Fine-Grained Semantically Aligned Vision-Language Pre-Training

Juncheng Li [1,2*†] **Xin He** [2†] **Longhui Wei** [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 a**L**igned visi**O**n-lang**U**age **P**r**E**-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 [18, 29, 38, 52] 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, 21, 23, 28, 33, 37, 46, 45] 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 [48]. However, they can only learn implicit alignment by end-to-end training, lacking explicit supervision to encourage semantic

*Work done when interning at Huawei Cloud.

†Equal Contribution.

‡Corresponding Authors.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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* [54], *image captioning* [51]), 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 a**L**igned visi**O**n-lang**U**age **P**r**E**-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, 43]) 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 [35], 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, 50] has been expanded to the joint domain of vision and language [2, 3, 25]. Dominant vision-language pre-training models can be categorized into two groups: *dual-encoder* and *fusion-encoder*. The dual-encoder methods [18, 29, 38, 52] 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, 16, 17, 21, 23, 27, 28, 33, 37, 45, 46, 53, 58] 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, 27, 28, 33, 46, 53, 58, 59] 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 [52] uses a token-wise maximum similarity to enhance the cross-modal interaction of dual-encoder methods. To learn explicit fine-grained semantic alignment, GLIP [26] and X-VLM [56] 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 [43] 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 [49]. With solid theoretic foundations, Shapley value has recently been studied as post-hoc explanation methods for Deep Neural Networks (DNN) [8, 34, 57]. Lundberg *et al.* [34] propose a unified attribution method based on Shapley value to interpret the predictions of DNN. Ren *et al.* [41] 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 $f_{\mathrm{I}}(I_i)$ and $f_{\mathrm{T}}(T_i)$ denote the global representations of the image and text. Then the cross-modal contrastive loss can be formulated as:

$$\mathcal{L}_{\mathrm{CMC}} = -\log \frac{\exp(f_{\mathrm{I}}(I_i)^\top f_{\mathrm{T}}(T_i)/\tau))}{\sum_j^B \exp(f_{\mathrm{I}}(I_i)^\top f_{\mathrm{T}}(T_j)/\tau)} - \log \frac{\exp(f_{\mathrm{I}}(I_i)^\top f_{\mathrm{T}}(T_i)/\tau))}{\sum_j^B \exp(f_{\mathrm{I}}(I_j)^\top f_{\mathrm{T}}(T_i)/\tau))} \tag{1}$$

where B is the batch size and τ 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

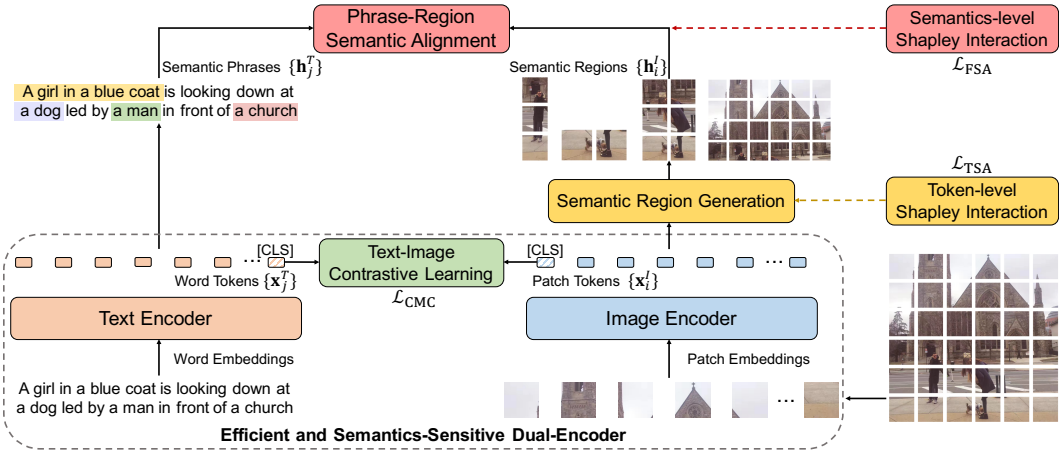


Figure 1: Overview of LOUPE. Our framework serves as a proxy training objective that encourages the image encoder to capture semantic regions and establishes the semantic alignment between region and phrase representations. The proxy training objective can be easily removed for downstream tasks, rendering an efficient and semantics-sensitive dual-encoder.

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 $\mathcal{L}_{\text{TSA}}$. 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 $\mathcal{L}_{\text{FSA}}$. Combined with the two newly proposed losses, the full objective of fine-grained semantically aligned vision-language pre-training can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{CMC}} + \mathcal{L}_{\text{TSA}} + \mathcal{L}_{\text{FSA}} \tag{2}$$

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 [43] 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 $\mathcal{N} = \{1, ..., n\}$ players, $\mathcal{S} \subseteq \mathcal{N}$ denotes a potential subset of players. A game $v(\cdot)$ is implemented as a function that maps each subset $\mathcal{S}$ of players to a score, modeling the outcome of a game when players in $\mathcal{S}$ participate in. Specifically, $v(\mathcal{N}) - v(\varnothing)$ denotes the contribution obtained by all players in the game. The Shapley value $\phi(i|\mathcal{N})$ for player i is defined as the average marginal contribution of player i to all possible coalitions $\mathcal{S}$ that are formed without i:

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

where $p(\mathcal{S})$ is the likelihood of $\mathcal{S}$ being sampled. The Shapley value has been proved to be the unique metric that satisfies the following axioms: *Linearity*, *Symmetry*, *Dummy*, and *Efficiency* [49]. 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

brings additional contributions to the game. The Shapley interaction [12] measures this additional contributions brought by the coalition compared with the case when the players work individually. For a coalition $\mathcal{S}$, we consider $[\mathcal{S}]$ as a single hypothetical player, which is the union of the players in $\mathcal{S}$. Then, the reduced game is formed by removing the individual players in $\mathcal{S}$ from the game and adding $[\mathcal{S}]$ to the game. The Shapley value $\phi([\mathcal{S}]|\mathcal{N} \setminus \mathcal{S} \cup \{[\mathcal{S}]\})$ for player $[\mathcal{S}]$ can be computed using Equation 3 over the reduced game. Similarly, we can obtain $\phi(i|\mathcal{N} \setminus \mathcal{S} \cup \{i\})$, where i is the individual player in S. Finally, the Shapley interaction for coalition $\mathcal{S}$ is formulated as:

$$\mathfrak{I}([\mathcal{S}]) = \phi([\mathcal{S}]|\mathcal{N} \setminus \mathcal{S} \cup \{[\mathcal{S}]\}) - \sum_{i \in \mathcal{S}} \phi(i|\mathcal{N} \setminus \mathcal{S} \cup \{i\}) \tag{4}$$

In this way, $\mathfrak{I}([\mathcal{S}])$ reflects the interactions inside $\mathcal{S}$. The higher value of $\mathfrak{I}([\mathcal{S}])$ indicates that players in $\mathcal{S}$ cooperate closely with each other.

3.2.2 Token-Level Shapley Interaction Modeling

Due to inherent semantic unit mismatch between texts and images, it is ineffective to directly compute the alignment between words and pixels (patches). A textual phrase usually refers to a specific image region, 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.

Input Representations. Given an image-text pair, the input image I is sliced into patches and flattened. Followed by linear projection layer and position embeddings, we obtain patch token sequence $\mathcal{X}^I = \{\mathbf{x}_i^I\}_{i=1}^{L_1}$ with an additional `[CLS_I]` token embedding. The input text T is tokenized and embedded into word token sequence $\mathcal{X}^T = \{\mathbf{x}_i^T\}_{i=1}^{L_2}$, added with position embeddings. We also prepend a learnable special token `[CLS_T]` to the word token sequence. Then, we adopt a dual-encoder structure to encode the patch token sequence and word token sequence, separately. On top of the image and text encoders, we obtain the representations of patch token sequence $\tilde{\mathcal{X}}^I = \{\tilde{\mathbf{x}}_i^I\}_{i=1}^{\tilde{L}_1}$ and word token sequence $\tilde{\mathcal{X}}^T = \{\tilde{\mathbf{x}}_i^T\}_{i=1}^{\tilde{L}_2}$. We take the learned representations of `[CLS_I]` and `[CLS_T]` tokens as the global representations for images and texts. And the global similarity of image-text pairs is measured by the cosine similarity between them.

Understanding Semantic Region via Shapley Interaction. Supposing a set of patches represent a complete 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 $\mathcal{X} = \mathcal{X}^I \cup \mathcal{X}^T$, and the global similarity between images and texts as the game score $v_1(\cdot)$. To compute $v_1(\mathcal{S})$, we keep tokens in $\mathcal{S}$ and mask input tokens in $\mathcal{X} \setminus \mathcal{S}$ to zeros. Thus, the global similarity only considers the tokens in $\mathcal{S}$, which reflects the contribution of the tokens in $\mathcal{S}$ to the global similarity judgment.

Semantic Region Generation. Inspired by *YOLOv3* [40], we design a lightweight region generation module. It takes each patch token representation $\tilde{\mathbf{x}}_i^I$ as input and generates a bounding box prediction centered on $\tilde{\mathbf{x}}_i^I$, which corresponds to a visual region $\mathcal{R}_i = \{\mathbf{x}_{i,k}^I\}_{k=1}^{K_i}$ with K_i patch tokens. The region generation module also predicts a confidence score $s(\mathcal{R}_i)$ for each region. We select the top-M predictions as the semantic regions. Then, the Shapley interaction of $\mathcal{R}_i$ can be defined as:

$$\mathfrak{I}([\mathcal{R}_i]) = \phi([\mathcal{R}_i]|X \setminus \mathcal{R}_i \cup \{[\mathcal{R}_i]\}) - \sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} \phi(\mathbf{x}_{i,k}^I|\mathcal{X} \setminus \mathcal{R}_i \cup \{\mathbf{x}_{i,k}^I\}) \tag{5}$$

According to the Equation 3, we can reformulate Shapley value into the form of expectation:

$$\phi([\mathcal{R}_i]|\mathcal{X} \setminus \mathcal{R}_i \cup \{[\mathcal{R}_i]\}) = \mathop{\mathbb{E}}_{c}\{\mathop{\mathbb{E}}_{\substack{\mathcal{S} \subseteq \mathcal{X} \setminus \mathcal{R}_i \\ |\mathcal{S}|=c}} [v_1(\mathcal{S} \cup \mathcal{R}_i) - v_1(\mathcal{S})]\} \tag{6}$$

where c represents the coalition size. $\phi(\mathbf{x}_{i,k}^I|X \setminus \mathcal{R}_i \cup \{\mathbf{x}_{i,k}^I\})$ can be defined in a similar manner, and the Shapley interaction of $\mathcal{R}_i$ can be reformulated as (we provide the proof in Appendix C):

$$\mathfrak{I}([\mathcal{R}_i]) = \mathop{\mathbb{E}}_{c}\{\mathop{\mathbb{E}}_{\substack{\mathcal{S} \subseteq \mathcal{X} \setminus \mathcal{R}_i \\ |\mathcal{S}|=c}} [v_1(\mathcal{S} \cup \mathcal{R}_i) - \sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} v_1(\mathcal{S} \cup \{\mathbf{x}_{i,k}^I\}) + (K-1)v_1(\mathcal{S})]\} \tag{7}$$

Taking normalized $\mathfrak{I}'([R_i])$ as the soft supervision label, the token-based semantic aggregation loss is defined as cross-entropy loss:

$$\mathcal{L}_{\text{TSA}} = -\frac{1}{M}\sum_{i=1}^{M}[\mathfrak{I}'([\mathcal{R}_i])\log(s(\mathcal{R}_i)) + (1-\mathfrak{I}'([\mathcal{R}_i]))\log(1-s(\mathcal{R}_i))] \tag{8}$$

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 $\mathcal{R}_i$ to obtain region representation $\mathbf{h}_i^I \in \mathbb{R}^d$. We employ an off-the-shelf constituency parser to extract phrases from text and obtain phrase representation $\mathbf{h}_i^T \in \mathbb{R}^d$ by `Avg-Pooling`. Totally, we obtain M regions $\mathcal{H}^I = \{\mathbf{h}_i^I\}_{i=1}^M$ and N phrases $\mathcal{H}^T = \{\mathbf{h}_j^T\}_{j=1}^N$. And the alignment matrix can be defined as: $\mathbf{A} = [a_{ij}]^{M\times N}$, where $a_{ij} = \mathbf{h}_i^{I\top}\mathbf{h}_j^T$ represents the alignment score between i-th region and j-th phrase. Next, we apply softmax-normalization over each row of $\mathbf{A}$, obtaining $\tilde{\mathbf{A}}$. For the i-th region, we calculate its maximum alignment score as $\max_j \tilde{a}_{ij}$. 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 $\mathcal{H} = \mathcal{H}^I \cup \mathcal{H}^T$ 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:

$$\mathfrak{I}([\mathcal{H}_{ij}]) = \phi([\mathcal{H}_{ij}]|\mathcal{H}\setminus\mathcal{H}_{ij}\cup\{[\mathcal{H}_{ij}]\}) - \phi(\mathbf{h}_i^I|\mathcal{H}\setminus\mathcal{H}_{ij}\cup\{\mathbf{h}_i^I\}) - \phi(\mathbf{h}_j^T|\mathcal{H}\setminus\mathcal{H}_{ij}\cup\{\mathbf{h}_j^T\}) \tag{9}$$

$$= \mathbb{E}_c\{\mathbb{E}_{\substack{\mathcal{S}\subseteq\mathcal{H}\setminus\mathcal{H}_{ij}\\|\mathcal{S}|=c}}[v_2(\mathcal{S}\cup\mathcal{H}_{ij}) - v_2(\mathcal{S}\cup\{\mathbf{h}_i^I\}) - v_2(\mathcal{S}\cup\{\mathbf{h}_j^T\}) + v_2(\mathcal{S})]\} \tag{10}$$

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

$$\mathcal{L}_{\text{FSA}} = -\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\mathfrak{I}'([\mathcal{H}_{ij}])\log(\tilde{a}_{ij}) \tag{11}$$

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 [35]. Previous methods mainly apply sampling-based method [6] 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 [6] estimates the expectation terms in Equation 7 and Equation 10 by sampling to compute the Shapley interaction. Inspired by noisy label learning [20], the UNSIL module learns to predict the Shapley interaction and the corresponding uncertainty $\sigma \in (0,1)$. Intuitively, if the UNSIL module makes a prediction with low uncertainty, we can directly apply its prediction to $\mathcal{L}_{\text{TSA}}$ and $\mathcal{L}_{\text{FSA}}$, 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 σ. Then, we sample a value ϵ from a uniform distribution on $(0,1)$. If $\epsilon > \sigma$, we directly use its prediction. If

Table 1: Results (%) of zero-shot image-text retrieval on Flickr30K and MSCOCO datasets.

	Flickr30K						*MSCOCO*					
	image-to-text			*text-to-image*			*image-to-text*			*text-to-image*		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
UNITER	83.6	95.7	97.7	68.7	89.2	93.9	-	-	-	-	-	-
CLIP	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN	88.6	98.7	99.7	75.7	93.8	**96.8**	58.6	83.0	89.7	45.6	69.8	78.6
FILIP	89.8	99.2	99.8	75.0	93.4	96.3	61.3	84.3	90.4	45.9	70.6	79.3
LOUPE	**90.5**	**99.5**	**99.8**	**76.3**	**93.9**	96.7	**62.3**	**85.1**	**91.2**	**50.1**	**75.4**	**83.3**

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

	CIFAR10	Food101	StanfordCars	SUN397	Flowers102	Country211	FER2013	Aircrafts	OxfordPets	Caltech101	ImageNet	Average
CLIP	**96.2**	92.9	77.3	67.7	78.7	34.9	**57.7**	36.1	93.5	92.6	75.3	73.0
LOUPE	95.9	**94.3**	**79.9**	**69.8**	**87.4**	**37.8**	53.3	**54.9**	**94.1**	**93.9**	**76.1**	**76.1**

$\epsilon \leq \sigma$, 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 $\mathfrak{I}^*$ and $\hat{\mathfrak{I}}$ denote the results from the sampling-based method and UNSIL module, respectively. Taking $\mathfrak{I}^*$ as the ground-truth, the UNSIL module is trained by:

$$\mathcal{L}_{\text{UNSIL}} = \frac{1}{\beta_1 \sigma} \mathcal{L}_{\text{MSE}}(\hat{\mathfrak{I}}, \mathfrak{I}^*) + \beta_2 \sigma \tag{12}$$

where the first term is mean squared error $\mathcal{L}_{\text{MSE}}$ weighted by the uncertainty, the second term serves as a regularization term for the prediction uncertainty, and β 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 [31] and the text encoder by BERT-Small [9]. The input images are resized to 224×224 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 [32] optimizer with a learning rate of 2×10^{-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 [30] and Flickr30K [36] 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.

	COCO		*PASCAL VOC*		*RefCOCO*			*RefCOCO+*		
	mAP@0.3	mAP@0.5	mAP@0.3	mAP@0.5	val	testA	testB	val	testA	testB
CLIP + Pixel-Wise	8.5	4.5	18.2	7.3	6.7	6.2	5.8	6.1	7.0	5.7
CLIP + K-Means	6.4	1.9	11.7	4.8	2.1	2.3	1.7	1.7	2.0	2.8
CLIP + Grad-CAM	7.1	3.2	19.1	8.2	5.5	5.2	4.8	4.4	5.6	4.9
AdaptCLIP	14.9	6.6	28.7	12.9	16.7	18.4	18.0	17.5	18.9	19.6
LOUPE	**25.3**	**12.1**	**30.3**	**19.5**	**25.2**	**26.8**	**24.5**	**22.9**	**23.3**	**23.6**

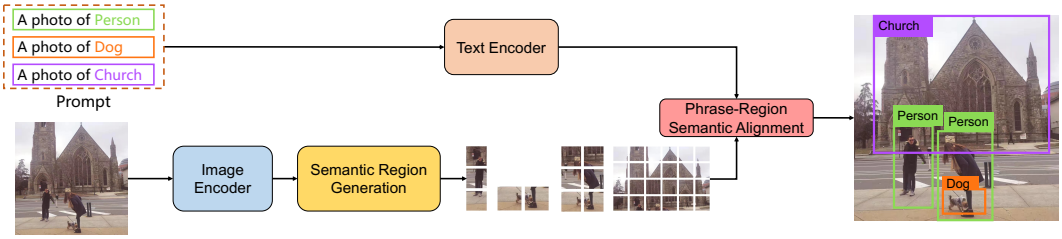


Figure 2: An example of LOUPE zero-shot transferring to object detection using prompt templates.

CLIP [38]. 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

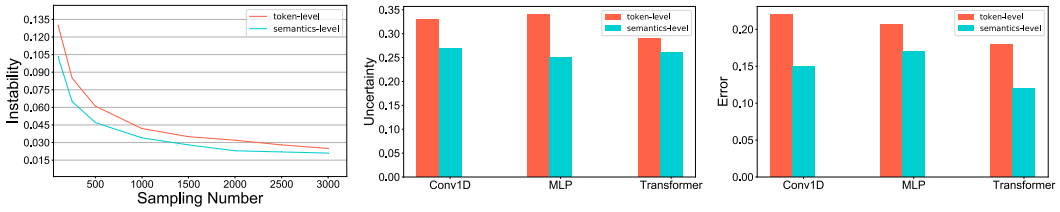
To answer whether our model has learned fine-grained semantics, we further evaluate LOUPE on object detection [42] and visual grounding [54], which require more fine-grained semantic understanding ability to identify specific visual regions in images according to the object labels or referring expressions. 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 [24], which is a concurrently unpublished method that leverages classic super-pixel (SLIC [1]) and bounding box proposal (selective search [47]) 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 [30] (65 classes) and PASCAL VOC [11] (20 classes). For visual grounding, we evaluate their top-1 accuracy at IoU thresholds of 0.5 on RefCOCO [54] and RefCOCO+ [54]. 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.

		MSCOCO		*COCO*		*RefCOCO*			Training Time
		I2T	T2I	mAP@0.3	mAP@0.5	val	testA	testB	(sec/iter)
1	Backbone	31.0	24.8	3.8	1.0	1.3	0.9	0.8	1.17
2	Backbone + $\mathcal{L}_{\text{TSA}}$	32.4	26.2	7.6	3.3	1.8	2.0	2.6	8.38
3	Backbone + $\mathcal{L}_{\text{TSA}}$ + $\mathcal{L}_{\text{FSA}}$	33.5	28.3	9.4	5.9	4.1	4.6	4.3	9.90
4	Backbone + $\mathcal{L}_{\text{TSA}}$ + $\mathcal{L}_{\text{FSA}}$ + UNSIL	33.3	28.1	9.0	5.6	4.5	4.9	4.4	1.93



(a) Instability (b) Uncertainty (c) Error

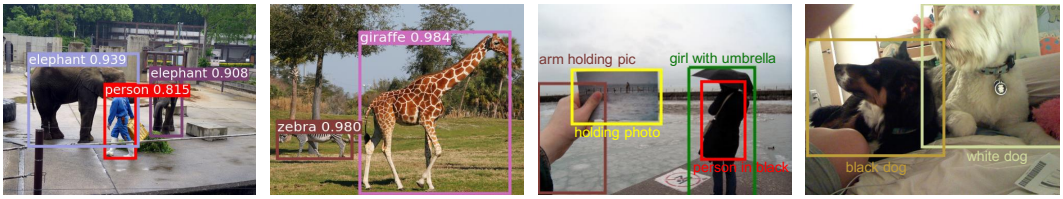
Figure 3: (a) Instability of the Shapley interaction approximation with respect to different sampling numbers. (b, c) Uncertainty and error of the UNSIL module with different structures.

As time-consuming human annotations are unscalable for massive object classes in the real world, some recent works [4, 39] 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, 55] 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 [44]). 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 $\mathcal{L}_{\text{TSA}}$ (Row 2), semantics-level Shapley interaction modeling supervision $\mathcal{L}_{\text{FSA}}$ (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 $\mathcal{L}_{\text{TSA}}$ and $\mathcal{L}_{\text{FSA}}$ bring significant improvement for all tasks. We observe that $\mathcal{L}_{\text{TSA}}$ 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 [57], 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 $\frac{\mathbb{E}_{u,v:u\neq v}|\mathfrak{I}_u-\mathfrak{I}_v|}{\mathbb{E}_w|\mathfrak{I}_w|}$, where $\mathfrak{I}_w$ 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*.



(a) Object Detection (b) Visual Grounding

Figure 4: Qualitative examples of object detection on COCO and visual grounding on RefCOCO+.

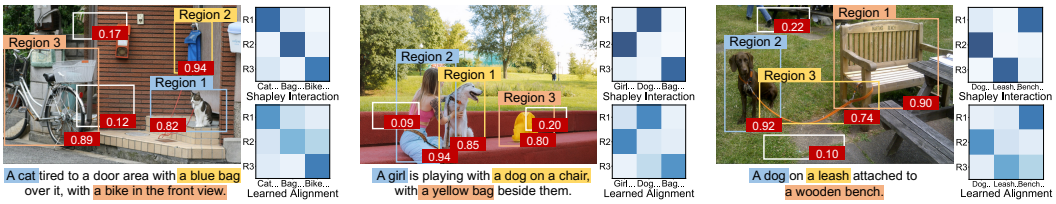


Figure 5: Visualization of learned fine-grained semantic alignment and corresponding Shapley interaction values. The values in the red boxes represent the Shapley interaction of regions.

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

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[4] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[7] 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 *European conference on computer vision*, pages 104–120. Springer, 2020.

[8] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Mark Everingham, Andrew Zisserman, Christopher KI Williams, Luc Van Gool, Moray Allan, Christopher M Bishop, Olivier Chapelle, Navneet Dalal, Thomas Deselaers, Gyuri Dorkó, et al. The pascal visual object classes challenge 2007 (voc2007) results. 2008.

[12] Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[15] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*, 2021.

[16] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[17] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[18] 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.

[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.

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

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

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[23] 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, 2021.

[24] Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. *arXiv preprint arXiv:2204.03647*, 2022.

[25] Juncheng Li, Xin Wang, Siliang Tang, Haizhou Shi, Fei Wu, Yueting Zhuang, and William Yang Wang. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12123–12132, 2020.

[26] 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. *arXiv*

preprint arXiv:2112.03857, 2021.
[27] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.
[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
[29] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.
[30] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.
[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
[33] 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.
[34] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
[35] Yasuko Matsui and Tomomi Matsui. Np-completeness for calculating power indices of weighted majority games. *Theoretical Computer Science*, 263(1-2):305–310, 2001.
[36] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
[37] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.
[38] 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.
[39] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11932–11939, 2020.
[40] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv e-prints*, 2018.
[41] Jie Ren, Die Zhang, Yisen Wang, Lu Chen, Zhanpeng Zhou, Yiting Chen, Xu Cheng, Xin Wang, Meng Zhou, Jie Shi, et al. A unified game-theoretic interpretation of adversarial robustness. *arXiv preprint arXiv:2111.03536*, 2021.
[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
[43] Lloyd S Shapley. A value for n-person games, contributions to the theory of games, 2, 307–317, 1953.
[44] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.
[45] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
[46] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
[47] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.
[48] 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.
[49] Robert J Weber. Probabilistic values for games. *The Shapley Value. Essays in Honor of Lloyd S. Shapley*, pages 101–119, 1988.
[50] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022.
[51] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.
[52] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[53] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.
[54] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.
[55] 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.
[56] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.
[57] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. *arXiv preprint arXiv:2010.05045*, 2020.
[58] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.
[59] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.
[60] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.
[61] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702, 2020.

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 are using 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.

A Appendix Overview

In this appendix, we present:

- Axiomatic Properties of Shapley Value (Section B).
- Proofs of Equation 7 and Equation 10 (Section C).
- Hyperparameters and Implementation Details (Section D).
- Pre-Training and Evaluation Details (Section E).
- More Experiment Results on Downstream Vision-Language Generation Task (Section F).
- Further Analysis on the Image Encoder (Section G).
- More Qualitative Examples on Object Detection and Visual Grounding (Section H).
- Linear Probing Evaluation (Section I).
- Training Efficiency Discussion (Section J).
- Detailed Discussion with Some Related Works (Section K).

B Axiomatic Properties of Shapley Value

In this section, we mainly introduce the axiomatic properties of Shapley value. *Weber et al.* [49] have proved that Shapley value is the unique metric that satisfies the following axioms: *Linearity*, *Symmetry*, *Dummy*, and *Efficiency*.

Linearity Axiom. If two independent games u and v can be linearly merged into one game $w(\mathcal{S}) = u(\mathcal{S}) + v(\mathcal{S})$, then the Shapley value of each player $i \in \mathcal{N}$ in the new game w is the sum of Shapley values of the player i in the game u and v, which can be formulated as:

$$\phi_w(i|\mathcal{N}) = \phi_u(i|\mathcal{N}) + \phi_v(i|\mathcal{N}) \tag{13}$$

Symmetry Axiom. Considering two players i and j in a game v, if they satisfy:

$$\forall \mathcal{S} \in \mathcal{N} \setminus \{i, j\}, v(\mathcal{S} \cup \{i\}) = v(\mathcal{S} \cup \{j\}) \tag{14}$$

then $\phi_v(i|\mathcal{N}) = \phi_v(j|\mathcal{N})$.

Dummy Axiom. The dummy player is defined as the player that has no interaction with other players. Formally, if a player i in a game v satisfies:

$$\forall \mathcal{S} \in \mathcal{N} \setminus \{i\}, v(\mathcal{S} \cup \{i\}) = v(\mathcal{S}) + v(\{i\}) \tag{15}$$

then this player is defined as the dummy player. In this way, the dummy player i has no interaction with other players, *i.e.*, $v(\{i\}) = \phi_v(i|\mathcal{N})$.

Efficiency Axiom. The efficiency axiom ensures that the overall reward can be assigned to all players, which can be formulated as:

$$\sum_{i \in \mathcal{N}} \phi_v(i) = v(\mathcal{N}) - v(\varnothing) \tag{16}$$

C Proofs of Equation 7 and Equation 10

In this section, we provide detailed proofs for Equation 7 in Section 3.2.2 and Equation 10 in Section 3.2.3.

We first provide proof for Equation 7. The token-level Shapley interaction for $\mathcal{R}_i$ can be decomposed as follows:

$$\mathfrak{I}([\mathcal{R}_i]) = \phi([\mathcal{R}_i]|X \setminus \mathcal{R}_i \cup \{[\mathcal{R}_i]\}) - \sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} \phi(\mathbf{x}_{i,k}^I|\mathcal{X} \setminus \mathcal{R}_i \cup \{\mathbf{x}_{i,k}^I\}) \tag{17}$$

$$= \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{X}\setminus\mathcal{R}_i \\ |\mathcal{S}|=c}} [v_1(\mathcal{S}\cup\mathcal{R}_i) - v_1(\mathcal{S})]\} - \sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{X}\setminus\mathcal{R}_i \\ |\mathcal{S}|=c}} [v_1(\mathcal{S}\cup\{\mathbf{x}_{i,k}^I\}) - v_1(\mathcal{S})]\} \tag{18}$$

$$= \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{X}\setminus\mathcal{R}_i \\ |\mathcal{S}|=c}} [v_1(\mathcal{S}\cup\mathcal{R}_i) - v_1(\mathcal{S})]\} - \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{X}\setminus\mathcal{R}_i \\ |\mathcal{S}|=c}} [\sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} (v_1(\mathcal{S}\cup\{\mathbf{x}_{i,k}^I\}) - v_1(\mathcal{S})\ )]\} \tag{19}$$

$$= \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{X}\setminus\mathcal{R}_i \\ |\mathcal{S}|=c}} [v_1(\mathcal{S}\cup\mathcal{R}_i) - v_1(\mathcal{S}) - \sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} (v_1(\mathcal{S}\cup\{\mathbf{x}_{i,k}^I\}) - v_1(\mathcal{S})\ )\ ]\ \} \tag{20}$$

$$= \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{X}\setminus\mathcal{R}_i \\ |\mathcal{S}|=c}} [v_1(\mathcal{S}\cup\mathcal{R}_i) - v_1(\mathcal{S}) - \sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} v_1(\mathcal{S}\cup\{\mathbf{x}_{i,k}^I\}) + \sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} v_1(\mathcal{S})\ ]\ \} \tag{21}$$

$$= \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{X}\setminus\mathcal{R}_i \\ |\mathcal{S}|=c}} [v_1(\mathcal{S}\cup\mathcal{R}_i) - \sum_{\mathbf{x}_{i,k}^I \in \mathcal{R}_i} v_1(\mathcal{S}\cup\{\mathbf{x}_{i,k}^I\}) + (K-1)v_1(\mathcal{S}\ )\ ]\ \} \tag{22}$$

We then provide proof for Equation 10. The semantics-level Shapley interaction between region i and phrase j can be decomposed as follows:

$$\mathfrak{I}([\mathcal{H}_{ij}]) = \phi([\mathcal{H}_{ij}]|\mathcal{H} \setminus \mathcal{H}_{ij} \cup \{[\mathcal{H}_{ij}]\}) - \phi(\mathbf{h}_i^I|\mathcal{H} \setminus \mathcal{H}_{ij} \cup \{\mathbf{h}_i^I\}) - \phi(\mathbf{h}_j^T|\mathcal{H} \setminus \mathcal{H}_{ij} \cup \{\mathbf{h}_j^T\}) \tag{23}$$

$$= \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{H}\setminus\mathcal{H}_{ij} \\ |\mathcal{S}|=c}} [v_2(\mathcal{S}\cup\mathcal{H}_{ij}) - v_2(\mathcal{S})]\} - \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{H}\setminus\mathcal{H}_{ij} \\ |\mathcal{S}|=c}} [v_2(\mathcal{S}\cup\{\mathbf{h}_i^I\}) - v_2(\mathcal{S})]\} \tag{24}$$

$$- \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{H}\setminus\mathcal{H}_{ij} \\ |\mathcal{S}|=c}} [v_2(\mathcal{S}\cup\{\mathbf{h}_j^T\}) - v_2(\mathcal{S})]\} \tag{25}$$

$$= \mathbb{E}_c\{ \mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{H}\setminus\mathcal{H}_{ij} \\ |\mathcal{S}|=c}} [v_2(\mathcal{S}\cup\mathcal{H}_{ij}) - v_2(\mathcal{S}\cup\{\mathbf{h}_i^I\}) - v_2(\mathcal{S}\cup\{\mathbf{h}_j^T\}) + v_2(\mathcal{S})\ ]\ \} \tag{26}$$

D Hyperparameters and Implementation Details

In this section, we summarize the hyperparameters in our LOUPE model in Table 5, including the hyperparameters of the image encoder, text encoder, and pre-training process. For the uncertainty-aware neural Shapley interaction learning module, we attempt three kinds of models (*i.e., Conv1D, 3-Layer MLP + Attention, 3-Layer Transformer*) to implement it for token-level and semantics-level Shapley interaction approximation.

For token-level Shapley interaction approximation, it takes the patch token sequence $\mathcal{X}^I = \{\mathbf{x}_i^I\}_{i=1}^{L_1}$, word token sequence $\mathcal{X}^T = \{\mathbf{x}_i^T\}_{i=1}^{L_2}$, and the visual region $\mathcal{R}_i = \{\mathbf{x}_{i,k}^I\}_{k=1}^{K_i}$ as input, and estimates the corresponding token-level Shapley interaction value for $\mathcal{R}_i$ along with the uncertainty σ.

Conv1D model first performs `Avg-Pooling` over learned patch representations of $\mathcal{R}_i$ to obtain the region representation $\mathbf{h}_i^I$, and then fuse the word and patch token representations with the region representation $\mathbf{h}_i^I$, respectively. Specifically, we project them into an unified semantic space by fully-connected layers and then fuse them through Hadamard product as:

$$\mathcal{F}^I = (\mathcal{W}_1 \mathbf{h}_i^I \mathbf{1}^T) \odot (\mathcal{W}_2 \mathcal{X}^I) \tag{27}$$

where $\mathcal{W}_1$ and $\mathcal{W}_2$ are the learnable projection parameters, $\mathbf{1}^T$ is the transpose of an all-ones vector, and $\odot$ represents Hadamard product. We can obtain $\mathcal{F}^T$ in a similar manner. Then, we apply 1D

Table 5: A summary of various hyperparameters in LOUPE.

Hyperparameter	**Value**
Image Encoder - Swin-L	
input image size	224×224
stage 1 - patch size	4×4
stage 1 - hidden size	192
stage 1 - window size	7×7
stage 1 - number of heads	6
stage 2 - patch size	8×8
stage 2 - hidden size	384
stage 2 - window size	7×7
stage 2 - number of heads	12
stage 3 - patch size	16×16
stage 3 - hidden size	768
stage 3 - window size	7×7
stage 3 - number of heads	24
stage 4 - patch size	32×32
stage 4 - hidden size	1536
stage 4 - window size	7×7
stage 4 - number of heads	48
Text Encoder - BERT-Small	
maximum length of word tokens	60
vocabulary size	30522
attention dropout probability	0.1
hidden activation function	GELU
hidden dropout probability	0.1
initializer range	0.02
intermediate size	2048
layer norm eps	$1e^{-12}$
hidden size	512
number of attention heads	8
number of hidden layers	4
Pre-Training	
number of epochs	20
batch size	512
learning rate	2e-4
learning schedule	`OneCycle`
cycle momentum	Ture
base momentum	0.85
max momentum	0.95
AdamW weight decay	0.01
AdamW β_1	0.9
AdamW β_2	0.999

convolution operation with kernel size = 4 and stride = 2 over $\mathcal{F}^I$ and $\mathcal{F}^T$, respectively. Following with `Max-Pooling` operation, we obtain $\tilde{\mathbf{f}}^I \in \mathbb{R}^d$ and $\tilde{\mathbf{f}}^T \in \mathbb{R}^d$. Next, we concatenate them with $\mathbf{h}_i^I$ and feed them to two separate 1-layer fully connected layers to get the Shapley interaction estimation and corresponding uncertainty.

3-Layer MLP + Attention model first performs `Avg-Pooling` over learned patch representations of $\mathcal{R}_i$ to obtain the region representation $\mathbf{h}_i^I$. Then, we use $\mathbf{h}_i^I$ as the query to attend the patch token sequence and compute a weighted sum of the patch token representations as:

$$\tilde{\alpha}_j^I = \mathcal{W}_3(tanh(\mathcal{W}_4 \mathbf{h}_i^I + \mathcal{W}_5 \mathbf{x}_j^I)) \tag{28}$$

$$\alpha^I = softmax([\tilde{\alpha}_1^I, ..., \tilde{\alpha}_{L_1}^I]) \tag{29}$$

$$\mathbf{e}^I = \sum_{j=1} \alpha_i^I \mathbf{x}_j^I \tag{30}$$

Where L_1 is the number of patch tokens. We can obtain $\mathbf{e}^T$ for word token sequence in a similar manner. Consequently, we concatenate them and $\mathbf{h}_i^I$ and feed them to two separate 3-layer fully connected layers to get the Shapley interaction estimation and corresponding uncertainty.

3-Layer Transformer model takes the concatenated sequence $\mathcal{X}^I$ and $\mathcal{X}^T$ as input. We add position embeddings and three kinds of token type embeddings (*i.e., word token, context patch token, region patch token*) to them. We then apply three layers of transformer blocks to jointly encode the input sequence and take the output `[CLS]` token to predict the Shapley interaction estimation and corresponding uncertainty, separately.

For semantics-level Shapley interaction approximation, it takes the M regions $\mathcal{H}^I = \{\mathbf{h}_i^I\}_{i=1}^M$, N phrases $\mathcal{H}^T = \{\mathbf{h}_j^T\}_{j=1}^N$, and the target region-phrase pair $< \mathbf{h}_i^I, \mathbf{h}_j^T >$ as input, and estimates the corresponding semantics-level Shapley interaction value for $< \mathbf{h}_i^I, \mathbf{h}_j^T >$ along with the uncertainty σ. The architectures of the three models are consistent with their token-level implementations.

E Pre-Training and Evaluation Details

E.1 Pre-Training Dataset Details

As recent works [18, 38, 52] have shown that pre-training models can obtain great performance gain by scaling up the dataset, we construct a large-scale dataset, which consists of 240 million image-text pairs and covers a broad set of visual concepts. Concretely, we elaborate more details in the following.

Raw image-text pair collection. We first harvest large-scale noisy image-text pairs from the web and design multiple filtering rules to improve the quality of the web data.

Image-based filtering. Following ALIGN [18], we remove pornographic images and keep only images where both dimensions are larger than 200 pixels. Also, we remove the images whose aspect ratio is larger than 10. To prevent from leaking testing data, we remove the images that appear in all downstream evaluation datasets (e.g., MSCOCO, Flickr30K).

Text-based filtering. We remove the repeated captions and keep only English texts. The texts that are shorter than 3 words or longer than 100 words are discarded. As ALIGN [18], we also remove the texts that contain any rare token (outside of 100 million most frequent unigrams and bigrams from the raw dataset).

Joint image-text filtering. Although the above filtering rules have filtered out many noisy data, it is hard to detect the mismatched image-text pairs, where the texts do not accurately describe the visual content of the images, resulting in undesirable noisy signals to vision-language pre-training. Inspired by BLIP [22], we train a discriminator as a filtering model to predict whether the text is matched to the image. Specifically, the filtering model consists of an image encoder and an image-grounded text encoder, which takes the cross-attention to fuse image features and text features. The filtering model is trained on CC12M dataset using image-text contrastive loss and image-text matching loss.

E.2 Evaluation Details

Zero-Shot Image-Text Retrieval. We evaluate the zero-shot performance of LOUPE on the image-text retrieval task over the widely used Flickr30K [36] and MSCOCO [30] datasets. The image-text retrieval consists of two subtasks: image-to-text retrieval and text-to-image retrieval, where a model is required to identify an image from candidates given a caption describing its content, or vice versa. The MSCOCO dataset consists of 123,287 images, and each image is aligned with five captions. The Flickr30K dataset contains 31,783 images and five captions for each image. Following previous works [18, 52], we evaluate the zero-shot performance on the 1K and 5K test sets of Flickr30K and MSCOCO, respectively. We take the final representation of `[CLS]` tokens as the global representations of images and texts, and use them to measure the image-text similarity. We first compute the similarity scores for all image-text pairs. Then, we take the top-K candidates for ranking and report the top-K retrieval results.

Zero-Shot Transfer to Object Detection. Without any fine-tuning, we evaluate the zero-shot transfer performance of LOUPE on the object detection task [42] over the COCO [30] and PASCAL VOC [11] datasets. For the COCO Objects dataset, we use their 2017 validation split for evaluation. Previous

Table 6: Image captioning evaluation results on COCO "Karpathy" test split.

	Image Captioning			
	BLEU@4	METEOR	CIDEr	SPICE
VLP [60]	36.5	28.4	117.7	21.3
$\text{OSCAR}_{\text{large}}$ [28]	37.4	30.7	127.8	23.5
$\text{VinVL}_{\text{large}}$ [58]	38.5	30.4	130.8	23.4
$\text{BLIP}_{\text{ViT-L}}$ [22]	40.4	-	136.7	-
$\text{LEMON}_{\text{large}}$ [15]	40.6	30.4	135.7	23.5
LOUPE	**40.9**	**31.5**	**137.8**	**24.3**

zero-shot object detection models [4, 39, 61] follow the split proposed by [4], which consists of 48 base classes and 17 novel classes. They train models on base classes and evaluate models on novel classes. Differently, we directly evaluate the zero-shot transfer performance on both the base and novel classes, without fine-tuning on the base classes. Totally, we evaluate models on 4,836 test images that contain 33,152 instances of 65 classes. PASCAL VOC is a widely used object detection dataset, which contains 20 object classes. For PASCAL VOC, we evaluate models on 9657 instances of 5072 images. To complete object detection, we first use the region generation module to generate a set of candidate regions and then use prompt text (*i.e., an image of [object class name].*) to expand each detection label to a sentence. Next, we encode sentences for each object class by the learned text encoder and measure their similarity with the candidate regions as the classification scores. Following most zero-shot object detection methods, we use mean Average Precision (mAP) at IoU of $\{0.3, 0.5\}$ as evaluation metrics.

Zero-Shot Transfer to Visual Grounding. Visual grounding [54] (also known as phrase localization and referring expression comprehension) aims to locate a specific visual region of the input image, according to the language referring expression. Visual grounding can be seen as generalized object detection, where the pre-defined class labels are replaced by language referring expression sentences. Without any fine-tuning, we evaluate the zero-shot transfer performance of LOUPE on the visual grounding task over the RefCOCO [54] and RefCOCO+ [54] datasets. These two datasets are collected by the ReferitGame [19], where a player is asked to write a language expression to refer to a specific object in the image, and another player is required to locate the target object given the image and the referring expression. RefCOCO dataset consists of 142,209 refer expressions for 50,000 objects in 19,994 images, which is split into train (120,624 expressions), val (10,834 expressions), testA (5,657 expressions), testB (5,095 expressions) sets. The images in testA set involve multiple persons and the images in testB set involve multiple objects. RefCOCO+ dataset consists of 141,564 expressions for 49,856 objects in 19,992 images, which is split into train (120,191 expressions), val (10,758 expressions), testA (5,726 expressions), testB (4,889 expressions) sets. We report the zero-shot transfer performance on the val, testA, and testB sets of both datasets.

F More Experiment Results on Vision-Language Generation Task

To further validate the generalization ability of the learned cross-modal representations by our LOUPE, we adapt the pre-trained LOUPE to vision-language generation task, *i.e.,* image captioning [2]. Image captioning is the task of describing images with natural languages, which requires models to identify and describe the fine-grained semantics of images. The input images are encoded by the learned image encoder. As BLIP [22], we train an image-grounded text decoder which shares the feed forward layers with the learned text encoder and adopts cross-attention to attend to the image features. The text decoder is trained with a language modeling loss to generate captions according to the images.

We evaluate the image captioning performance on the MSCOCO [30] dataset, which is split into train (113, 287 images), val (5,000 images), "Karpathy" test split (5,000 images). Each image has 5 captions. We use the train split to train the image-grounded text decoder and report the performance on the public "Karpath" 5k test split. Following standard metrics, we use BLEU@4, METEOR, CIDEr, and SPICE as evaluation metrics. We compare our LOUPE model with recent vision-language pre-training generation models [15, 22, 28, 58, 60]. All methods are fine-tuned with cross-entropy loss only, without CIDEr optimization. As shown in Table 6, our LOUPE achieves competitive performance on all metrics, which verifies the strong generalization ability of our model on downstream vision-language generation tasks.

Table 7: Further ablation results (R@1) with respect to different image encoders.

		Image Encoder	*Flickr30K*		*MSCOCO*	
			image-to-text	*text-to-image*	*image-to-text*	*text-to-image*
1	ALIGN [18]	EfficientNet	88.6	75.7	58.6	45.6
2	FILIP [52]	ViT-L	89.8	75.0	61.3	45.9
3	CLIP [38]	ViT-L	88.0	68.7	58.4	37.8
4	CLIP*	Swin-L	88.7	74.3	59.3	46.2
5	**LOUPE**	Swin-L	**90.5**	**76.3**	**62.3**	**50.1**

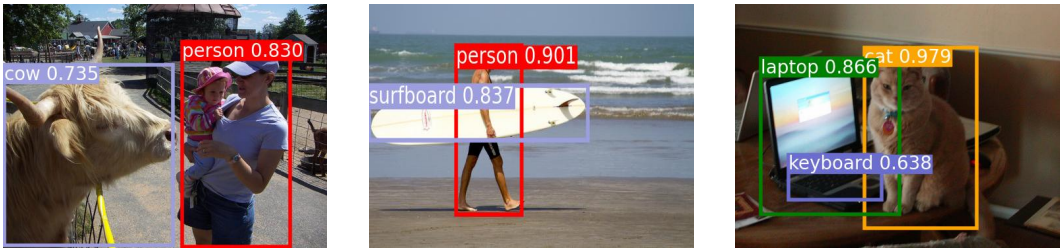


Figure 6: Qualitative examples of object detection on COCO Objects dataset.

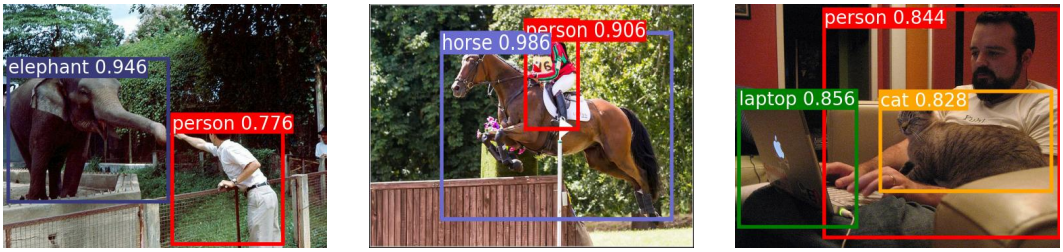


Figure 7: Qualitative examples of object detection on PASCAL VOC dataset.

G Further Analysis on the Image Encoder

In our work, we adopt the Swin-L [31] as our image encoder due to the following considerations. (1) The shifted windowing scheme of Swin Transformer achieves linear computational complexity with respect to image size, which is more efficient than ViT [10]. This merit is particularly beneficial to the vision-language pre-training as we need to process large-scale images (240M). (2) The hierarchical architecture of Swin Transformer is more flexible to model semantic regions at various scales.

To further verify the performance gain from our proposed fine-grained semantically aligned vision-language pre-training framework, we implement a variant version of CLIP that adopts Swin-L as the image encoder (Row 4 in Table 7), using the same training dataset as our LOUPE. It can also be viewed as the backbone of our LUOPE (without optimization from our token-level and semantics-level Shapley interaction modeling). As shown in Table 7, comparing CLIP* with CLIP, the Swin-L image encoder does bring some improvements over CLIP. However, there is still a clear performance gap between CLIP* and our LOUPE. With the same architecture, our LOUPE has 2.68 points higher average R@1 than the CLIP* over two datasets. This further verifies that the main performance gain comes from our proposed fine-grained semantically aligned vision-language pre-training framework. Notably, we observe that the text-to-image retrieval of our implementation is obviously higher than CLIP. This phenomenon has also been confirmed by [18, 52] (see Row 1 and Row 2 in Table 7). We suppose that it might be caused by some training details or the dataset collection of CLIP.

H More Qualitative Examples on Object Detection and Visual Grounding

For a more intuitive view of the performance of our model on object detection and visual grounding, we visualize more qualitative examples. Concretely, Figure 6 and Figure 7 show more object detection examples on the COCO [30] and PASCAL VOC [11] datasets. Figure 8 and Figure 9 show more visual grounding examples on the RefCOCO [54] and RefCOCO+ [54] datasets.

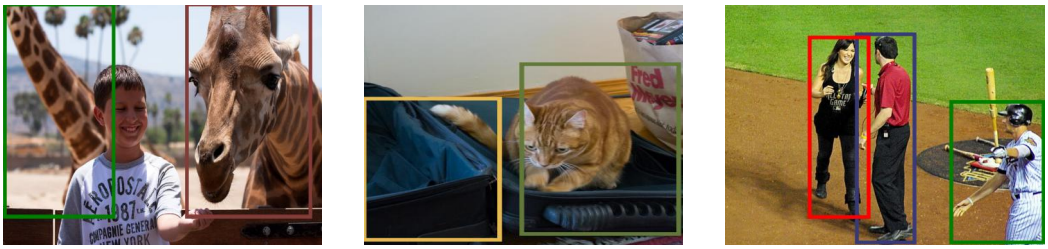


Figure 8: Qualitative examples of visual grounding on RefCOCO dataset.

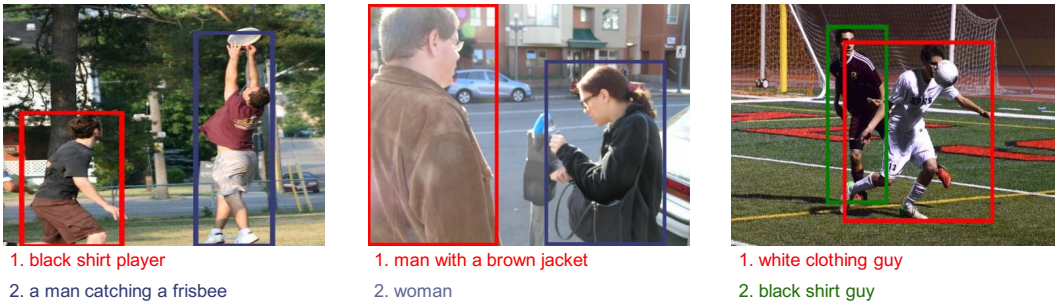


Figure 9: Qualitative examples of visual grounding on RefCOCO+ dataset.

Table 8: Linear probing performance (top-1 accuracy) over 11 datasets.

	CIFAR10	Food101	StanfordCars	SUN397	Flowers102	Country211	FER2013	Aircrafts	OxfordPets	Caltech101	ImageNet
CLIP	**98.0**	95.2	90.9	81.8	99.2	46.4	**72.9**	69.4	95.1	96.5	83.9
LOUPE	97.6	**96.0**	**92.1**	**82.6**	**99.5**	**49.3**	70.7	**80.2**	**95.5**	**97.5**	**85.7**

Table 9: Comparison of training cost and architecture parameters.

	Pre-Training Image-Text Pairs	Parameters	GPUs	Days	GPU Days
CLIP	400M	425M	256 V100	12 days	3072
ALIGN	1800M	820M	1024 TPUv3	-	-
FILIP	340M	417M	192 V100	24 days	4608
LOUPE	240M	226M	128 V100	20 days	2560

I Linear Probing Evaluation

In this section, we evaluate the linear probing performance of our LOUPE on image classification. Following the same evaluation setting as CLIP [38], we freeze the whole backbone network and only fine-tuning the last linear classification layer, which takes the `[CLS]` token as input. We report the linear probing performance over 11 datasets in Table 8. Our LOUPE outperforms CLIP with average improvement of 1.6%. Notably, on ImageNet, the largest dataset among 11 datasets, our LOUPE surpasses CLIP by 1.8%.

J Training Efficiency Discussion

Although our proposed Shapley interaction modeling increases the training time per iteration, it enables our model to converge with fewer total iterations by encouraging our model to learn fine-

grained region-phrase alignment beyond coarse image-text alignment. As shown in Table 9, our LOUPE achieves the best performance while using relatively small GPU days (128 GPUs $\times$ 20 days).

Indeed, the proposed Shapley interaction modeling increases the training time per iteration, but it enables our model to learn fine-grained region-phrase alignment from raw image-text pairs without any object-level human annotations. Our LOUPE can be used as a zero-shot object detector without any fine-tuning. Compared with the expensive cost of human annotations, the increased training time might be acceptable. Meanwhile, manual annotations for extremely diverse object categories in the real world are unscalable and even impossible while our model 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. For example, the right case of Figure 4 in the main paper shows that LOUPE successfully recognizes the leash region and aligns it with the "a leash" phrase. Note that the "leash" category has never appeared in any existing object detection datasets.

On the other hand, our method is much more efficient than methods that rely on off-the-shelf object detectors (e.g., Faster R-CNN) to extract visual features. Recent studies [21, 52] have noticed that extracting visual features using object detectors greatly slows down the training (about 20 FPS per GPU) and requires more GPU memory. Thus, our model avoids such a heavy burden while being able to identify semantic-rich visual regions without any pre-training detectors or human annotations.

Methods	**Coarse-grained image-text alignment**	**Fine-grained region-phrase alignment**	**Ways to learn fine-grained region-phrase alignment**
CLIP, ALIGN, DeCLIP	✓	-	-
ImageBERT, UNITER, FILIP, ViLT, ALBEF	✓	✓	Implicit supversion signals from end-to-end training (*e.g.,* Image-Text Contrastive loss)
GLIP, X-VLM, RegionCLIP	✓	✓	Human bounding-box annotations and supervised pre-trained Region Proposal Network
LOUPE	✓	✓	Explicit alignment information quantified by game-theoretic interactions

Figure 10: Comparison of the LOUPE with existing methods.

K Detailed Discussion with Some Related Works

In this section, we first provide comparison table to highlight key differences of our LOUPE with various methods. Then, we provide a detailed discussion with three recent works (*i.e.*, FILIP [52], RegionCLIP [59], X-VLM [56]), which also investigate fine-grained semantic alignment.

We highlight key differences in Figure 10. Our LOUPE differs as it explicitly learns fine-grained region-phrase alignment from the novel perspective of game-theoretic interactions, without resorting to any object-level human annotations and pre-trained Region Proposal Network (RPN). Notably, the human bounding-box annotations are usually limited to the pre-defined object categories, and the RPN can only detect regions belonging to the pre-defined categories of pre-training object detection datasets. Thus, the methods that use human bounding-box annotations or pre-trained RPN usually

suffer from detecting novel objects beyond the pre-defined categories while LOUPE learns from large-scale raw image-text pairs, which are more scalable and contain a broader set of visual concepts.

Compared with FILIP, the superiorities of using Shapley Interaction modeling are mainly three-fold: **1)** We suppose that directly computing token-wise alignment between every patch token and word token is not efficient and meaningful because an individual word token or patch token might not contain complete semantics. A semantic-rich phrase (e.g., "a girl in a blue coat") usually consists of multiple words, and its corresponding visual region is composed of multiple patches. Also, some words (e.g., "is", "the") and patches (e.g., background pixels) are not meaningful. Based on this insight, our LOUPE differs as we first propose token-level Shapley interaction modeling to aggregate patches into semantic-meaningful regions, and then introduce semantics-level Shapley interaction modeling to explicitly model the fine-grained semantic alignment between semantic-meaningful regions and phrases. **2)** Although FILIP computes token-wise similarity to simulate the fine-grained alignment, it can only learn implicit alignment from the supervision of image-text contrastive loss, lacking training signals to explicitly encourage semantic alignment between visual regions and textual phrases. In contrast, our Shapley interaction modeling provides explicit supervision signals (e.g., the alignment matrices visualized in Figure 4) to learn the fine-grained alignment. The consistently superior performance of our LOUPE than FILIP over all metrics also demonstrates the benefit of explicit fine-grained alignment learning. **3)** FILIP can not be directly applied to object detection and visual grounding through implicit token-wise alignment learning while our LOUPE can immediately transfer to these tasks without any fine-tuning. It is because the proposed Shapley interaction modeling enables our model to identify semantic regions and align these regions with language. As shown in Table 2, without any bounding-box annotations and fine-tuning, our LOUPE achieves competitive performance across four object detection and visual grounding benchmarks.

Our LOUPE is different from RegionCLIP in the following aspects: **1)** RegionCLIP uses pre-trained Region Proposal Network (RPN) to detect regions in images. However, RPN is usually pre-trained on pre-defined object categories (e.g., 80 classes for MSCOCO), which can not cover extensive categories of objects in the large-scale pre-training dataset. Furthermore, since the RPN casts excessive demand on memory and computation, existing methods (i.e., RegionCLIP) usually fix the parameters of RPN and regard region detection as pre-processing step, disconnected with vision-language pre-training. Thus, the performance of RegionCLIP is also restricted by the quality of the RPN. In contrast, our LOUPE learns to identify semantic regions of images by token-level Shapley interaction modeling, which is more scalable and enables our LOUPE to learn a broader set of visual concepts from large-scale pre-training dataset. **2)** RegionCLIP constructs a pool of object concepts from image-text corpus and aligns visual regions with these concepts. These concepts are usually individual nouns (e.g., boy, kite, bus). In contrast, our LOUPE focuses on phrases that involve rich context (e.g., "a boy running on the grass"). By aligning visual regions with phrases that contain rich semantic context, our LOUPE can learn a boarder set of visual concepts (e.g., objects, actions, relations) from the large-scale pre-training dataset.

As for X-VLM, the main differences lie in three-fold: **1)** X-VLM is trained on well-annotated datasets, where regions with bounding-box annotations are provided and each of them is associated with a description text. Such a manner is time-consuming and hard to scale to larger raw image-text data from the web. Our LOUPE differs as we are trained on noisy image-text pairs from the Internet. **2)** X-VLM takes ground-truth regions as input and is trained to predict the bounding-box supervised by the regression loss on the ground-truth coordinates. In contrast, our LOUPE learns to identify semantic regions of images without such strong supervision signals from human annotations. **3)** X-VLM has ground-truth alignment information between regions and their corresponding description texts, which provide strong supervision signals for region-text matching. By comparison, our LOUPE learns the fine-grained region-phrase alignment from game-theoretic interactions.