Semi-supervised Grounding Alignment for Multi-modal Feature Learning

Shih-Han Chou[1,2], Zicong Fan[3], James J. Little[1], Leonid Sigal[1,2,4]
[1]*Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada*
[2]*Vector Institute for AI, Toronto, Ontario, Canada*
[3]*Department of Computer Science, ETH Zürich, Rämistrasse, Zürich, Switzerland*
[4]*Canada CIFAR AI Chair*
{*shchou75, little, lsigal*}*@cs.ubc.ca; zicong.fan@inf.ethz.ch*

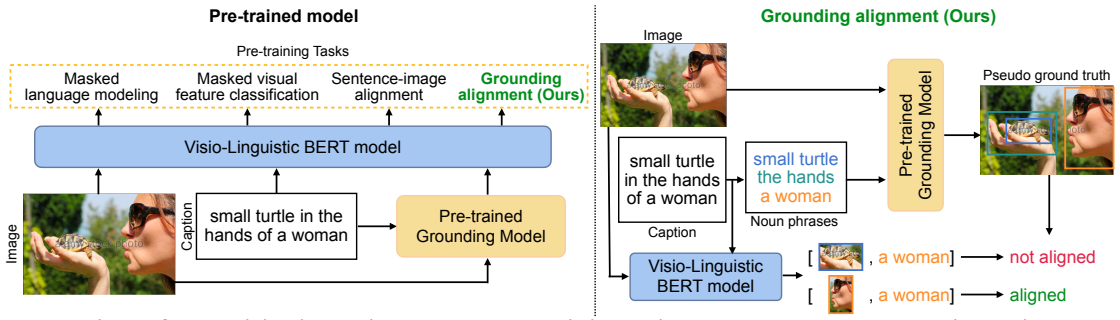

Figure 1: **Illustration of the visio-lingustic BERT pre-training with the proposed grounding alignment.** The left-hand side shows the model architecture and the pre-training tasks. The inputs are the image-caption pairs, and we leverage masked language modeling, masked visual feature classification, sentence-image alignment, and the proposed *semi-supervised grounding alignment* in the pre-training stage. The right-hand side shows the proposed grounding alignment in detail. The noun phrases are first extracted from the caption input. Then a pre-trained grounding model takes the image and noun phrases as inputs to generate the *pseudo* ground truth for the grounding alignment task. In the end, the model predicts the compatibility between the selected image regions and noun phrases, e.g. aligned or not aligned.

***Abstract*—Self-supervised transformer-based architectures, such as ViLBERT [1] and others, have recently emerged as dominant paradigms for multi-modal feature learning. Such architectures leverage large-scale datasets (e.g., Conceptual Captions [2]) and, typically, image-sentence pairings, for self-supervision. However, conventional multi-modal feature learning requires huge datasets and computing for both pre-training and fine-tuning to the target task. In this paper, we illustrate that more granular semi-supervised alignment at a region-phrase level is an additional useful cue and can further improve the performance of such representations. To this end, we propose a novel semi-supervised grounding alignment loss, which leverages an off-the-shelf pre-trained phrase grounding model for pseudo-supervision (by producing region-phrase alignments). This semi-supervised formulation enables better feature learning in the absence of any additional human annotations on the large-scale (Conceptual Captions) dataset. Further, it shows an even larger margin of improvement on smaller data splits, leading to effective data-efficient feature learning. We illustrate the superiority of the learned features by fine-tuning the resulting models to multiple vision-language downstream tasks: visual question answering (VQA), visual commonsense reasoning (VCR), and visual grounding. Experiments on the VQA, VCR, and grounding benchmarks demonstrate the improvement of up to 1.3% in accuracy (in visual grounding) with large-scale training; up to 5.9% (in VQA) with 1/8 of the data for pre-training and fine-tuning[1].**

***Keywords*-grounding; multi-modal feature learning; VQA; VCR;**

[1]We will release the code and all pre-trained models upon acceptance.

I. Introduction

Visio-linguistic tasks (e.g., visual grounding [3], [4], [5], [6], image captioning [7], [2], visual question answering [7], [8], [9], [10], *etc.*) have emerged as important problems in high-level visual understanding. Traditionally, such approaches have leveraged encoder-decoder architectures, where components (e.g., CNN encoders, language encoders, and decoders) have typically been pre-trained on large corpora of unimodal data (e.g., ImageNet classification, language modeling) and then transferred and fine-tuned to the final visio-linguistic target task in question. More recently, however, the focus has shifted to learning joint visio-linguistic representations, by leveraging recent Transfromer[11]-based neural architectures with multi-modal co-attentive mechanisms (e.g., ViLBERT [1], VL-BERT [12], UNITER [13],), that use so-called *proxy* self-supervised objectives for pre-training. As illustrated in Figure 1, such proxy objectives most often include *masked language modeling* [1], [12] and *masked visual classification* [1], [12], as well as, *sentence-image alignment* [1]. The benefit of such models is their ability to learn general, expressive, and implicitly aligned multi-modal representations that, when fine-tuned to the target/downstream task, result in considerably improved performance.

Despite the enormous recent success of these multi-modal BERT-based architectures across a broad range of visio-

linguistic tasks, challenges remain. First, such architectures use coarse, if any, image-sentence alignment proxy objectives [1]. It is reasonable to assume that more granular alignment would be beneficial. This intuition has also been borne out in recent work that presented the 12-in-1 model [14], learned across 12 visio-linguistic datasets and supervised tasks in a multi-task manner. Multi-task learning leverages several datasets to train the model, improving the performance, but relies on the labeled data aggregated across datasets and tasks. This exposes the second key challenge, mainly, that conventional feature learning approaches require huge datasets and substantial computation for both pre-training and fine-tuning to the target task. A data-efficient semi-supervised formulation, that could work well even with limited data, would ultimately be much more desirable and flexible. Our core goal is to address the aforementioned challenges with such an objective.

Specifically, we propose a semi-supervised approach and loss to effectively distill the information from an off-the-shelf, pre-trained, language grounding model to arrive at the improved visio-linguistic representation. Specifically, unlike multi-task learning of [14], we do not assume any additional annotations on the visio-linguistic BERT pre-training dataset, beyond image-sentence pairings (e.g., coming from Conceptual Captions [2] dataset) used in ViLBERT, VL-BERT, and others. However, we do assume access to a pre-trained off-the-shelf grounding model (e.g., a one-stage model in [15]). By parsing sentences in the pre-training dataset into a set of noun phrases and grounding those phrases to regions, we generate a series of granular pseudo-annotations which we can then use as part of the grounding alignment (see Figure 2). We show that these granular pseudo-labels benefit feature learning for a variety of downstream tasks. Moreover, the proposed semi-supervised formulation has a larger margin of improvement on a smaller data split, indicating the efficiency and effectiveness of the proposed approach. Most importantly, we illustrate that the features learned by leveraging the off-the-shelf model, when fine-tuned to the task of grounding itself, improve on both the off-the-shelf and original ViLBERT performance.

The above observation suggests an intriguing potential for self-training, where a model trained for a supervised visio-linguistic task (grounding in this case) can improve feature learning, which in turn improves the supervised (grounding) model performance and so on. We note that the proposed semi-supervised framework is general and neither relies on a specific off-the-shelf grounding model nor on any specific BERT-like feature learning architecture. To the best of our knowledge, this is the first semi-supervised formulation of generic visio-linguistic feature learning. Further, distillation through pseudo-labeling of the form proposed can potentially be used to assimilate knowledge from many *expert* task models without increasing the complexity of the feature learning pipeline or requiring data amassing.

Contributions: Our core contribution is a novel semi-supervised grounding alignment mechanism and loss, which leverages an off-the-shelf pre-trained phrase grounding model for pseudo-supervision by using it to inferring region-phrase alignments. This semi-supervised formulation enables better feature learning in the absence of any additional (human) annotations on the target large-scale (Conceptual Captions [2]) dataset, which would be costly to ascertain. Our formulation is agnostic to both the off-the-shelf grounding model and the core visio-linguistic BERT formulation; we validate this by applying our semi-supervised method atop of ViLBERT [1] and VL-BERT [12]. The resulting semi-supervised variants of these visio-linguistic feature learning techniques achieve improved performance on a wide array of downstream tasks (e.g., grounding, VQA, VCR) when fine-tuned to them in a supervised manner. We conduct the experiments with large-scale training and small splits training. The results show the improvement of our approach is up to 1.3% in accuracy (in visual grounding) with large-scale training and up to 5.9% (in VQA) with $1/8$ of the data for pre-training and fine-tuning.

II. Related Work

Visio-Linguistic Representation Learning: Vision and language representation learning is an active area of research. The majority of recent methods extend the BERT [16] architecture to a multi-modal setting by processing both visual and language inputs [17], [13], [18], [19], [1], [12], [20]. Although the pre-training datasets are different (e.g., Conceptual Captions Dataset [2] vs. MS COCO [21]), the proposed training procedures are similar across architectures. The input tokens are language tokens (words) and image tokens (image region of interest). The pre-training tasks are masked language modeling, masked visual feature modeling, and sentence-image alignment. In [20], they also perform masked visual-feature classification and visual question answering in the pre-training stage.

In detail, ViLBERT [1] and VL-BERT [12] introduce a Co-Transformer layer that refines both modalities jointly via attention. The experiments demonstrate that ViLBERT and VL-BERT outperform single-stream models, which illustrates the importance of joint multi-modal feature learning. In B2T2 [17], the authors propose an early fusion architecture where visual features are embedded on the same level as input word tokens. The experiments show that early integration of the visual features into the text analysis is key to their architecture's effectiveness. Similarly, UNITER [13] proposes an elegant self-attention mechanism designed for learning contextualized representations in order to develop universal image-text representations for visio-linguistic tasks. However, none of these methods model region-phrase alignment. In this paper, we propose a semi-supervised grounding alignment that can benefit feature learning for a variety of downstream tasks.

Semi-Supervised Learning: Semi-supervised learning [22] refers to the class of models that leverage unlabelled data to improve supervised model performance (often trained with limited labeled data). While review of semi-supervised method is beyond the scope of the paper, we want to highlight the use of semi-supervision in language grounding. Specifically, in [5], to solve the problem of only a subset of language annotations and bounding boxes being available, the authors propose a novel semi-supervised approach that learns grounding by reconstructing a given phrase using an attention mechanism. In [23], the authors study the case of objects without labeled queries. They propose a learned location and subject embedding predictors to generate the corresponding language embeddings for objects lacking annotated queries in the training set. With the assistance of the detector, they also apply the predictors to train a grounding model on images without any annotation. In this paper, we formalize a semi-supervised approach that produces pseudo-annotations for region-phrase alignments at a scale (for a large dataset), by employing an off-the-shelf pre-trained grounding model.

Weakly-Supervised Learning: Weakly-supervised learning is used when granularity of the labels doesn't match the task at hand. Existing works [24], [25], [26], [27], [28] leverage image-caption pairs to address the visual grounding task. However, the mapping between inputs and outputs is still necessary. On the other hand, semi-supervised learning is used when only some input and output relations are given. In our work, we leverage semi-supervised learning to distill the information from the off-the-shelf language grounding model during the pre-training stage. This is beneficial as we do not need any additional annotations.

Knowledge Distillation: Knowledge Distillation was first proposed by Hinton *et al.* [29] where knowledge is transferred from a cumbersome model to a small model for efficient deployment. It has been used in various tasks, including model compression, transfer learning, life-long learning and others [30], [31], [32], [33]. In [33], the authors train a student network that is deeper and thinner by using not only the outputs but also the intermediate representations learned by the teacher network as hints. Similarly, [32] introduces multi-step knowledge distillation that employs an intermediate-sized network to bridge the gap between the small and the large network. For lifelong learning, Li *et al.* [31] propose an algorithm and first introduce knowledge distillation to preserve performance on old tasks. Based on [31], Hou *et al.* [30] propose a novel approach trying to seek a better balance between preservation and adaptation by adapting to the new task from an intermediate expert while preserving a small subset of data for old tasks. Similar to knowledge distillation, we leverage a pre-trained grounding network to distill region-to-phrase alignment decisions into visio-linguistic BERT feature learning architecture.

III. Approach

The proposed semi-supervised grounding alignment is an additional pre-training task for *any* existing visio-linguistic BERT-based architecture (see Figure 2 (green)). Importantly, grounding alignment is the task that requires no additional annotations beyond coarse image-sentence pairings. The goal of this task is to predict whether selected image regions are aligned with noun phrases. To train this granular grounding alignment head, the representations from the selected visual and language tokens are used to predict grounding scores. We leverage semi-supervised proxy annotations as a way to generate pseudo ground truth for this task. We overview components of the visio-linguistic feature learning in Section III-A and then formalize our semi-supervised grounding alignment method in Section III-B. Moreover, we explain the spatial positional encoding for image proposals, which we also find to be useful.

A. Visio-Linguistic BERT

Before describing proposed semi-supervised alignment pre-training task, we first overview recent visio-linguistic BERT models atop of which it is designed to be applied. Such architectures, e.g., [1] and [12], abstractly illustrated in Figure 2, share many aspects of architectural design. Specifically, the backbone is a modified multi-layer visio-linguistic bidirectional Transformer encoder taking both visual and language tokens as inputs. Visual tokens are the features of regions-of-interest (RoIs) and language tokens are the encoded words from the captions. During the pre-training stage, the visual tokens are the output bounding boxes of the pre-trained object detector. The typical network is trained with masked visual feature classification, masked language modeling and sentence-image alignment.

Masked language modeling: This pre-trained task [1], [12] is very similar to the masked language modeling in the original BERT model [16]. However, rather than only using the linguistic content, the model also leverages visual clues. In detail, during the pre-training stage, the input word tokens are randomly masked and are replaced by a special token, $\langle \texttt{Mask} \rangle$. Then the model needs to predict the $\langle \texttt{Mask} \rangle$ tokens based on the unmasked words and the visual features. This task is trained with a cross entropy loss, $\mathcal{L}_{word}$.

Masked visual feature classification: Similar to masked language modeling, the visual tokens are randomly masked out [1], [12]. The model's task is to predict the categories of the masked visual tokens by leveraging the features from other unmasked tokens. The ground truth object categories are obtained from the output of the pre-trained object detector. This task is trained with a KL-divergence loss, $\mathcal{L}_{img}$.

Sentence-image alignment: Since the inputs to the model are image-caption pairs, the goal of this pre-trained task is to predict whether the input image and caption are aligned [1], e.g., whether the caption describes the image. To train

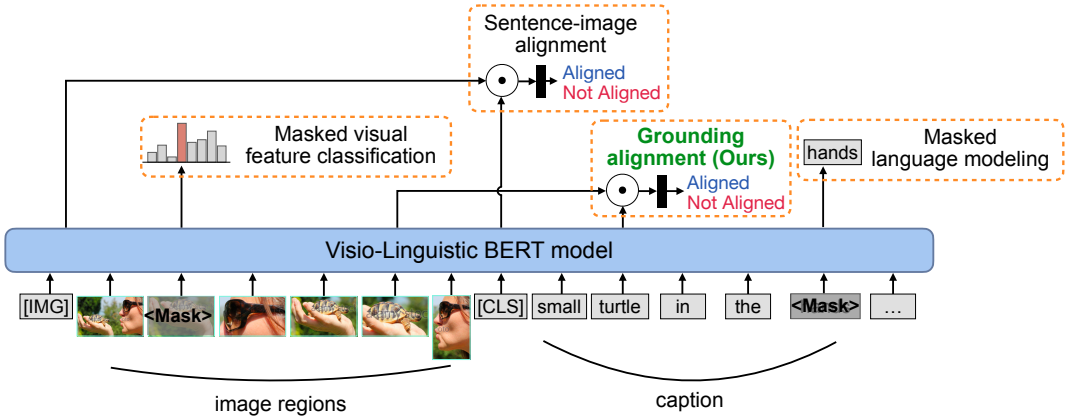

Figure 2: **Pre-training of visio-linguistic BERT with semi-supervised grounding alignment.** The model takes the image-caption pairs as inputs and is trained under four tasks: masked visual feature classification, masked language modeling, sentence-image alignment, and proposed grounding alignment. In masked visual feature classification and masked language modeling, the model needs to reconstruct image region categories or words for masked inputs. In sentence-image alignment, the model must predict whether the caption describes the image regions. In grounding alignment, the model has to predict if the selected image region and phrase are aligned or not.

the model, the holistic representations taken from `[IMG]` and `[CLS]` are used to predict the alignment between the image and caption. We use image-caption pairs from the Conceptual Captions dataset [2] as positive samples and randomly replace the images, or the captions, to form negative samples; training with binary cross entropy loss $\mathcal{L}_{align}$.

B. Grounding Alignment

To have more granular alignment between visual and linguistic domains, we introduce grounding alignment. Given the visual and linguistic sequences V and L, our model will first form the alignment matrix M that represents which noun phrase corresponds to which bounding box(es) (see Figure 3). However, the Conceptual Captions dataset does not have grounding alignment annotations. Motivated by ideas from semi-supervised visual grounding, we leverage the state-of-the-art visual grounding model to generate the *pseudo* ground truth. The Conceptual Captions dataset contains an image with a corresponding caption. We first take the caption C as the input to the noun phrase extractor for which we use the TextBlob library [34]. The outputs will be the set of noun phrases $\{N_p\}$ in the caption. After extracting the noun phrases, images I coupled with corresponding noun phrase N_p are used as the input to the state-of-the-art grounding model [15], the output of which is the bounding box which represents the region that refers to the specified noun phrase:

$$B_{gnd} = f_{gnd}(I, N_p), \tag{1}$$

where f_{gnd} denotes the pre-trained state-of-the-art visual grounding model and B_{gnd} is the output bounding box(es).

Since most visio-linguistic BERT models use pre-extracted image regions, the output bounding boxes from the grounding model might not have exact matches. To address this, we calculate the intersection over union (IoU) of the grounding output and pre-extracted bounding boxes. If the IoU is larger than a threshold[2], we treat these bounding boxes as the matches of the corresponding noun phrase.

To form the alignment matrix M, we initialize a binary matrix of size #word tokens $\times$ #visual tokens. The entries of the alignment matrix represent whether the indexed word corresponds to the indexed image region. For example, if the entry $(2, 3)$ is 1, then the image token 2 corresponds to the word token 3. Otherwise, if the entry is 0, then they are not matched. During training, we take this alignment matrix as the input of the pre-training model. Because the alignment matrix is unbalanced (fewer 1s than 0s in general), we use hierarchical sampling to avoid the model being affected by this bias. In detail, we split the alignment matrix into positive (0's) and negative (1's) set. We then sample a balanced subset of grounding alignments, or miss-alignments, by drawing equal number of positive and negative region and noun phrase pairs respectively.

In grounding alignment, we test two variations. One is token-level grounding, and the other is phrase-level grounding. Token-level grounding means that we only do the word-region match. The *language* representation, in this case, will be the chosen word token representation. On the other hand, the phrase-level grounding means the phrase-region match. To combine the word tokens into a noun phrase representation, we use an LSTM [35] to get the phrase-level language representation. As we show in experiments the latter is considerably better in practice.

Once having the corresponding language representations H_L^* and visual representations H_V^*, we use the grounding

[2]We use an IoU threshold of 0.5, which is motivated by [15].

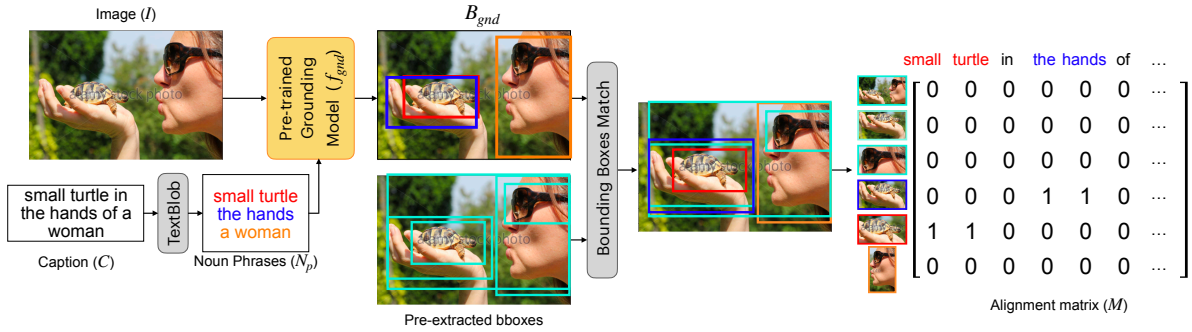

Figure 3: **Grounding Alignment.** The grounding alignment leverages an off-the-shelf pre-trained phrase grounding model to generate pseudo-supervision by producing region-phrase alignment. After having the pseudo ground truth, B_{gnd}, we perform bounding boxes matching to get the alignment matrix M for the grounding alignment pre-training task.

alignment model f_{align} to calculate the score g_{score},

$$g_{score} = f_{align}(H_V^*, H_L^*). \tag{2}$$

The grounding alignment model f_{align} contains a feed-forward network followed by a ReLU layer and a grounding layer. The feed-forward network and the ReLU layer are used to map features to the same dimension. The grounding layer is used to calculate the score of the chosen word/phrase and regions. The score measures the compatibility of the word/phrase and the region. A higher score implies the chosen word/phrase is more likely to belong to the chosen region. We train this alignment model with a binary cross-entropy loss objective, L_{CE}, on the predicted score and the pseudo ground truth labels discussed earlier,

$$\mathcal{L}_{gnd} = L_{CE}(g_{score}, M^*), \tag{3}$$

where M^* denotes the (positive or negative) ground truth sampled from the alignment matrix M.

C. Spatial Positional Encoding

Motivated by the object detection transformer paper [36], we add spatial positional encoding (SPE) in the visual representation. We follow the same setting as described in [36]. In doing so, we use a fixed absolute encoding to represent these spatial positions. Specifically, for both spatial coordinates of each embedding, we independently use sine and cosine functions with different frequencies. We then add them to get the final visual representation.

D. Loss Function

The overall training loss is a combination of the original visio-linguistic BERT model loss and the proposed semi-supervised grounding alignment:

$$\mathcal{L} = \mathcal{L}_{word} + \mathcal{L}_{img} + \lambda_{align}\mathcal{L}_{align} + \lambda_{gnd}\mathcal{L}_{gnd}, \tag{4}$$

where λ_{align} and λ_{gnd} are hyper-parameters which modulate relative importance of the sentence-image alignment and the proposed grounding alignment.

IV. Experiments

Our model learns visio-linguistic representations that can account for more granular alignment between vision and language in a semi-supervised manner. To assess their quality, as common in the literature, we employ and measure performance on a series of proxy tasks (see Fig. 4). We use accuracy as the evaluation metric for these tasks.

Visual Grounding: The referring expression, or grounding, task is designed to localize an image region given a natural language reference (Fig. 4 (a)). For this task, we fine-tune the model on the RefCOCO+ dataset [37]. A common approach to this task is to rank a set of image region proposals given the natural language description. We follow the same settings as ViLBERT [1] using the bounding box proposals provided by [38]. During the fine-tuning stage, we pass the final representation for each image region into a learned linear layer to predict a matching score (Fig. 4 (a)). We set the proposal boxes having 0.5 Intersection over Union (IoU) with the ground truth boxes as true labels. We fine-tune the model with a binary cross-entropy loss for a maximum of 20 epochs and 256 batch size. We use the Adam optimizer with an initial learning rate $4e{-}5$. In inference, we evaluate on the val set and take the highest-scoring region as the prediction.

Visual Question Answering (VQA): In the VQA task, given an image and a question, the model needs to answer the question based on the content of the image. We fine-tune the pre-trained model on the VQA 2.0 dataset [8] which consists of 1.1 million questions on COCO images [21]. Each image contains approximately 10 answers. To fine-tune the model, we add a two-layer MLP on the top of the pre-trained model (see Figure 4 (b)). We feed the element-wise product of the visual and linguistic feature to the two-layer MLP to map the representation to $3,129$ candidate answers. The fine-tuning stage is trained with a multi-label classification loss. A soft target score is assigned to each answer based on the majority of 10 human candidate answers. Then, we use a binary cross-entropy loss on the soft target scores. The model is fine-tuned with 256 batch size

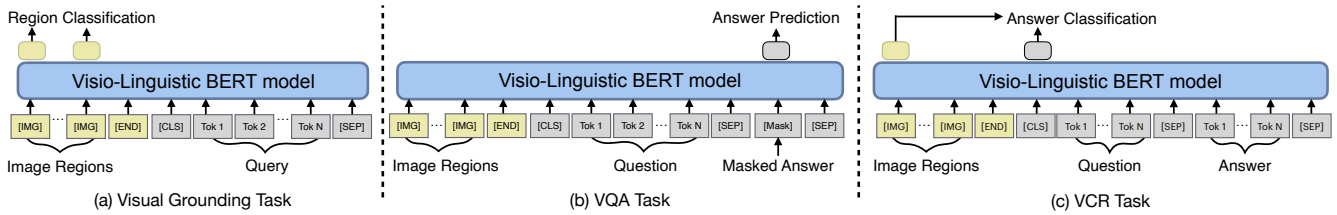

Figure 4: **Fine-tuning stage.** (a) The visual grounding task. The inputs are the concatenation of image regions and the query. (b) The VQA task. To fine-tune to the VQA task, the inputs are the concatenation of image regions, question, and masked answer. (c) VCR task. The inputs of the VCR task are image regions, question, and one of the four candidate answers.

for 20 epochs. The Adam optimizer with an initial learning rate $4e-5$ is used. During inference, the output answer is an $\arg\max$ of the softmax prediction.

Visual Commonsense Reasoning (VCR): The VCR task consists of two problems: visual question answering (Q$\rightarrow$ A) and answer justification (QA$\rightarrow$ R). In the Q$\rightarrow$ A problem, given an image, a question, and a set of answers, the model needs to choose among the answers in a multiple-choice manner. Similarly, in QA$\rightarrow$ R, given an image, a question, and a correct answer, the model has to select the rationale. The QA$\rightarrow$ R problem is also a multiple-choice QA problem. We fine-tune the pre-trained model on the VCR dataset [39] which consists of 290k multiple-choice Q-A pairs and 110k movie scenes. To fine-tune on VCR, we concatenate the question and each answer to form 4 different text inputs as the input to the pre-trained model (see Fig. 4 (c)). A linear layer is added on top of the base pre-trained model, ensuring that the model predicts the score of the QA pair. The final prediction is a softmax score over the 4 QAs. The fine-tuning is achieved with respect to the binary cross-entropy loss with batch size of 64 and for 20 epochs. The Adam optimizer with an initial learning rate $2e-5$ is used.

A. Implementation Details

Training Details: We apply the pre-training stage on the Conceptual Captions Dataset [2]. The Conceptual Captions Dataset contains 3.3 million image-caption pairs. The data from the Conceptual Captions Dataset is automatically scraped from web images. Although noisy, it comes with a huge diversity of visual and language content. Since there are some broken links when downloading the data, we were only able to obtain about 3 million image-caption pairs in total. During pre-training, we conduct all experiments on 8 GPUs (6 GeForce RTX 2080Ti and 2 Quadro RTX 6000) on a single server for the full dataset, which takes approx. 110 hours. Following the settings in [1], the batch size is set to 512 for 10 epochs and the Adam optimizer is used with an initial learning rate $1e-4$. A linear decay learning rate schedule with a warm-up is also used.

B. Model Setup

Given our focus on data efficiency and the desire to see the effect dataset size plays, we experiment with two variants for pre-training: a full Conceptual Captions dataset and a random $1/8$ split of the same dataset for the majority of experiments. Note that while we take $1/8$ of data to do the pre-training, we use the whole RefCOCO+, VQA, and VCR datasets for fine-tuning. Later, in Table III, we experiment with even smaller splits for pre-training and fine-tuning.

The goal of our work is to illustrate the effectiveness of semi-supervised learning through the proposed grounding alignment loss. Hence, we apply our semi-supervised alignment on top of the two common and simple architectures: ViLBERT [1] and VL-BERT [12]. However, the proposed loss and methodology can be easily added to any vision-language pre-training pipeline as long as the pre-training dataset contains image-caption pairs.

Baselines: Since our model is a direct extension of the ViLBERT [1] and the VL-BERT [12] models, we use them as our natural baselines.

Ablation Study: We consider the effects our design choices make on performance in Table II. Specifically, we consider incrementally adding one component at a time to the baseline ViLBERT [1] model trained on 1/8 of the Conceptual Captions dataset. This results in the following variants:

[+ SPE]: ViLBERT [1] baseline model with added Spatial Positional Encoding described in Section III-C.
[+ SS Ground (token)]: ViLBERT [1] with SPE and our semi-supervised grounding alignment implemented at the token level (see Section III-B).
[+ SS Ground (phrase)]: Our final model, consisting of ViLBERT [1] with SPE and our semi-supervised grounding alignment implemented at the phrase level.

The hyper-parameters of $\lambda_{align} = 1$ and $\lambda_{ground} = 20$ in Rows 3 and 4 were chosen through cross-validation.

C. Results and Discussion

Effectiveness of grounding alignment: As shown in Table I, adding semi-supervised grounding (at the phrase level) consistently shows improved performance against the baselines in all settings and for all downstream tasks.

Effect of dataset size: Comparing the models pre-trained on different sizes of Conceptual Caption Dataset ($1/8$ and full dataset), semi-supervised grounding appears to be more

Table I: **Quantitative Results.** We evaluate the proposed semi-supervised grounding alignment loss on two architectures: ViLBERT [1] and VL-BERT [12]. Our models include SPE and implement semi-supervised grounding at the phrase level. Training with two dataset sizes is explored (full and 1/8 of Conceptual Captions); best results for given dataset size are bold.

	Method	Visual Grounding	VQA	VCR (Q→A)	VCR (QA→R)
	One-Stage Visual Grounding [15]	72.05	-	-	-
	12-in-1 Multi-task ViLBERT [14] (GT $\xrightarrow{\text{finetune}}$ ST)	72.12	-	-	-
1/8 dataset	VL-BERT [12]	66.35	69.78	70.11	70.32
	VL-BERT [12] + Semi-supervised Grounding (our)	67.17	**70.09**	70.51	70.75
	ViLBERT [1]	70.92	67.85	70.83	72.47
	ViLBERT [1] + Semi-supervised Grounding (our)	**72.23**	68.98	**71.88**	**73.62**
full dataset	VL-BERT [12]	68.90	70.86	71.51	72.92
	VL-BERT [12] + Semi-supervised Grounding (our)	69.36	**70.89**	72.02	73.59
	ViLBERT [1]	72.22	69.17	72.15	73.61
	ViLBERT [1] + Semi-supervised Grounding (our)	**72.47**	69.63	**72.49**	**73.73**

Table II: **Ablation Study.** We conduct ablations using $1/8$ amount of the Conceptual Caption Dataset. The best result of each column is marked by the bold black color. SPE denotes spatial positional encoding and SS Ground represents semi-supervised grounding alignment. Hyper-parameters for λ_{align} and λ_{gnd} were chosen through cross-validation.

Settings	SPE	λ_{align}	λ_{gnd}	token	phrase	Visual Grounding	VQA	VCR (Q→A)	VCR (QA→R)
ViLBERT [1]	-	1	-	-	-	70.92	67.85	70.83	72.47
+ SPE	✓	1	0			71.19	68.12	71.67	73.58
+ SS Ground (token)	✓	1	20	✓		69.02	67.64	71.72	73.06
+ SS Ground (phrase)	✓	1	20		✓	**72.23**	**68.98**	**71.88**	**73.62**

Table III: **Efficiency of the Proposed Semi-supervised Grounding Alignment.** We evaluate the proposed semi-supervised grounding alignment loss with different amounts of pre-training (vertical) and fine-tuning (horizontal) data. With less data the margin of improvements of our model over the baseline ViLBERT [1] increases.

			Visual Grounding			**VQA**			**VCR: Q → A**			**VCR: QA → R**			**Avg.**
			Full	1/4	1/8	Full	1/4	1/8	Full	1/4	1/8	Full	1/4	1/8	
Conceptual Captions	Full	Baseline	72.22	71.10	69.77	69.17	68.15	60.41	72.15	72.02	70.24	73.61	73.5	71.29	70.30
		Ours	72.47	71.72	70.57	69.63	68.67	60.97	72.49	72.22	70.90	73.73	73.57	71.79	70.73
			(+0.25)	(+0.62)	(+0.80)	(+0.46)	(+0.52)	(+0.56)	(+0.34)	(+0.20)	(+0.66)	(+0.12)	(+0.07)	(+0.50)	(+0.43)
	1/8	Baseline	70.92	70.00	67.42	67.85	66.13	55.67	70.83	69.15	68.71	72.47	69.91	69.85	68.24
		Ours	72.23	71.51	70.46	68.98	68.02	61.61	71.88	71.11	70.02	73.62	72.95	71.22	70.30
			(+1.31)	**(+1.51)**	**(+3.04)**	**(+1.13)**	**(+1.89)**	**(+5.94)**	**(+1.05)**	**(+1.96)**	**(+1.31)**	**(+1.15)**	**(+3.04)**	**(+1.37)**	**(+2.06)**
	1/16	Baseline	69.9	69.28	65.88	67.21	64.39	54.66	70.36	69.26	66.55	72.28	70.32	67.57	67.26
		Ours	70.71	69.64	66.81	68.48	65.82	55.45	70.97	70.83	68.82	72.69	72.36	69.34	68.49
			(+0.81)	(+0.84)	(+0.93)	**(+1.27)**	**(+1.43)**	(+0.79)	(+0.61)	**(+1.57)**	**(+2.27)**	(+0.41)	**(+2.04)**	**(+1.77)**	**(+1.27)**

effective for a smaller dataset. In other words, the proposed approach allows competitive performance that is data-efficient (requires a fraction of the original dataset). Part of the reason is a trade-off between the size/quality of supervised and unsupervised pseudo-annotated data inherent in all semi-supervised methods. To study the effectiveness of the proposed semi-supervised grounding alignment with limited data, we conduct additional experiments in Table III. We do so with the ViLBERT model as the baseline and add the semi-supervised grounding alignment to it for "Our" model. We experiment with (full, $1/8$ and $1/16$) fractions of Conceptual Captions dataset for pre-training and (full, $1/4$ and $1/8$) fractions of target datasets for fine-tuning. Notably, the margin of improvement our model has, over the baseline, increases as the amount of training data decreases for both pre-training and fine-tuning. Using $1/8$ of the Conceptual Captions dataset for pre-training leads to the largest improvement of 2.06%, on average, and up to 5.94% with $1/8$ VQA dataset fine-tuning. The improvement with $1/4$ of the VCR dataset is also sizable at 3.04%. Extensive experiments in Table III illustrate that the proposed semi-supervised alignment is not only effective but also highly data-efficient.

Comparison to multi-task learning: Results in Table I are suggestive that semi-supervision through distillation might be more effective than multi-task learning; this observation

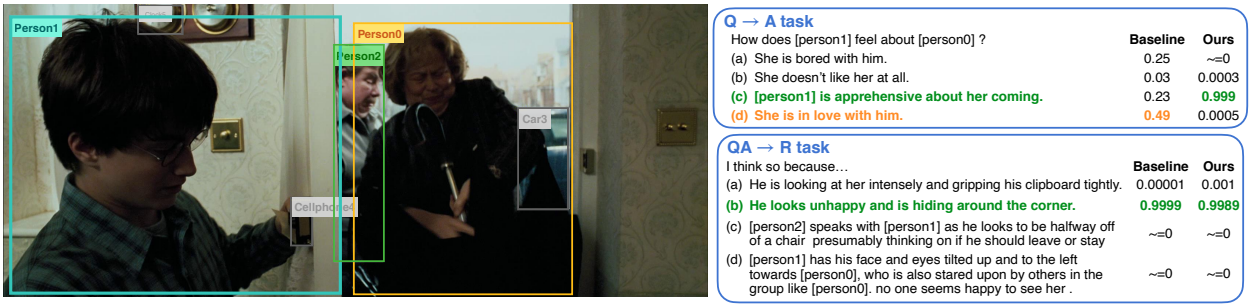

Figure 5: **Qualitative results**. Qualitative results for VCR task. The numbers represent the probabilities across the answers. During inference, the answer with the highest score is chosen. The green color represents the correct answers and the orange color denotes the wrong answer. More examples can be found in the supplementary material.

is consistent with other distillation literature [29]. In the context of our experiments, the evidence is twofold. First, both ours and [14] utilize the same grounding data for pre-training and fine-tuning the visual grounding task. The results show that with the same amount of data, our method performs better. Second, we compare to 72.12 from [14] because this is the most similar setting with ours, which is pre-trained using 5 visual grounding datasets [37], [40], [41], [42] in a multi-task manner and then fine-tuned to RefCOCO+ dataset. The result demonstrates that our method is still better. We note that this improved performance holds even when we pre-train on $1/8$ of the Conceptual Captions.

Ablation observations: In Table II, we conduct the ablation study of the model components. Experiments show that adding SPE improves performance and further addition of semi-supervised grounding, at the phrase level, achieves the best results. This, along with the results in Table I clearly indicates that the grounding alignment helps the model learn better visio-linguistic representations. Comparing two variants of the semi-supervised grounding alignment, at the `token` and at the `phrase` level, we see a clear superiority of the latter. In other words, the form of semi-supervised alignment task and loss plays an important role. Furthermore, we also show the quantitative results of the models w/o SPE in the Suppl., which show the same trend.

Generality: We provide results with two visio-linguistic backbones, ViLBERT [1] and VL-BERT [12], to demonstrate the applicability of the proposed grounding alignment. Table I summarizes the results of baseline models and our method. In all cases, we use our final model which includes SPE and semi-supervised grounding at the phrase level. We observe a similar trend as in Table II: adding grounding alignment outperforms the baselines. Another point worth noting is that our method outperforms the one-stage visual grounding model [15]. Therefore, the features learned by leveraging the off-the-shelf model, when fine-tuned to the task of grounding itself, improve on both the off-the-shelf and the original baseline models' performance.

Qualitative Results: We show the qualitative result of the VCR task in Figure 5. The baseline [1] predictions and ours as well as the final prediction probabilities are listed. The green color of the answers denotes the ground truth answer. The green color of numbers indicates the predicted answer is correct and the orange color is the wrong prediction. We can see that the baseline and ours predict the correct answer in most cases. However, in the Q $\rightarrow$ A example, our method predicts the correct answer while the baseline fails. This shows that our method can focus more on the correct answers. Please see the supplementary material for more qualitative results for different tasks.

Limitations and future directions: One limitation of the proposed approach is that since the pseudo-supervision is pre-extracted from the off-the-shelf region-phrase grounding model, we cannot do end-to-end training. Although the outputs of the off-the-shelf region-phrase grounding model are reasonable and improve feature learning, we believe joint training of the Visio-Linguistic BERT model and the off-the-shelf model might boost performance. Thus, to advance grounding alignment, we suggest developing a model that can combine the two models and train in an end-to-end manner. Moreover, we also suggest using other supervision to facilitate feature learning. For example, we can use scene graph generation or human pose estimation as a large portion of the dataset contains humans and objects.

V. Conclusion

In this paper, we propose a novel semi-supervised grounding alignment mechanism and loss, which leverages an off-the-shelf pre-trained phrase grounding model to generate pseudo grounding truths. This formulation enables better feature learning on the large-scale dataset without any additional human annotations and illustrates that more granular semi-supervised alignment at a region-phrase level is useful. The proposed grounding alignment shows the effectiveness of the learned features by fine-tuning the visio-linguistic BERT models to multiple downstream vision-language tasks. Experiments manifest the improvements in the visual grounding, VQA, and VCR benchmarks.

REFERENCES

[1] Lu, Jiasen and Batra, Dhruv and Parikh, Devi and Lee, Stefan, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019. 1, 2, 3, 5, 6, 7, 8

[2] Sharma, Piyush and Ding, Nan and Goodman, Sebastian and Soricut, Radu, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018. 1, 2, 4, 6

[3] Bajaj, Mohit and Wang, Lanjun and Sigal, Leonid, "G3raphground: Graph-based language grounding," in *CVPR*, 2019. 1

[4] Deng, Chaorui and Wu, Qi and Wu, Qingyao and Hu, Fuyuan and Lyu, Fan and Tan, Mingkui, "Visual grounding via accumulated attention," in *CVPR*, 2018. 1

[5] Rohrbach, Anna and Rohrbach, Marcus and Hu, Ronghang and Darrell, Trevor and Schiele, Bernt, "Grounding of textual phrases in images by reconstruction," in *ECCV*, 2016. 1, 3

[6] Xiao, Fanyi and Sigal, Leonid and Jae Lee, Yong, "Weakly-supervised visual grounding of phrases with linguistic structures," in *CVPR*, 2017. 1

[7] Anderson, Peter and He, Xiaodong and Buehler, Chris and Teney, Damien and Johnson, Mark and Gould, Stephen and Zhang, Lei, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018. 1

[8] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015. 1, 5

[9] Ben-Younes, Hedi and Cadene, Rémi and Cord, Matthieu and Thome, Nicolas, "Mutan: Multimodal Tucker fusion for visual question answering," in *ICCV*, 2017. 1

[10] Goyal, Yash and Khot, Tejas and Summers-Stay, Douglas and Batra, Dhruv and Parikh, Devi, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017. 1

[11] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia, "Attention is all you need," in *NeurIPS*, 2017. 1

[12] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," in *ICLR*, 2020. 1, 2, 3, 6, 7, 8

[13] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020. 1, 2

[14] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *CVPR*, 2020. 2, 7, 8

[15] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *ICCV*, 2019. 2, 4, 7, 8

[16] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[17] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," *arXiv preprint arXiv:1908.05054*, 2019. 2

[18] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training." in *AAAI*, 2020. 2

[19] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019. 2

[20] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *EMNLP*, 2019. 2

[21] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015. 2, 5

[22] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, 2020. 3

[23] H. Zhu, A. Sadhu, Z. Zheng, and R. Nevatia, "Utilizing every image object for semi-supervised phrase grounding," in *WACV*, 2021. 3

[24] Q. Wang, H. Tan, S. Shen, M. W. Mahoney, and Z. Yao, "Maf: Multimodal alignment framework for weakly-supervised phrase grounding," *arXiv preprint arXiv:2010.05379*, 2020. 3

[25] Y. Liu, B. Wan, L. Ma, and X. He, "Relation-aware instance refinement for weakly supervised visual grounding," in *CVPR*, 2021. 3

[26] A. Arbelle, S. Doveh, A. Alfassy, J. Shtok, G. Lev, E. Schwartz, H. Kuehne, H. B. Levi, P. Sattigeri, R. Panda *et al.*, "Detector-free weakly supervised grounding by separation," in *ICCV*, 2021. 3

[27] T. Gupta, A. Vahdat, G. Chechik, X. Yang, J. Kautz, and D. Hoiem, "Contrastive learning for weakly supervised phrase grounding," in *ECCV*. Springer, 2020. 3

[28] S. Datta, K. Sikka, A. Roy, K. Ahuja, D. Parikh, and A. Divakaran, "Align2ground: Weakly supervised phrase grounding guided by image-caption alignment," in *ICCV*, 2019. 3

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. 3, 8

[30] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Lifelong learning via progressive distillation and retrospection," in *ECCV*, 2018. 3

[31] Z. Li and D. Hoiem, "Learning without forgetting," *PAMI*, 2017. 3

[32] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *AAAI*, 2020. 3

[33] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014. 3

[34] S. Loria, "textblob documentation," *Release 0.15*, 2018. 4

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997. 4

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020. 5

[37] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *EMNLP*, 2014. 5, 8

[38] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *CVPR*, 2018. 5

[39] Zellers, Rowan and Bisk, Yonatan and Farhadi, Ali and Choi, Yejin, "From recognition to cognition: Visual commonsense reasoning," in *CVPR*, 2019. 6

[40] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016. 8

[41] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *CVPR*, 2016. 8

[42] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "Guesswhat?! visual object discovery through multi-modal dialogue," in *CVPR*, 2017. 8