Semantic VL-BERT: Visual Grounding via Attribute Learning

Prashan Wanigasekara[§]
Alexa AI-NU
Cambridge, USA
wprasha@amazon.com

Kechen Qin[§]
Alexa AI-NU
Cambridge, USA
qinkeche@amazon.com

Emre Barut
Alexa AI-NU
Cambridge, USA
ebarut@amazon.com

Fan Yang
Alexa AI-NU
Cambridge, USA
fyaamz@amazon.com

Weitong Ruan
Alexa AI-NU
Cambridge, USA
weiton@amazon.com

Chengwei Su
Alexa AI-NU
Cambridge, USA
chengwes@amazon.com

***Abstract*—In recent years, Smart Home Assistants have expanded into tens of thousands of devices and transformed from a voice only assistant to a much more versatile smart assistant, that uses a connected display to provide a multi-modal customer experience. In order to further improve on the multi-modality experience, comprehension systems need models that can work with multisensory inputs. We focus on the problem of visual grounding, which allows customers to interact with and manipulate items displayed on a screen via voice. We propose a novel learning approach that improves upon a lightweight single stream transformer architecture by adjusting it to better align the visual input features with the referring expressions. Our approach learns to cluster parts of the image along spatial and channel dimensions based on descriptive attributes in the query, and takes advantage of the information in separate clusters more efficiently, as demonstrated by a 1.32% absolute accuracy improvement on a public dataset over the baseline. Given that modern-day Smart Home Assistants have very stringent memory and latency requirements, we restrict our focus to a family of lightweight single stream transformer architectures – our focus is not to beat the ever improving state-of-the-art in visual grounding but to improve upon a lightweight transformer architecture which leads to a model that is easy to train and deploy while having improved semantic awareness.**

***Index Terms*—vision, language, multi-modality, single stream transformer**

I. Introduction

In visual grounding, a model is tasked to identify an object in an image based on a natural language query that describes the object of interest. It is one of the fundamental problems in multi-modality and embodied AI, as it can be utilized in numerous use cases in practice, including item selection and voice-based navigation. Some of the current generic architectures for visual grounding are shown in Fig. 1. In Fig. 1a) and b) the image is fed through an object detector (e.g., Faster RCNN, YOLO [1, 2]), which identifies the different objects, and produces region features for each such object. The text query is run through a separate language model (e.g., a bi-LSTM, BERT), and then these two embeddings are fused in a multi-modality network. Figure 1a) uses a single transformer for both images and text, while Fig. 1b) uses modality dependent transformers first followed by a cross modal fusion transformer. In contrast, Fig. 1c) shows an end-to-end visual grounding architecture. There, image features are obtained via a vanilla Convolutional Neural Network (CNN). Then, the image and text embeddings are jointly trained in an end-to-end fashion using an encoder-decoder transformer setup. In all of these architectures, the multi-modal neural network produces the final classification score that represents the match between a specific object in the image and the input query. Recent work [3, 4, 5, 6, 7, 8, 9, 10, 11] has shown that integrating pre-trained embedding layers, custom loss functions and downstream fine-tuning tasks into this general framework can significantly improve its performance.

Another architecture that was used prior to transformers was Modular Attention Networks(MattNet) [12]. In contrast to the transformer-based models that use a single deep structure to capture multi-modal representations, Modular Networks use multiple hand-crafted modules that explicitly capture various aspects of the input like location, object attributes, and relationship information. Hand-crafted modules tend to capture attribute references better and have the added advantage of interpretability. However, using hand-crafted modules in practice is not scalable, since visual grounding use cases evolve in complexity at a rapid pace as user interactions become more nuanced over time. Each such evolution would either require a manual tweak to an existing module or adding a new handcrafted module.

In our study, we learn object attributes similar to MattNet [12] but in an unsupervised manner, coupled with the benefits of a simple single stream transformer architecture similar to VL-BERT [3]. In a practical visual grounding setting, most users utilize a set of specific referring attributes (e.g., color, size, shape, location etc.) which need to be matched with a displayed image with high accuracy in real time. In such a setting, focusing on object attribute semantics together with visual grounding is uniquely beneficial. Multi-modal

[§] Equal contribution

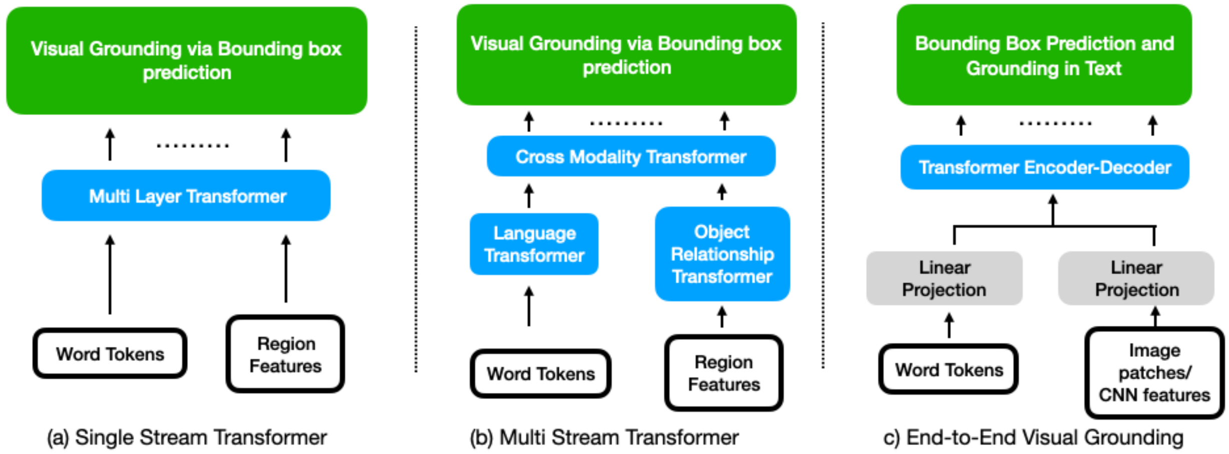

Visual Grounding via Bounding box prediction
Multi Layer Transformer
Word Tokens
Region Features
(a) Single Stream Transformer
Visual Grounding via Bounding box prediction
Cross Modality Transformer
Language Transformer
Object Relationship Transformer
Word Tokens
Region Features
(b) Multi Stream Transformer
Bounding Box Prediction and Grounding in Text
Transformer Encoder-Decoder
Linear Projection
Linear Projection
Word Tokens
Image patches/ CNN features
c) End-to-End Visual Grounding

Fig. 1: Commonly used transformer architectures for visual grounding. Single (a), dual (b) and (c) end-to-end encoder-decoder architectures are presented.

transformer architectures like VL-BERT [3] strive to learn semantics via various pre-training tasks (e.g., masked language modelling with visual cues, masked region-of-interest classification with linguistic cues) and an additional fine-tuning task. In our work shown in Fig. 2, we propose an extension to the single stream architecture that incorporates learning of object attribute semantics in the training stage without the additional burden of the time-consuming pre-training step. Our solution relies on the fact that certain channels in the input are related to specific colors and similar attributes [13]. If such attributes are fed separately to the network, the visual transformer can better relate these attributes in the query to their regions-of-interest in its self-attention scheme, and thus have improved accuracy. Unlike MattNet [12], we perform the object attribute matching task in an unsupervised manner. Notably, we do not supply the network with information on the properties of each visual channel, and instead train the network to find an optimal clustering pattern for the provided data. This results in a learning setup in which the neural network implicitly learns which channels of the visual input relate to which attributes in the query.

We also notice that the object attributes that a user refers to can be associated with multiple sub-regions of a region-of-interest. Similar to utilizing channels, we provide an unsupervised clustering approach to group grid based pixels groups within a bounding box region of interest. We term this as spatial attention. The unsupervised spatial clustering aids in aligning referring attributes in the query with sub-regions of the region-of-interest. In our setup, learning the alignment between the text, object attributes, and pixel regions within a region of interest happens in an end-to-end, unsupervised fashion. Our contributions are as follows:

- We introduce a **channel attention** module that allows unsupervised clustering of object semantics in the channel dimension.
- We introduce a **spatial attention** module that groups sub-regions of the detected regions of interests and facilitates alignment between object semantics and geometric locations.
- We propose a framework that uses **relative positioning** and **attribute prediction loss** and demonstrate in our experiments that the neural network behaves as expected: clusters corresponds to different types of objects, attributes and spatial groups.

In the following, we outline related work in Section II. We present our Semantic VL-BERT model in Section III, and the results of our experiments on public datasets (RefCOCO, RefCOCO+) [14] are presented in Section IV.

II. Related Work

Multi-modal transformers allow both visual and linguistic embedded features as inputs. Each input either represents a word from the query (e.g., via a WordPiece embedding [15]), or a region-of-interest from the input image that has been identified with an object detector, such as a Faster R-CNN. Through the multi-head self-attention mechanism, the different layers of the transformer aggregate information from all the other elements. By stacking multiple layers of these multi-modal transformer attention modules, a rich representation that aggregates and aligns visual-linguistic clues is obtained as the final output. During inference, task-specific branches can be utilized at the last layer for various visio-linguistic tasks.

Over the last couple of years, tens of different such visual transformers have been proposed in the literature. A comprehensive survey of multi-modal transformers can be found in [17]. Most of these approaches use a single stream transformer as in Fig. 1a) to fuse information from image and language [6, 11, 8, 3], while others, such as ViLBERT [4] and LXMERT [5], use two transformers followed by a co-attentional transformer layer to generate joint representations as shown in Fig. 1b). There have been studies conducted [18] to create a general framework from these single vs.

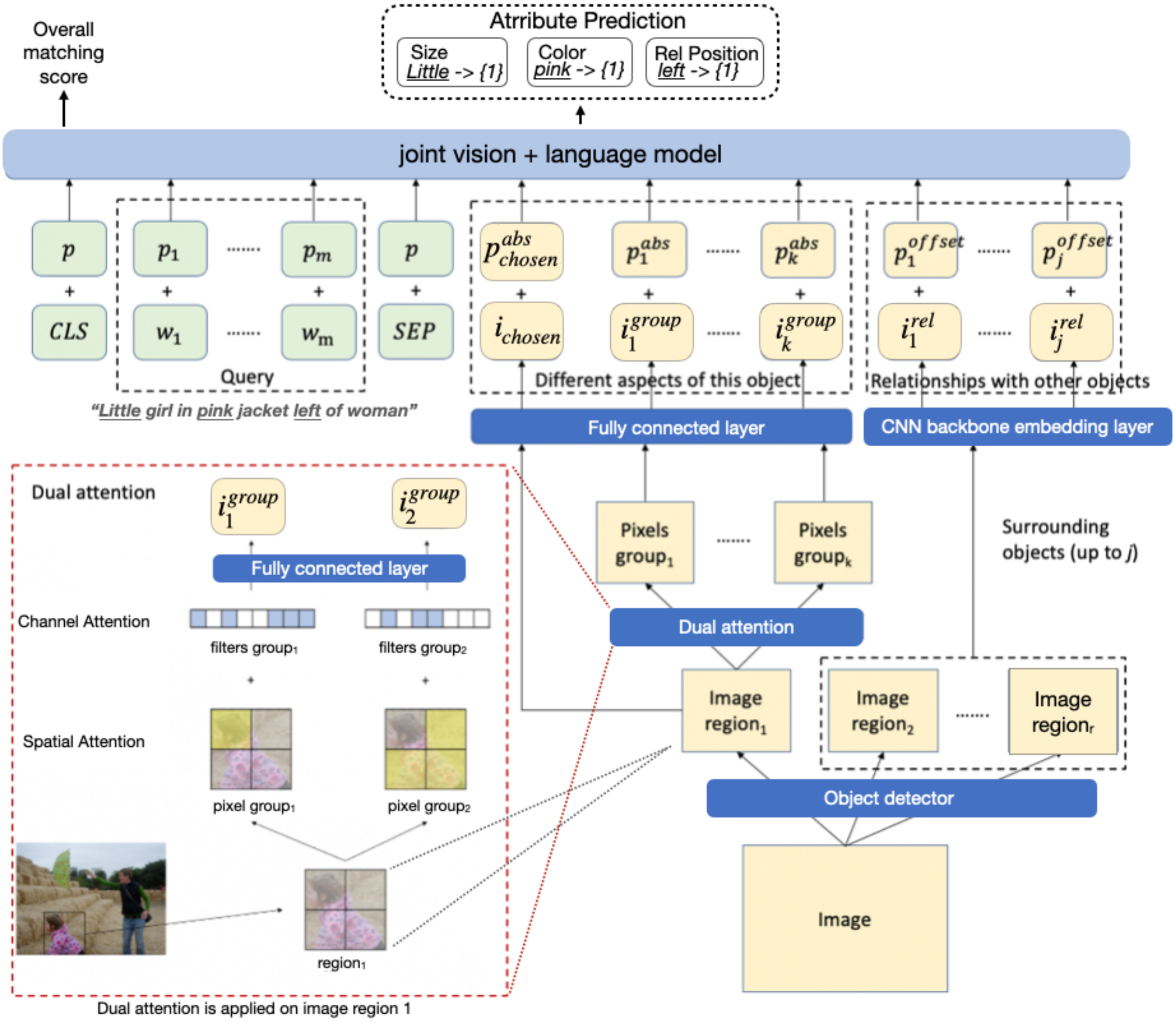


Fig. 2: Architecture of Semantic VL-BERT. The textual referring query is *"Little girl in pink jacket left of woman"*. We show the original image in the bottom left. On the bottom right of the image, we show the image being represented by $\{1, ..r\}$ image regions-of-interest (RoI). The figure shows **Dual Attention** being applied to the first "chosen" region of interest "Image region$_1$" where the region is further segmented into k pixel groups. The pixel groups for the first region are fed into a fully connected layer. This captures semantics within the first region of interest. The result is then concatenated with embeddings of nearby up to j RoIs $\{1, ..j\}$ with relative position offsets $p^{offset}_{1..j}$ to capture inter RoI semantics. On the bottom left of the figure, we show a zoomed in conceptual view of the **Dual Attention** module. It shows the "Image region$_1$" being clustered into spatial subgroups on a grid. The spatial attention is combined with channel attention to get the pixel groups. In the very top part of the figure we see the standard single stream multi-modal transformer architecture with textual embeddings w, image embeddings i, position embeddings p with the [CLS], [SEP] tokens from BERT [16].

multi stream transformer models. In other work, OSCAR [9] uses object tags detected in images as anchor points to significantly ease the learning of alignment. Similarly, ERNIE-ViL [10] pre-trains the model with scene graph annotations to construct detailed semantic connections across vision and language. Pixel-BERT [19] replaces region features with grid features, which relieves the cost of bounding box annotations and produces detailed visual representations. These models use a CNN based image backbone for the image processing stage. Recently vision transformers (ViT) [20, 21] have been applied to visual question answering, visual reasoning, image captioning, and retrieval [22, 23, 24, 25, 26] but they have limited success in visual grounding. The current state-of-the-art performance in visual grounding is held by MDETER [7] which is an end-to-end text-modulated detection system. Instead of utilizing region proposals from an object detector MDETER uses a CNN, Roberta [27] backbone to extract features, then projects them to a shared embedding space and directly predicts bounding boxes for objects in the referring expression in an encoder-decoder setup. MDETER utilizes 3 loss functions: (i) a loss based on bipartite matching between predicted and ground truth objects from prior work [28]; (ii) a soft token prediction loss which aligns predicted objects with token spans in the query; and (iii) a contrastive loss that aligns visual and textual embeddings. We show this architecture using Fig. 1c).

In this work, we propose to improve on single stream vision and language architectures similar to Fig. 1a) by utilizing the semantic meaning of visual features. We select VL-BERT [3] as our candidate single stream transformer model, as it has a generic, adaptable design with minimal pre-training tasks and custom losses that demonstrates the effectiveness of our approach. We note that our approach is applicable to any multi-modal single stream architecture similar to VL-BERT. As mentioned earlier the current state of the art is held by MDETER[7] which has a more complex architecture, loss functions and pre-training tasks than ours. ERNIE-ViL [10] is a comparable model to ours but requires pre-training on scene graphs which is an added ground truth information source. In this study, we focus on improving a minimal viable generic architecture that works well for visual grounding that is guided by the semantic attribute related components of the query, as opposed to beating an ever-evolving benchmark. Picking a simple single stream architecture leads to models that can satisfy stringent memory and latency requirements of modern-day Smart Home Assistants while providing accuracy values that are comparable to state-of-the-art models.

III. Semantic-VL-BERT

In this section we first describe a generic single stream multi-modal transformer architecture in Section III-A. Then, we describe our novel extensions with the aide of Fig. 2 in Section III-B.

A. Single Stream Multi-modal transformer architecture

The general multi-modal single stream transformer architecture uses $w_1, ..., w_m$ word embeddings, $i_1, ...i_j$ image embeddings and concatenates them with a [SEP] token to create a single sequence. A [CLS] token is inserted at the beginning of the sequence to capture the overall prediction outcome (e.g., predicted bounding box, classification score). These visio-linguistic embeddings are added to accompanying position embeddings; $\tilde{p}_1, ...\tilde{p}_m$ for text and $p_1, ...p_j$ for image. Finally, an additional embedding to denote the type of embedding (e.g., visual, text) is added, and the final representation is fed into a transformer model, e.g., BERT [16].

B. Semantic VL-BERT Enhancements

Our proposed architecture is shown in Fig. 2. It can be applied to any single stream transformer model as described in Section III-A that models vision and language jointly by (i) adding a dual attention module, (ii) using relative location information, and (iii) adding attribute prediction as an auxiliary task. Next, we detail each of these adjustments.

Dual Attention: Current models directly feed the tokenized query and regions-of-interest (RoI) to the transformer model. We propose to use a dual attention model to group pixels within an RoI and to select the most appropriate visual features for each group. As detailed previously, we expect each pixel group to capture a different type of visual semantics. For this goal, we utilize *Spatial* and *Channel Attention* schemes.

In ***Spatial Attention***, each image region-of-interest is divided into K grids. Spatial Attention clusters pixel grids into groups and generates a new vector representation for each group. More specifically, we let $V \in \mathbb{R}^{G \times D}$ represent grid features of an RoI [1]. We compute attention scores $a \in \mathbb{R}^{G \times K}$ for each grid location with

$$a = \sigma\left(VC^T\right), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function and $C \in \mathbb{R}^{K \times D}$ represents a matrix that consists of K cluster centroids. The visual representation of the k-th pixel group V^k is constructed by multiplying the grid features with the attention learned for that group;

$$V^k_{g,d} = a_{g,k} V_{g,d}, \tag{2}$$

where g indexes a position in the 14×14 grid G, d indexes a position in dimension D and k indexes the cluster centroids.

For ***Channel Attention***, we use guided attention over the channels to capture the various semantic components. We first aggregate spatial information of the visual feature vector of k-th group, V^k by using max-pooling operations over the grid positions G, giving $V^{k,\max} \in \mathbb{R}^{1 \times D}$. We empirically found max pooling to give the best results. Then, the max-pooled descriptors are forwarded to a multi-layer perceptron (MLP) parameterized by $W \in \mathbb{R}^{D \times D}$. The channel attention score for group k, $b_k \in \mathbb{R}^{1 \times D}$, is computed with

$$b_k = \sigma(V^{k,\max} W). \tag{3}$$

Then, as in Spatial Attention, the visual features, V_k are multiplied with the weights b_k and are passed to the transformer.

Relative Location: In order to model object relationships, we feed objects around the chosen RoI with their offset coordinates (relative to the RoI) to the model to provide context information. More specifically, for the j-th surrounding object of the chosen RoI, we calculate the offset coordinates by mapping the 4-d vector

$$\left[\frac{[\Delta x_{tl}]_j}{w_j}, \frac{[\Delta y_{tl}]_j}{h_j}, \frac{[\Delta x_{br}]_j}{w_j}, \frac{[\Delta y_{br}]_j}{h_j}\right], \tag{4}$$

into a high-dimensional representation p_j^{offset}, and concatenate it with its visual feature embedding i_j. Here, (x_{tl}, y_{tl}), (x_{br}, y_{br}), w, h denote the top-left coordinate, bottom-right coordinate, image width and image height for the j-th surrounding object of the chosen RoI. For each such element, we aggregate its visual features, element type, and relative position to build the final embedding feature.

[1] We take $G = 14 \times 14$ and $D = 1024$ in this paper. Please note the abuse of notation, as we assume that V is vectorized (i.e., flattened) in its second dimension.

Attribute prediction as an auxiliary task: We add an additional optional attribute prediction auxiliary task on top of the visual grounding task. Attributes can be obtained from parsing the referring expression [29, 30], scene graphs [31], or via using object attribute detection modules such as VinVL [32]. Each image is associated with a binary vector that is 1 if the attribute is present and 0 otherwise. When this auxiliary task is present, there is an auxiliary attribute prediction loss. We conduct experiments modelling the loss using binary cross-entropy and in a multi label fashion. For our experiments with the RefCOCO, RefCOCO+ data sets, out of the many attributes that are available, we pick color, size and relative location since they are the key referring terms used in referring expressions. For our experiments, the attributes are parsed from the referring expression using a template parser [30]. The attribute prediction loss is given below where n ranges over the data points, a ranges over the attributes and λ_{attr} is the weight assigned to the attribute prediction loss. The ground truth and predictions are denoted by y, $\hat{q}$ respectively

$$\begin{aligned} L_{attr} = \lambda_{attr} \sum_{n} \sum_{a} y_{na} \log(\hat{q}_{na}) \\ +(1 - y_{na}) \log(1 - \hat{q}_{na}). \end{aligned} \tag{5}$$

This attribute prediction loss is added to the loss of correctly predicting the ground truth bounding box L_{gt} to get the final loss. For L_{gt} each region is labelled as positive only when the intersection-over-union between the predicted box and the ground truth box is over 0.5.

$$L_{final} = L_{gt} + L_{attr}. \tag{6}$$

L_{gt} and L_{attr} are optimized based on a *joint vision + language* internal multi-modal representation, as seen in the top part of Fig. 2. In other words, this creates a multi-task setting where the two task related loss entities are linked because they optimize over the same underlying multi-model internal representation space. This in turn links the K pixel groups with the attributes, as their associated losses are jointly optimized over the above-mentioned multi-modal representation space. As part of future work, we plan to do an in-depth analysis of multi-modal feature representation spaces similar to [33, 13] to solidify the link between the query attributes and visual sub-regions.

IV. Experiments

A. Datasets

In this section, we present the results of our experiments on the RefCOCO and RefCOCO+ [29, 34], two datasets that are the most commonly used data sets for the visual grounding task. Both of the datasets consist of more than 140,000 referring expressions for 50,000 different objects from 20,000 different images [34]. The datasets are generated via the ReferItGame [29] where 2 players play an online object referring game. In the RefCOCO dataset, there are no restrictions placed on the type of language used in the referring expressions, while for the RefCOCO+ dataset the players are disallowed from using location related words. Thus, RefCOCO+ is expected to contain more words related to attributes than location. The creators of the data sets have split the test sets into two splits. TestA contains images of multiple people, and TestB contains multiple instances of all other objects. For more detailed information related to data sets and splits, we refer to [29, 34].

B. Experimental Setup

We follow the same framework as in VL-BERT$_{\text{base}}$ in our implementation: the transformer uses the BERT-base structure with 12 layers, and the visual features are computed using a Faster R-CNN with a ResNet-101 backbone. We use ADAM with learning rate and weight decay parameters, both given by 10^{-4}. Similar to VL-BERT, the model is only pretrained on Conceptual Captions dataset [35] as the visual linguistic corpus. The 3 main hyperparameters in our architecture are: K, the number of spatial attention groups; j, number of relative location components; and λ_{attr} the weight of the auxiliary attribute prediction task's loss. We set $K = 3$, and $j = 5$. We empirically find $K = 3$ groups to give the best results for RefCOCO and RefCOCO+ datasets. We limit j to 5 due to memory constraints during training and inference time, as each additional RoI object bring an associated memory cost. To tune the loss weight parameter, λ_{attr}, we use a grid-search and present the results of the best parameter as chosen by the validation score.

C. Results and case study

We compare our approach with the baseline model (VL-BERT) and a non-transformer visual grounding model (MAttNet) [3, 12] in Table I. From Table I it can be seen that our approach provides a 1.32% absolute improvement in accuracy over VL-BERT. A comparison with the current state-of-the-art is provided towards the end of the section in Table V for completeness.

TABLE I: Accuracy on RefCOCO validation dataset with ground truth bounding boxes.

Model	RefCOCO Val Acc.
MAttNet	85.5%
VL-BERT[2]	89.1%
Semantic-VL-BERT	**90.42%**

TABLE II: Results of the ablation study on the RefCOCO validation set.

Model	Abs. Improv	RefCOCO Val Acc.
Baseline (VL-BERT)[2]	-	89.1 %
Baseline + Relative Location Offset	*+0.9%*	90.0 %
Baseline + Relative Location Offset + Spatial Attention	*+0.2%*	90.2%
Baseline + Relative Location Offset + Spatial & Channel Attention	*+0.22%*	**90.42%**

[2] Numbers recreated by the authors based on original code given in [3]

TABLE III: Accuracy results on RefCOCO & RefCOCO+ datasets using the ground truth object bounding boxes.

Model	RefCOCO Val	Δ	RefCOCO TestA	Δ	RefCOCO TestB	Δ	RefCOCO+ Val	Δ	RefCOCO+ TestA	Δ	RefCOCO+ TestB	Δ
Baseline (VL-BERT)[3]	89.1%	-	90.2%	-	88.0%	-	79.3%	-	82.1%	-	73.5%	-
Semantic VL-BERT *(w/o attribute prediction)*	**90.4%**	+1.3%	91.0%	+0.8%	88.6%	+0.6%	80.0%	+0.7%	**83.6%**	+1.5%	73.8%	+0.3%
Semantic VL-BERT *(w/ attribute prediction)*	**90.4%**	+1.3%	**91.3%**	+1.1%	**89.5%**	+1.5%	**80.1%**	+0.8%	82.9%	+0.8%	**74.4%**	+0.9%

TABLE IV: Accuracy results on RefCOCO & RefCOCO+ datasets using the object bounding box proposals by MattNet [12].

Model	RefCOCO Val	Δ	RefCOCO TestA	Δ	RefCOCO TestB	Δ	RefCOCO+ Val	Δ	RefCOCO+ TestA	Δ	RefCOCO+ TestB	Δ
Baseline (VL-BERT)[3]	79.6%	-	85.2%	-	**73.0%**	-	71.3%	-	77.2%	-	60.9%	-
Semantic VL-BERT *(w/o attribute prediction)*	80.0%	+0.4%	**85.6%**	+0.4%	72.7%	-0.3%	71.7%	+0.4%	78.1%	+0.9%	**62.4%**	+1.5%
Semantic VL-BERT *(w/ attribute prediction)*	**80.2%**	+0.6%	85.2%	0.0%	72.7%	-0.3%	**72.0%**	+0.7%	**78.6%**	+1.4%	62.0%	+1.1%

To analyze the source of the improvement, we conduct an ablation study in which we add our suggested improvements step by step. The results of the ablation study are given in Table II. It is seen that the largest improvements are provided by adding relative locations of nearby objects, followed by channel and spatial attention, respectively.

We then compare our approach with the baseline VL-BERT [3] model with and without the attribute prediction loss on the RefCOCO and RefCOO+ datasets in Table III. The results are based on ground truth bounding boxes. Semantic VL-BERT improves over VL-BERT in all the cases and the most significant improvements are seen on the TestB dataset. This is expected as TestB consists of multiple non-person objects, its referring expressions contain more attribute words compared to others, and Semantic VL-BERT is trained to make better use of such attributes.

In Table IV we switch the grounding boxes with the MattNet detected bounding box proposals[12] and evaluate the performance. Here, Semantic VL-BERT improves over VL-BERT in all the cases except RefCOCO Test B. We hypothesize that there is a larger misalignment between referring expressions and MattNet proposed regions in RefCOCO Test B which leads to this degradation.

We see that Semantic-VL-BERTs with and without the auxiliary attribute prediction task are better than the baseline in almost all instances based on Table III, IV. To better understand the driving factors for the improvement, we perform a qualitative case study by inspecting bounding boxes picked by Semantic-VL-BERT with its accompanying spatial attention groups in Fig. 3 and Fig. 4. The queries are displayed in the figure captions. Over the image, the VL-BERT prediction is plotted in red and Semantic-VL-BERT prediction is given in blue. Next to the image are the spatial attention maps that correspond to different pixel groups. We can see in Fig. 3, the third pixel group aligns with *the brown bruised spot*, and in Fig. 4 the second pixel group aligns with *sunglasses on her head* and the third group aligns with *the color white*.

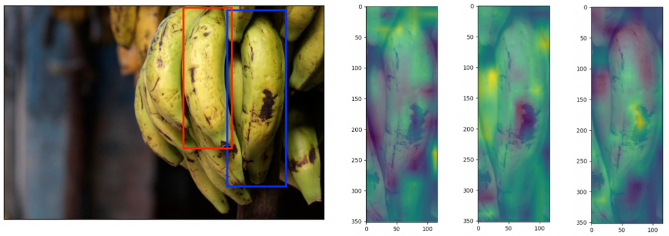


Fig. 3: Query: *"Banana with large brown spot close to the right"*. The original image is shown to the left and the $K = 3$ pixel groups to the right. The parsed attributes are: "color:brown", "size:large", "relative_location:to the right".

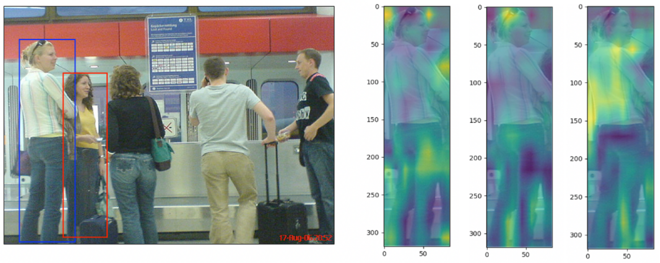


Fig. 4: Query: *"Lady in white with sunglasses on her head"*. The original image is shown to the left and the $K = 3$ pixel groups to the right. The parsed attributes are: "color:white".

We find that the additional spatial attention pixel groups that are created from our dual attention method act as a data augmentation scheme. The additional pixel groups, channel attention coupled with the attribute prediction loss help the model to group and better align the referring expression with the visual features provided initially via the Faster R-CNN.

We add more positive and negative examples accompanied by spatial attention maps for $K = 3$ spacial attention groups to illustrate our point further in Appendix A.

In order to paint a complete picture, we also list down the

[3] Numbers recreated by the authors based on original code given in [3]

TABLE V: Accuracy Comparison with state-of-the-art models on RefCOCO+ dataset using MattNet object bounding box proposals [12]

Model	**RefCOCO+ Acc.**		
	Val	TestA	TestB
MattNet	65.3%	71.6%	56.0%
Baseline (VL-BERT)[4]	71.3%	77.2%	60.9%
Semantic VL-BERT	72.0%	78.6%	62.0%
ERNIE-ViL	74.0%	80.3%	64.7%
MDETER	**81.1%**	**85.5%**	**72.9%**

current state-of-the-art results for visual grounding in Table V. We select the results on RefCOCO+ dataset using object bounding box proposals by MattNet, as those results are the most common denominator among reported results for the state-of-the-art models for visual grounding. As mentioned in Section II, MDETER [7] holds the current state-of-the-art performance but is a more complex model with an encoder-decoder setup. ERNIE-ViL [10] is reliant on scene graph annotations and pretrains on a larger dataset (1M additional image-text pairs).

Conclusion

We introduce a dual attention mechanism with relative positioning and attribute prediction loss that enables the multi-modal model to better capture semantics and align visual features with text. Our model does not require any additional pre-training and can be added on top of any single stream joint vision and language model, similar to VL-BERT. We demonstrate our model's effectiveness on the RefCOCO and RefCOCO+ datasets and show significant improvements over the baseline VL-BERT model.

References

[1] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.

[2] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016.

[3] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[4] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 13–23, 2019.

[5] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019.

[6] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019.

[7] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. *CoRR*, abs/2104.12763, 2021.

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *CoRR*, abs/1909.11740, 2019.

[9] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *CoRR*, abs/2004.06165, 2020.

[10] 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. *CoRR*, abs/2006.16934, 2020.

[11] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press, 2020.

[12] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[13] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti,

[4] Numbers recreated by the authors based on original code given in [3]

Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL, 2014.

[15] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[17] Salman H. Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *CoRR*, abs/2101.01169, 2021.

[18] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal Pretraining Unmasked: Unifying the Vision and Language BERTs. *arXiv e-prints*, page arXiv:2011.15124, November 2020.

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

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

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv e-prints*, page arXiv:2012.12877, December 2020.

[22] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. *arXiv e-prints*, page arXiv:2102.03334, February 2021.

[23] Socratis Gkelios, Yiannis Boutalis, and Savvas A. Chatzichristofis. Investigating the Vision Transformer Model for Image Retrieval Tasks. *arXiv e-prints*, page arXiv:2101.03771, January 2021.

[24] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training Vision Transformers for Image Retrieval. *arXiv e-prints*, page arXiv:2102.05644, February 2021.

[25] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. CPTR: Full Transformer Network for Image Captioning. *arXiv e-prints*, page arXiv:2101.10804, January 2021.

[26] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating nested transformers. *CoRR*, abs/2105.12723, 2021.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[28] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.

[29] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics.

[30] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[32] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Making Visual Representations Matter in Vision-Language Models. *arXiv e-prints*, page arXiv:2101.00529, January 2021.

[33] Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. https://distill.pub/2021/multimodal-neurons.

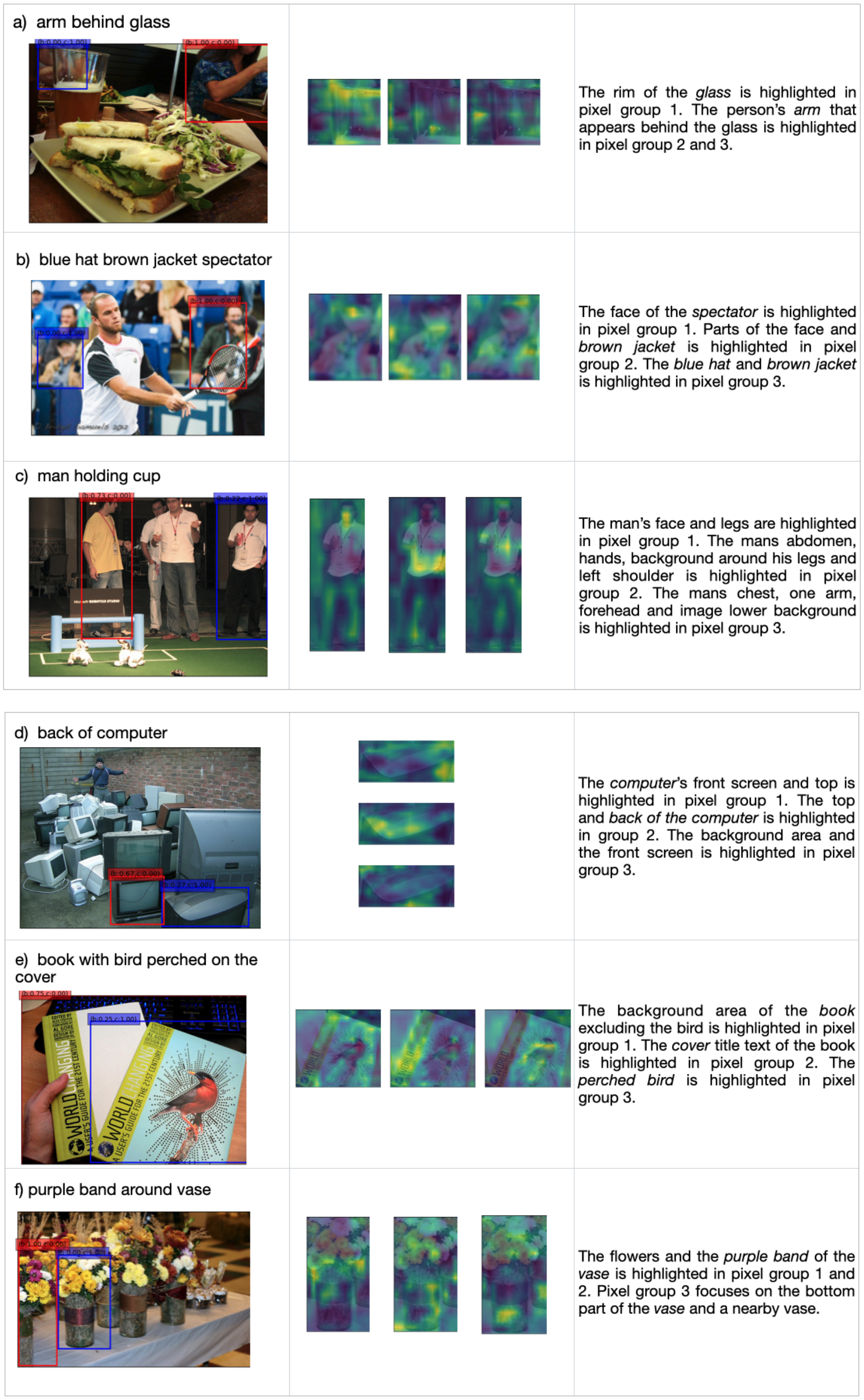
[34] Licheng Yu, Patric Poirson, Shan Yang, Alexander Berg, and Tamara Berg. Modeling context in referring expressions. 07 2016.

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu 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)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Appendix

A. Additional positive and negative examples

In Fig. 5 we show positive examples where the Semantic-VL-BERT infers the correct bounding box as opposed to the baseline. In the column adjoining each image, we show the spatial attention maps for $K = 3$ pixel groups for the RoI chosen by Semantic-VL-BERT. In both figures, the box colored in red corresponds to the baseline model and the box in blue corresponds to the Semantic VL-BERT model.



a) arm behind glass

The rim of the *glass* is highlighted in pixel group 1. The person's *arm* that appears behind the glass is highlighted in pixel group 2 and 3.

b) blue hat brown jacket spectator

The face of the *spectator* is highlighted in pixel group 1. Parts of the face and *brown jacket* is highlighted in pixel group 2. The *blue hat* and *brown jacket* is highlighted in pixel group 3.

c) man holding cup

The man's face and legs are highlighted in pixel group 1. The mans abdomen, hands, background around his legs and left shoulder is highlighted in pixel group 2. The mans chest, one arm, forehead and image lower background is highlighted in pixel group 3.

d) back of computer

The *computer*'s front screen and top is highlighted in pixel group 1. The top and *back of the computer* is highlighted in group 2. The background area and the front screen is highlighted in pixel group 3.

e) book with bird perched on the cover

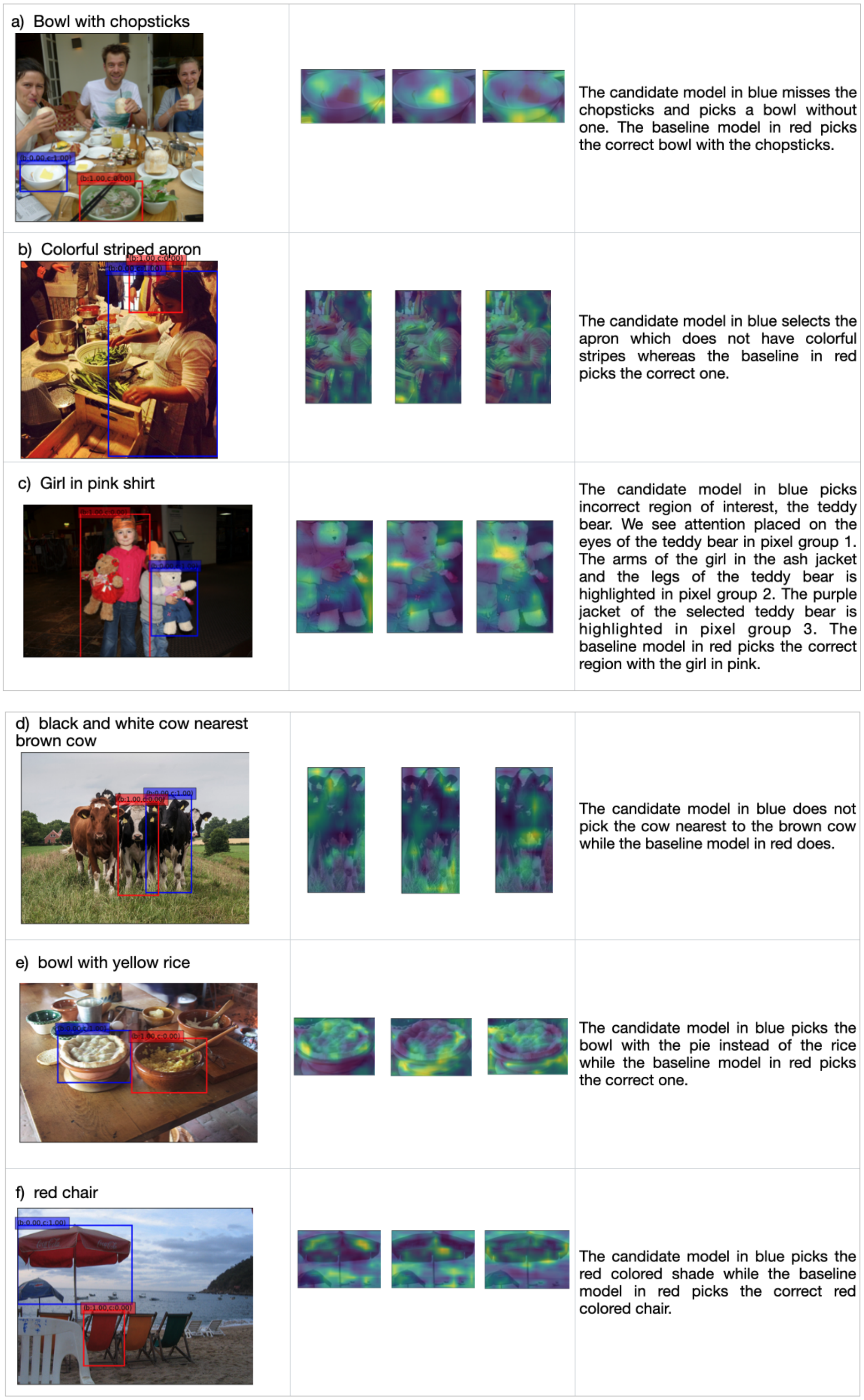
The background area of the *book* excluding the bird is highlighted in pixel group 1. The *cover* title text of the book is highlighted in pixel group 2. The *perched bird* is highlighted in pixel group 3.

f) purple band around vase

The flowers and the *purple band* of the *vase* is highlighted in pixel group 1 and 2. Pixel group 3 focuses on the bottom part of the *vase* and a nearby vase.

Fig. 5: Examples where Semantic-VL-BERT prediction is correct and the VL-BERT prediction is wrong

In Fig. 6 we show negative examples where Semantic-VL-BERT infers the incorrect bounding box.



a) Bowl with chopsticks

The candidate model in blue misses the chopsticks and picks a bowl without one. The baseline model in red picks the correct bowl with the chopsticks.

b) Colorful striped apron

The candidate model in blue selects the apron which does not have colorful stripes whereas the baseline in red picks the correct one.

c) Girl in pink shirt

The candidate model in blue picks incorrect region of interest, the teddy bear. We see attention placed on the eyes of the teddy bear in pixel group 1. The arms of the girl in the ash jacket and the legs of the teddy bear is highlighted in pixel group 2. The purple jacket of the selected teddy bear is highlighted in pixel group 3. The baseline model in red picks the correct region with the girl in pink.

d) black and white cow nearest brown cow

The candidate model in blue does not pick the cow nearest to the brown cow while the baseline model in red does.

e) bowl with yellow rice

The candidate model in blue picks the bowl with the pie instead of the rice while the baseline model in red picks the correct one.

f) red chair

The candidate model in blue picks the red colored shade while the baseline model in red picks the correct red colored chair.

Fig. 6: Examples where Semantic-VL-BERT prediction is wrong and the VL-BERT prediction is correct