Technical Summary of Semi-supervised Grounding Alignment for Multi-modal Feature Learning

Part I: Problem Formulation, Methods, and Evidence

Kai-Yu Lu

2025/09/25

1 Research Problem and Motivation

Problem. The paper addresses data-efficient vision–language representation learning within Transformer-based visio-linguistic encoders such as ViLBERT and VL-BERT. Existing pre-training uses coarse image–sentence objectives that underutilize fine-grained region–phrase correspondences. The central idea is to add a semi-supervised grounding alignment objective at region–phrase level without requiring any extra human annotations on large web-scale pre-training corpora. This is achieved by distilling pseudo labels from an off-the-shelf phrase grounding model to guide granular alignment during pre-training. Performance is evaluated on visual grounding, VQA, and VCR.

Motivation. Coarse sentence–image alignment may fail to learn explicit cross-modal alignment, and collecting region–phrase labels at scale is expensive. Semi-supervised pseudo-labeling enables data efficiency and better inductive bias toward grounding.

Claim. The method improves downstream accuracy and the gains are larger in low-data regimes.

2 Related Work

Visio-linguistic pre-training. ViLBERT, VL-BERT, UNITER and related models extend BERT to joint visual and textual tokens using masked language modeling, masked visual feature classification, and sentence–image alignment.

Semi-supervised and weakly supervised grounding. Prior work leverages incomplete labels or image–caption pairs for grounding.

Knowledge distillation. Transferring knowledge from a teacher model to a student is well-studied. The present paper distills region–phrase alignment decisions from a pre-trained grounding model into a BERT-style encoder during pre-training to improve representations.

3 Dataset Construction

Pre-training. Conceptual Captions (approximately 3.0M usable image–caption pairs obtained from a nominal 3.3M set) scraped from the web. Two settings are used for ablation: full dataset and a random one-eighth split.

Fine-tuning and evaluation sets. RefCOCO+ for visual grounding, VQA 2.0 for question answering with 1.1M questions on COCO images, and VCR for visual commonsense reasoning with approximately 290k multiple-choice Q–A pairs and approximately 110k movie scenes.

Hardware and training schedule. Pre-training on 8 GPUs for approximately 110 hours with batch size 512 for 10 epochs and Adam with initial learning rate 1×10^{-4} and linear warmup–decay. Fine-tuning settings are task-specific as detailed in Section 4.

4 Query Protocol and Task Definitions

Pre-training tasks.

- *Masked language modeling* (MLM): predict masked word tokens conditioned on unmasked words and visual regions using cross-entropy loss L_{word}.
- *Masked visual feature classification*: predict categories for masked visual tokens using KL-divergence loss L_{img} with detector-provided labels.
- *Sentence–image alignment*: binary prediction of whether a caption matches an image using holistic [CLS] and [IMG] representations with binary cross-entropy loss L_{align}.
- *Grounding alignment* (proposed): binary prediction of whether a selected region aligns with a noun phrase using pseudo supervision from a grounding teacher.

Fine-tuning tasks.

- *Visual grounding on RefCOCO+*: rank region proposals and count a hit if top prediction has IoU at least 0.5 with ground truth. Metric is accuracy.
- *VQA 2.0*: multi-label classification over 3,129 candidate answers using a two-layer MLP on the element-wise product of fused features. Soft targets come from 10 human answers. Metric is accuracy.
- *VCR*: two multiple-choice subtasks Q$\rightarrow$A and QA$\rightarrow$R using a linear head over fused features. Metric is accuracy.

5 Modeling Approach

Backbone and tokens

Backbone. The method augments standard visio-linguistic BERT encoders (ViLBERT and VL-BERT). Visual tokens are region-of-interest features from a pre-trained detector. Text tokens are wordpiece embeddings.

Spatial Positional Encoding (SPE) adds fixed sine–cosine encodings to visual coordinates to inject geometry.

Semi-supervised Grounding Alignment

Terminology. Let I be an input image. Let C be its caption. Noun phrases $\{N_p\}$ are extracted from C by a phrase extractor. Let f_{gnd} denote an off-the-shelf phrase grounding model that outputs bounding boxes likely to correspond to a phrase. Let M denote a binary alignment matrix between language and visual tokens. Let H_V^* and H_L^* denote selected visual and language representations for a candidate region–phrase pair.

Pseudo-label generation.

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

Explanation. B_{gnd} is the teacher-predicted bounding box set for phrase N_p in image I. f_{gnd} is any pre-trained grounding model. Teacher boxes are matched to detector proposals by Intersection over Union with a threshold of 0.5. The matches populate the alignment matrix M whose entries indicate whether a word or phrase aligns with a region. Due to imbalance, hierarchical sampling balances positive and negative pairs for training.

Phrase representation and scoring. Phrases are encoded either at token level or phrase level. Phrase level concatenates word tokens through an LSTM to obtain H_L^* which performed better empirically.

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

Explanation. f_{align} is a feed-forward projector with ReLU followed by a grounding layer that outputs a compatibility score between the selected visual feature H_V^* and phrase feature H_L^*. A larger g_{score} indicates higher likelihood of alignment.

Grounding alignment loss.

$$L_{\text{gnd}} \;=\; L_{\text{CE}}(g_{\text{score}},\, M^*)\,. \tag{3}$$

Explanation. L_{CE} is the binary cross-entropy between the predicted score g_{score} and the sampled pseudo label $M^* \in \{0,1\}$ from the alignment matrix for the candidate region–phrase pair.

Total pre-training objective.

$$L \;=\; L_{\text{word}} \;+\; L_{\text{img}} \;+\; \lambda_{\text{align}}\, L_{\text{align}} \;+\; \lambda_{\text{gnd}}\, L_{\text{gnd}}. \tag{4}$$

Explanation. L_{word} is cross-entropy for masked word prediction. L_{img} is KL-divergence for masked visual category prediction. L_{align} is binary cross-entropy for sentence–image alignment. λ_{align} and λ_{gnd} are scalar weights tuned by cross-validation. This combination encourages both coarse image–caption consistency and granular region–phrase alignment during pre-training.

6 Empirical Results

Overall effectiveness

Full-data regime. Adding phrase-level grounding alignment with SPE improves over ViLBERT and VL-BERT baselines on all tasks. For example, with ViLBERT pre-trained on the full Conceptual Captions set, visual grounding improves from 72.22 to 72.47 accuracy, VQA improves from 69.17 to 69.63, VCR Q→A improves from 72.15 to 72.49, and VCR QA→R improves from 73.61 to 73.73.

Low-data regime. With one-eighth pre-training data, gains are larger. ViLBERT improves from 70.92 to 72.23 on visual grounding, from 67.85 to 68.98 on VQA, from 70.83 to 71.88 on VCR Q$\rightarrow$A, and from 72.47 to 73.62 on VCR QA$\rightarrow$R.

Data efficiency

When varying both pre-training and fine-tuning data, the average improvement peaks with one-eighth pre-training data. The average gain is approximately 2.06 points across tasks, and the largest per-cell improvement reaches approximately 5.94 points on VQA when fine-tuning with one-eighth of the VQA data. This shows the pseudo-label signal is especially valuable when human supervision is scarce.

Ablation findings

Spatial Positional Encoding consistently helps geometry-aware alignment.

Phrase-level grounding outperforms token-level grounding, which indicates that phrase composition is important for alignment learning. Hyperparameters $\lambda_{\text{align}}=1$ and $\lambda_{\text{gnd}}=20$ are selected via cross-validation in one-eighth data ablations.

Comparison to multi-task learning

Compared with a multi-task ViLBERT trained across numerous labeled grounding datasets, the proposed semi-supervised alignment yields higher visual grounding accuracy despite using the same or less supervision. This suggests distillation-style pseudo-labeling can be a competitive alternative to multi-task supervision while retaining simplicity in the pre-training pipeline.

7 Summary

Contributions. A general semi-supervised grounding alignment objective plugs into visio-linguistic BERT encoders, uses an external grounding teacher to create pseudo region–phrase labels, and improves downstream tasks with notable gains in low-data regimes.

Limitations. The pipeline relies on pre-extracted pseudo labels and cannot perform joint end-to-end training of the teacher and student.

Future work. Joint optimization with the grounding teacher, and incorporation of other structured signals such as scene graphs or human pose, may further improve alignment learning.

Technical Glossary and Settings

ViLBERT. Dual-stream co-attentional Transformer for vision and language.

VL-BERT. Single-stream early fusion of visual features with text tokens inside BERT layers.

UNITER. Unified pre-training for universal image–text representations.

RefCOCO+. Referring expression comprehension benchmark that restricts absolute location words to encourage fine-grained grounding.

VCR. Visual Commonsense Reasoning with multiple-choice QA and rationale selection.

Hyperparameters. Pre-training learning rate 1×10^{-4} with linear warmup and decay, batch size 512 for 10 epochs. Fine-tuning uses Adam with task-specific learning rates: 4×10^{-5} for grounding and VQA, 2×10^{-5} for VCR, batch sizes 256 for grounding and VQA, and 64 for VCR.

Evaluation metric. Accuracy for all tasks with IoU threshold 0.5 for grounding hits.