H_{ij}) - v_2(S \cup \{h^i_I\}) - v_2(S \cup \{h^j_T\}) + v_2(S) \right]$$  \hspace{1cm} (10)$$

where $|H_{ij}|$ represents the single player formed by the coalition of $i$-th region and $j$-th phrase. Taking normalized $\mathcal{Y}([H_{ij}])$ as soft labels, the fine-grained semantic alignment loss can be defined as:

$$L_{TSA} = -\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \mathcal{Y}([H_{ij}]) \log(\hat{a}_{ij})$$  \hspace{1cm} (11)$$

### 3.3 Uncertainty-Aware Neural Shapley Interaction Learning

According to Equation 3 and Equation 4, computing exact the Shapley value is an NP-hard problem. Previous methods mainly apply sampling-based method to approximate it. While sampling-based approximation is unbiased, an accurate approximation requires thousands of model evaluations. To reduce the computational cost, we propose an uncertainty-aware neural Shapley interaction learning (UNSil) module to cooperate with the sampling-based method, rendering an efficient and effective hybrid strategy.

Specifically, the sampling-based method estimates the expectation terms in Equation 7 and Equation 10 by sampling to compute the Shapley interaction. Inspired by noisy label learning, the UNSIL module learns to predict the Shapley interaction and the corresponding uncertainty $\sigma$. Intuitively, if the UNSIL module makes a prediction with low uncertainty, we can directly apply its prediction to $L_{TSA}$ and $L_{TSA}$, avoiding thousands of model evaluations. If the uncertainty is high, we then resort to the sampling-based method for a more accurate estimation.

During training, the UNSIL module first predicts the target interaction with uncertainty $\sigma$. Then, we sample a value $\epsilon$ from a uniform distribution on $(0, 1)$. If $\epsilon > \sigma$, we directly use its prediction.
Table 1: Results (%) of zero-shot image-text retrieval on Flickr30K and MSCOCO datasets.

<table>
<thead>
<tr>
<th></th>
<th>Flickr30K</th>
<th>MSCOCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>image-to-text</td>
<td>text-to-image</td>
</tr>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>ImageBERT</td>
<td>70.7</td>
<td>90.2</td>
</tr>
<tr>
<td>UNITER</td>
<td>83.6</td>
<td>95.7</td>
</tr>
<tr>
<td>CLIP</td>
<td>88.0</td>
<td>98.7</td>
</tr>
<tr>
<td>ALIGN</td>
<td>88.6</td>
<td>98.7</td>
</tr>
<tr>
<td>FILIP</td>
<td>89.8</td>
<td>99.2</td>
</tr>
<tr>
<td>LOUPE</td>
<td>90.5</td>
<td>99.5</td>
</tr>
</tbody>
</table>

Table 2: Top-1 accuracy (%) of zero-shot image classification over 11 datasets.

<table>
<thead>
<tr>
<th></th>
<th>CIFAR10</th>
<th>Food101</th>
<th>StanfordCars</th>
<th>SUN397</th>
<th>Flowers102</th>
<th>Country211</th>
<th>Aircrafts</th>
<th>OxfordPets</th>
<th>Caltech101</th>
<th>ImageNet</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>96.2</td>
<td>92.9</td>
<td>77.3</td>
<td>67.7</td>
<td>78.7</td>
<td>34.9</td>
<td>57.7</td>
<td>36.1</td>
<td>93.5</td>
<td>92.6</td>
<td>75.3</td>
</tr>
<tr>
<td>LOUPE</td>
<td>95.9</td>
<td>94.3</td>
<td>79.9</td>
<td>69.8</td>
<td>87.4</td>
<td>37.8</td>
<td>53.3</td>
<td>54.9</td>
<td>94.1</td>
<td>93.9</td>
<td>76.1</td>
</tr>
</tbody>
</table>

\( \epsilon \leq \sigma \), we then use the sampling-based method to compute the Shapley Interaction and update the UNSIL module based on the sampling-based results. Note that, for the first few iterations, we employ the sampling-based method directly, and use its results to train the UNSIL module.

Let \( \hat{I} \) and \( \hat{I}^* \) denote the results from the sampling-based method and UNSIL module, respectively. Taking \( \hat{I}^* \) as the ground-truth, the UNSIL module is trained by:

\[
L_{UNSI} = \frac{1}{\beta_1 \sigma} L_{MSE}(\hat{I}, \hat{I}^*) + \beta_2 \sigma
\]  

where the first term is mean squared error \( L_{MSE} \) weighted by the uncertainty, the second term serves as a regularization term for the prediction uncertainty, and \( \beta \) is the weight hyper-parameter. The UNSIL module implicitly learns the uncertainty from the regression loss function. We discuss the implementation details of the UNSIL module in Section 4.5 and Appendix D.

4 Experiments

4.1 Pre-training Details

As sufficient data is a prerequisite for vision-language pre-training, we construct a dataset with 240M image-text pairs from the Internet. We implement the image encoder by Swin-L [28] and the text encoder by BERT-Small [9]. The input images are resized to \( 224 \times 224 \) and the input texts are tokenized by WordPiece with a maximum length of 60. We pre-train the model for 20 epochs using a batch size of 512 on 128 NVIDIA V100 GPUs. We utilize AdamW [29] optimizer with a learning rate of \( 2 \times 10^{-4} \) and a weight decay of 0.01. More pre-training and evaluation details are provided in Appendix D, E. We also analyze the image encoder and training efficiency in Appendix G, J.

4.2 Zero-Shot Image-Text Retrieval

We compare LOUPE on the widely used MSCOCO [27] and Flickr30K [33] datasets. First, the results in Table 1 show that LOUPE achieves new state-of-the-art zero-shot performance on most metrics of the two datasets, demonstrating the stronger generalizability of our pre-training framework. Second, while previous works mainly pre-train on larger datasets (CLIP 400M, ALIGN 1800M, FILIP 340M), LOUPE still achieves superior performance using less training data (240M). Third, compared with FILIP which directly computes token-wise similarity, our model captures semantic alignment between visual regions and textual phrases, which is more semantically meaningful. For text-to-image retrieval on MSCOCO, LOUPE significantly outperforms FILIP by 4.2% on recall@1.

4.3 Zero-Shot Image Classification

In this section, we evaluate LOUPE on the zero-shot image classification task. We compare LOUPE with CLIP on 11 downstream classification datasets, following the same evaluation setting as
Table 3: Without fine-tuning, zero-shot transfer performance on object detection and visual grounding.

<table>
<thead>
<tr>
<th></th>
<th>COCO</th>
<th>PASCAL VOC</th>
<th>ReCOCO</th>
<th>ReCOCO+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP@0.3</td>
<td>mAP@0.5</td>
<td>val</td>
<td>testA</td>
</tr>
<tr>
<td>CLIP + Pixel-Wise</td>
<td>8.5</td>
<td>4.5</td>
<td>18.2</td>
<td>7.3</td>
</tr>
<tr>
<td>CLIP + K-Means</td>
<td>6.4</td>
<td>1.9</td>
<td>11.7</td>
<td>4.8</td>
</tr>
<tr>
<td>CLIP + Grad-CAM</td>
<td>7.1</td>
<td>3.2</td>
<td>19.1</td>
<td>8.2</td>
</tr>
<tr>
<td>AdaptCLIP</td>
<td>14.9</td>
<td>6.6</td>
<td>28.7</td>
<td>12.9</td>
</tr>
<tr>
<td>LOUPE</td>
<td>25.3</td>
<td>12.1</td>
<td>30.3</td>
<td>19.5</td>
</tr>
<tr>
<td></td>
<td>mAP@0.5</td>
<td>val</td>
<td>testA</td>
<td>testB</td>
</tr>
<tr>
<td>CLIP + Pixel-Wise</td>
<td>6.7</td>
<td>6.2</td>
<td>5.8</td>
<td>5.7</td>
</tr>
<tr>
<td>CLIP + K-Means</td>
<td>2.1</td>
<td>2.3</td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td>CLIP + Grad-CAM</td>
<td>5.5</td>
<td>5.2</td>
<td>4.8</td>
<td>4.4</td>
</tr>
<tr>
<td>AdaptCLIP</td>
<td>16.7</td>
<td>18.4</td>
<td>18.0</td>
<td>17.5</td>
</tr>
<tr>
<td>LOUPE</td>
<td>25.2</td>
<td>26.8</td>
<td>24.5</td>
<td>22.9</td>
</tr>
</tbody>
</table>

Figure 2: An example of LOUPE zero-shot transferring to object detection using prompt templates.

CLIP [35]. Table 2 summarizes the results. As shown in Table 2, our LOUPE outperforms CLIP with average improvement of 3.1%. Notably, on ImageNet, the largest dataset among 11 datasets, our LOUPE surpasses CLIP by 0.8%. Also, we observe that LOUPE achieves substantial performance gains on several fine-grained image classification datasets (i.e., Flowers102 and Aircrafts). It demonstrates the superiority of our LOUPE on fine-grained semantics understanding.

We also evaluate the linear probing performance of our LOUPE on image classification. The detailed results can be found in Appendix I.

4.4 Zero-Shot Transfer to Object Detection and Visual Grounding

To answer whether our model has learned fine-grained semantics, we further evaluate LOUPE on object detection [39] and visual grounding [51], which require more fine-grained semantic understanding ability to identify specific visual regions in images according to the object labels or referring expressions. Visual grounding can be seen as generalized object detection, where the pre-defined class labels are replaced by language referring expression sentences. As LOUPE can generate a set of semantic regions that are aligned with textual phrases, it can be easily applied to object detection and visual grounding without structure modification. For visual grounding, we take referring expressions as input text. For object detection, as illustrated in Figure 2, we use prompt to expand detection labels to input text. Then, we encode input text by the learned text encoder, and these tasks can be completed by measuring the similarity between candidate region representations and text representations.

For comparison, we zero-shot transfer CLIP (ViT-L/14) to object detection and visual grounding by applying several non-parametric approaches on the spatial feature maps of CLIP. We also compare with AdaptCLIP [21], which is a concurrently unpublished method that leverages classic superpixel (SLIC [1]) and bounding box proposal (selective search [44]) methods to zero-shot transfer CLIP to phrase localization. We use its public official implementations to get the experiment results. For object detection, we evaluate their mean Average Precision (mAP) at IoU thresholds of {0.3, 0.5} on COCO [27] (65 classes) and PASCAL VOC [11] (20 classes). For visual grounding, we evaluate their top-1 accuracy at IoU thresholds of 0.5 on ReCOCO [51] and ReCOCO+ [51]. The experiment details of CLIP variants and LOUPE are provided in Appendix E.

Table 3 summarizes the results. 1) Overall, LOUPE outperforms all CLIP variants by a large margin. The significantly higher performance illustrates the stronger zero-shot transfer ability of our fine-grained semantically aligned pre-training paradigm. 2) Second, all CLIP variants rely on pre-processing steps on CLIP’s feature map (e.g., AdaptCLIP first uses SLIC to group pixels and then uses selective search to generate a large number of proposals), which is time-consuming. In contrast, our method directly predicts the semantic regions based on the patch token representations. 3) Third, the consistently competitive performance across four benchmarks validates that LOUPE can learn fine-grained semantics from raw text supervision. LOUPE demonstrates a promising alternative, that is, learning fine-grained semantics from large-scale raw image-text pairs, which are easily available and contain a broader set of visual concepts.
### Table 4: Ablation study of each component across three tasks.

<table>
<thead>
<tr>
<th></th>
<th>MSCOCO</th>
<th>COCO</th>
<th>ReCO</th>
<th>Training Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I2T</td>
<td>T2I</td>
<td>mAP@0.3</td>
<td>mAP@0.5</td>
</tr>
<tr>
<td>Backbone</td>
<td>31.0</td>
<td>24.8</td>
<td>3.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Backbone + $L_{\text{TSA}}$</td>
<td>32.4</td>
<td>26.2</td>
<td>7.6</td>
<td>3.3</td>
</tr>
<tr>
<td>Backbone + $L_{\text{TSA}}$ + $L_{\text{FSA}}$</td>
<td>33.5</td>
<td>28.3</td>
<td>9.4</td>
<td>5.9</td>
</tr>
<tr>
<td>Backbone + $L_{\text{TSA}}$ + $L_{\text{FSA}}$ + UNSIL</td>
<td>33.3</td>
<td>28.1</td>
<td>9.0</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Figure 3: (a) Instability of the Shapley interaction approximation with respect to different sampling numbers. (b, c) Uncertainty and error of the UNSIL module with different structures.

As time-consuming human annotations are unscalable for massive object classes in the real world, some recent works [4, 36] target at training object detectors with annotations on base object classes to generalize to the remaining object classes of the same dataset. The latest works [13, 52] leverage the generalizability of vision-language pre-training models to further improve the zero-shot performance on novel classes. However, these zero-shot approaches still require bounding box annotations on base classes for task-specific supervised learning. In contrast, our LOUPE is trained on large-scale raw image-text pairs, which are already accessible on the Internet and contain more diverse semantics.

### 4.5 Ablation Study

#### Effectiveness of Individual Components.

In this section, we investigate the effectiveness of each component in Table 4. Given the costly training time, all ablation studies are based on a relatively small dataset (Conceptual Captions 3M [41]). We start with the backbone model that consists of a dual-encoder trained by cross-modal contrastive loss. We then gradually add token-level Shapley interaction modeling supervision $L_{\text{TSA}}$ (Row 2), semantics-level Shapley interaction modeling supervision $L_{\text{FSA}}$ (Row 3), and UNSIL module (Row 4). For Row 2 and Row 3, the Shapley interaction is only computed by the sampling-based method. The results in Table 4 show that both $L_{\text{TSA}}$ and $L_{\text{FSA}}$ bring significant improvement for all tasks. We observe that $L_{\text{TSA}}$ boosts a 3.8% improvement on object detection. And the improved fine-grained visual semantic understanding further facilitates the cross-modal retrieval performance (+1.4%). The semantics-level Shapley interaction modeling further improves the performance on all tasks by modeling the semantic alignment between visual regions and textual phrases. Comparing Row 3 and Row 4, we observe that the UNSIL module maintains the estimation accuracy while avoiding intensive computations. The averaged training time is reduced from 9.90 seconds per iteration to 1.93 seconds per iteration.

#### Accuracy of the Shapley Interaction Learning.

Since we use the sampling-based method [6] to compute the Shapley Interaction and train the UNSIL module, we conduct a study to evaluate the accuracy of the sampling-based method and the error of the UNSIL module. As [54], we compute the interaction multiple times and measure the instability of them. A lower instability means that we obtain similar interactions from different sampling processes. It indicates a high accuracy. Specifically, the instability is defined as $\frac{\mathbb{E}_{u \neq w} [I_u - I_w]}{\mathbb{E}_w [I_w]}$, where $I_w$ denotes the interaction computed in the $w$-th time. We average the instability values over Shapley interaction of 100 image-text pairs. We report the average instability values with respect to different sampling numbers. As shown in Figure 3 (a), the instability decreases along with the increase of the sampling number. When the sampling number is larger than 500, the approximated Shapley interaction is stable enough with instability less than 0.06. Further, we attempt different models (i.e., Conv1D, 3-Layer MLP + Attention, 3-Layer Transformer) to implement the UNSIL module (please see Appendix D for more details). We test them on 1000 samples and report their mean uncertainty and relative error in Figure 3 (b) and (c). We observe that MLP + Attention is good enough to predict the interaction with lower complexity. Thus, we implement the UNSIL module by MLP + Attention.
4.6 Qualitative Analysis

**Qualitative Examples.** As shown in Figure 4, LOUPE successfully captures the regions that correspond to the detected objects, and grounds the referring expressions onto the referred regions.

**Visualization of Learned Fine-Grained Semantic Alignment.** In Figure 5, we visualize some key semantic regions and corresponding alignment matrices inferred by LOUPE. We present the regions with top-3 confidence (Region 1 – 3) and two randomly sampled regions (white boxes). The red boxes at the bottom of bounding boxes indicate their normalized token-level Shapley interaction values. Comparing their Shapley interaction values, we observe that the token-level Shapley interaction successfully distinguishes semantic regions from randomly sampled regions. The semantically meaningful regions tend to have stronger interaction. It indicates that token-level Shapley interaction can provide correct supervision for semantic region generation. Further, we show the alignment matrices inferred by semantics-level Shapley interaction and LOUPE, respectively. As shown in the right case of Figure 5, LOUPE successfully recognizes the leash region and aligns it with the “a leash” phrase. Note that existing object detection datasets do not contain the “leash” category.

5 Conclusion

This paper introduces a novel vision-language pre-training framework, **LOUPE**, which models the fine-grained semantic alignment between visual regions and textual phrases by game-theoretic interactions. To efficiently compute the interactions, we further propose an uncertainty-aware neural Shapley interaction learning module. Comprehensive experiments show that LOUPE achieves new state-of-the-art on image-text retrieval datasets and can transfer to object detection and visual grounding in a zero-shot manner. This work demonstrates a new promising direction of learning fine-grained semantics from large-scale raw image-text data.

**Limitations.** 1) The phrases are extracted by off-the-shelf constituency parsers, whose predictions might not be completely accurate. 2) The web data might inevitably contain mismatched image-text pairs, leading to noisy supervision.

**Social Impacts.** Our model is trained on noisy data from the Internet, which may contain unsuitable images, violent text, or private information. Thus, additional analysis of the data is necessary. Further, the use of our model for privacy surveillance or other nefarious purposes should be prohibited.

**Acknowledgment.** This work has been supported in part by the National Key Research and Development Program of China (2018AAA0101900), Zhejiang NSF (LR21F020004), Key Research and Development Program of Zhejiang Province, China (No. 2021C01013), Chinese Knowledge Center of Engineering Science and Technology (CKCEST). We thank all the reviewers for their valuable comments.
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Please see Appendix B, C.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The pre-trained LOUPE model is coming soon.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see Section 4.1 and Appendix D, E.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see Section 4.1 and Appendix D, E, J.

4. If you used existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] We do not use crowdsourcing or conducted research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] We do not use crowdsourcing or conducted research with human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] We do not use crowdsourcing or conducted research with human subjects.