Fine-Grained Semantically Aligned Vision-Language Pre-Training

Anonymous Author(s)
Affiliation
Address
email

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 on image-text retrieval benchmarks. 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.

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 [12, 20, 29, 42] 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 [4, 14, 15, 19, 24, 28, 37, 36] 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 [39]. However, they can only learn implicit alignment by end-to-end training, lacking explicit supervision to encourage semantic 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* [44], *image captioning* [41]), 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

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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* [8, 34]) 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 [26], existing methods mainly employ sampling-based method [3] 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 4.4.

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. Dominant vision-language pre-training models can be categorized into two groups: *dual-encoder* and *fusion-encoder*. The dual-encoder methods [12, 20, 29, 42] 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 [4, 10, 11, 14, 15, 18, 19, 24, 28, 36, 37, 43, 47]

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 [4, 18, 19, 24, 37, 43, 47] attempt to leverage off-the-shelf object detection models to extract object features for pre-training. However, they still fail to tell the correspondence between detected objects and phrases. FILIP [42] uses a token-wise maximum similarity to enhance the cross-modal interaction of dual-encoder methods. However, the complexity grows quadratically with the number of tokens, and it also lacks explicit training signals to guide the fine-grained semantic alignment. To gain fine-grained visual understanding, GLIP [17] utilizes human-annotated visual grounding and object detection datasets and takes them as pre-training tasks. In contrast, our proposed framework explicitly learns the fine-grained semantic alignment between visual regions and textual phrases from raw image-text data and at the same time maintains the high efficiency of dual-encoder.

Shapley Values. The Shapley value [34] 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 [40]. With solid theoretic foundations, Shapley value has recently been studied as post-hoc explanation methods for Deep Neural Networks (DNN) [5, 25, 46]. Lundberg *et al.* [25] propose a unified attribution method based on Shapley value to interpret the predictions of DNN. Ren *et al.* [32] 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 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 $\mathcal{L}_{\mathrm{TSA}}$. 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 $\mathcal{L}_{\mathrm{FSA}}$. Combined with the two newly proposed pre-training losses, the full objective of fine-grained semantically aligned vision-language pre-training can be formulated as:

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

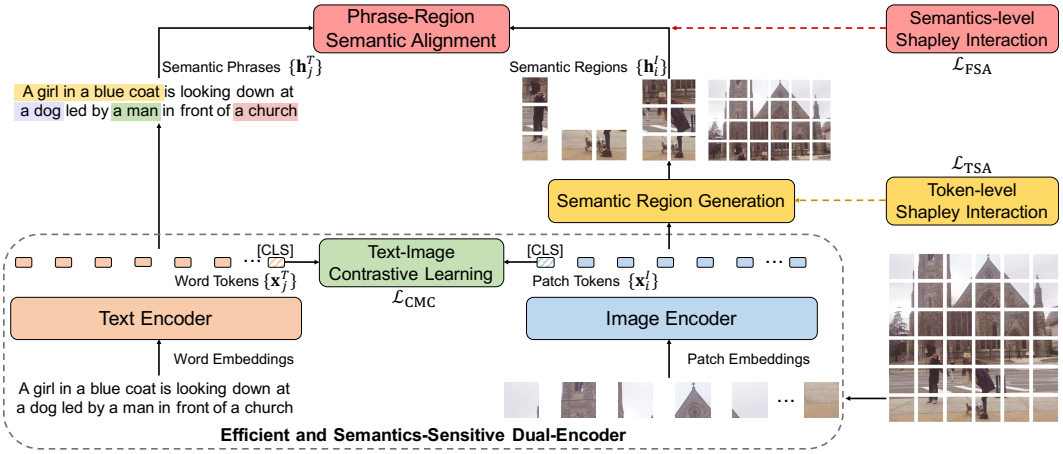


Figure 1: Overview of LOUPE. Our framework serves as a proxy training objective that encourages the image encoder to capture semantic regions and establishes the semantic alignment between region and phrase representations. The proxy training objective can be easily removed for downstream tasks, rendering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

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 [34] 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* [40]. 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 [8] 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* [31], 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 $\mathcal{R}_i = \{\mathbf{x}_{i,k}^I\}_{k=1}^{K}$ with K 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:

$$\begin{aligned}
\mathfrak{I}([\mathcal{H}_{ij}]) &= \phi([\mathcal{H}_{ij}] | \mathcal{H} \setminus \mathcal{H}_{ij} \cup \{[\mathcal{H}_{ij}]\}) - \phi(\mathbf{h}_i^I | \mathcal{H} \setminus \mathcal{H}_{ij} \cup \{\mathbf{h}_i^I\}) - \phi(\mathbf{h}_j^T | \mathcal{H} \setminus \mathcal{H}_{ij} \cup \{\mathbf{h}_j^T\}) \quad (9) \\
&= \mathbb{E}_c \{ \mathbb{E}_{\substack{\mathcal{S} \subseteq \mathcal{H} \setminus \mathcal{H}_{ij} \\ |\mathcal{S}| = c}} [v_2(\mathcal{S} \cup \mathcal{H}_{ij}) - v_2(\mathcal{S} \cup \{\mathbf{h}_i^I\}) - v_2(\mathcal{S} \cup \{\mathbf{h}_j^T\}) + v_2(\mathcal{S})] \} \quad (10)
\end{aligned}$$

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:

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

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 [26]. Previous methods mainly apply sampling-based method [3] 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 [3] estimates the expectation terms in Equation 7 and Equation 10 by sampling to compute the Shapley interaction. Inspired by noisy label learning [13], 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 $\epsilon \leq \sigma$, 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:

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

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 4.4 and Appendix D.

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
FIILIP	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: 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**

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 [22] and the text encoder by BERT-Small [6]. 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 [23] optimizer with a learning rate of 2×10^{-4} and a weight decay of 0.01. More details are provided in Appendix D, E.

4.2 Zero-Shot Image-Text Retrieval

We compare LOUPE on the widely used MSCOCO [21] and Flickr30K [27] 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.

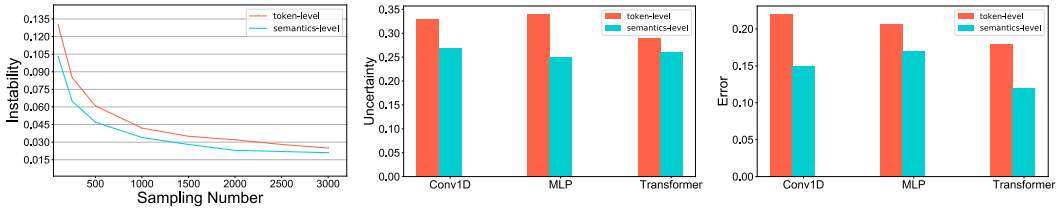
4.3 Zero-Shot Transfer to Object Detection and Visual Grounding

To answer whether our model has learned fine-grained semantics, we further evaluate LOUPE on object detection [33] and visual grounding [44], 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, 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 [16], which is a concurrently unpublished method that leverages classic super-pixel (SLIC [1]) and bounding box proposal (selective search [38]) 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 [21] (65 classes) and PASCAL VOC [7] (20 classes). For visual grounding, we evaluate their top-1 accuracy at IoU thresholds of 0.5 on RefCOCO [44] and RefCOCO+ [44]. The experiment details of CLIP variants and LOUPE are provided in Appendix E.

Table 3: 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 2: (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.

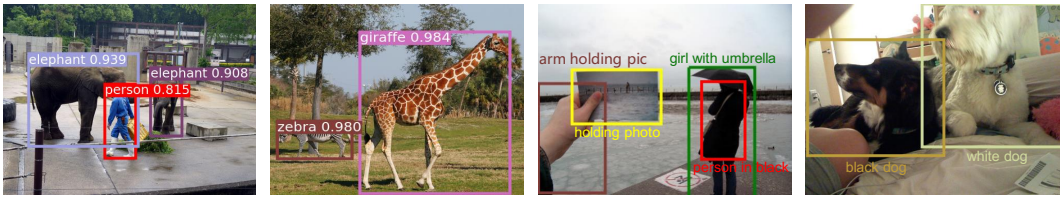
Table 2 summarizes the results. **1)** Overall, LOUPE outperforms all CLIP variants by a large margin. The significantly higher performance illustrates the stronger zero-shot transfer ability of our fine-grained semantically aligned pre-training paradigm. **2)** Second, all CLIP variants rely on pre-processing 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.

As time-consuming human annotations are unscalable for massive object classes in the real world, some recent works [2, 30] 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 [9, 45] 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.

4.4 Ablation Study

Effectiveness of Individual Components. In this section, we investigate the effectiveness of each component in Table 3. Given the costly training time, all ablation studies are based on a relatively small dataset (Conceptual Captions 3M [35]). 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 3 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 proposed 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 [3] 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 [46], 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



(a) Object Detection (b) Visual Grounding

Figure 3: Qualitative examples of object detection on COCO and visual grounding on RefCOCO+.

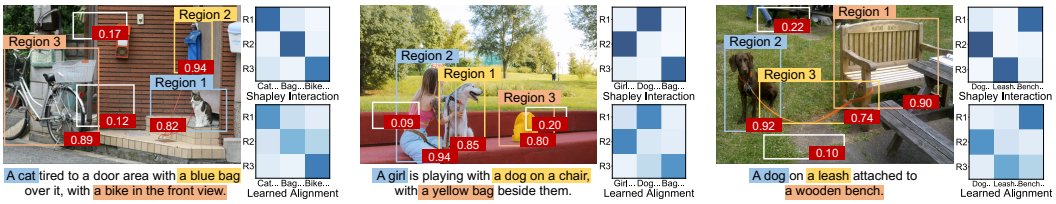


Figure 4: Visualization of learned fine-grained semantic alignment and corresponding Shapley interaction values. The values in the red boxes represent the Shapley interaction of regions.

the average instability values with respect to different sampling numbers. As shown in Figure 2 (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 2 (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*.

4.5 Qualitative Analysis

Qualitative Examples. As shown in Figure 3, 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 4, 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 4, 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.

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] 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.

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

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv e-prints*, pages arXiv–2104, 2021.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[24] 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.

[25] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[26] Yasuko Matsui and Tomomi Matsui. Np-completeness for calculating power indices of weighted majority games. *Theoretical Computer Science*, 263(1-2):305–310, 2001.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv e-prints*, 2018.

[32] 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.

[33] 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.

[34] Lloyd S Shapley. A value for n-person games, contributions to the theory of games, 2, 307–317, 1953.

[35] 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.

[36] 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.

[37] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[38] 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.

[39] 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.

[40] Robert J Weber. Probabilistic values for games. *The Shapley Value. Essays in Honor of Lloyd S. Shapley*, pages 101–119, 1988.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. *arXiv preprint arXiv:2010.05045*, 2020.

[47] 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.

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 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)? [No] Unfortunately, according to the company policy, we can not provide the code in the review stage.
   (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.
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.