Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts

Yan Zeng [1] Xinsong Zhang [1] Hang Li [1]

Abstract

Most existing methods in vision language pre-training rely on object-centric features extracted through object detection and make fine-grained alignments between the extracted features and texts. It is challenging for these methods to learn relations among multiple objects. To this end, we propose a new method called X-VLM [1] to perform 'multi-grained vision language pre-training.' The key to learning multi-grained alignments is to locate visual concepts in the image given the associated texts, and in the meantime align the texts with the visual concepts, where the alignments are in multi-granularity. Experimental results show that X-VLM effectively leverages the learned multi-grained alignments to many downstream vision language tasks and consistently outperforms state-of-the-art methods.

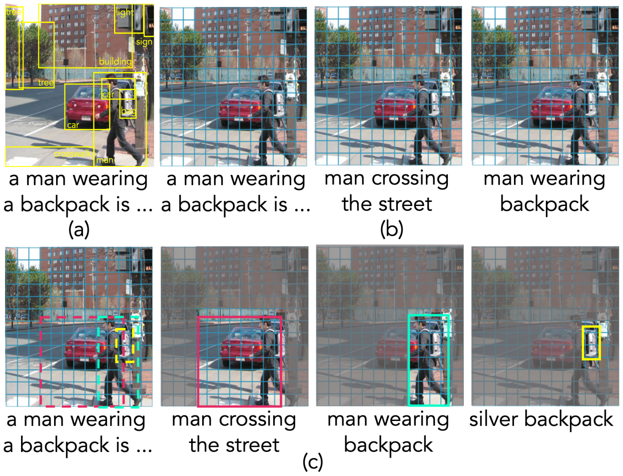

Figure 1. A comparison of (a) the existing methods relying on object detection, (b) the methods aligning the texts with the whole image, and (c) our approach.

1. Introduction

Vision language pre-training aims to learn vision language alignments from a large number of image-text pairs. A pre-trained Vision Language Model (VLM) fine-tuned with a small amount of labeled data has shown the state-of-the-art performances in many Vision Language (V+L) tasks such as visual question answering and image-text retrieval.

Existing methods learning vision language alignments fall into two approaches as shown in Figure 1 (a, b). Most of them detect objects in the image and align the text with fine-grained (object-centric) features. They either utilize pre-trained object detectors (Tan & Bansal, 2019; Lu et al., 2019; Li et al., 2019; 2020a; Chen et al., 2020; Li et al., 2020b; Gan et al., 2020) or conduct object detection on-the-fly in the pre-training process (Su et al., 2020; Xu et al., 2021a). The other methods do not rely on object detection and only learn alignments between the texts and coarse-grained (overall) features of the image (Huang et al., 2020; 2021; Kim et al., 2021; Li et al., 2021a).

Both the fine-grained and coarse-grained approaches have drawbacks. Object detection identifies all possible objects in the image, and some of them might not be relevant to the text. Object-centric features cannot easily represent relations among multiple objects, e.g. "man crossing the street". Moreover, it is challenging to pre-define the categories of objects suitable for downstream tasks. On the other hand, the coarse-grained approaches cannot effectively learn fine-grained alignments between vision and language, e.g. object-level, which has shown to be critical for some downstream tasks such as visual reasoning, visual grounding, and image captioning.

Ideally, we want a VLM to learn multi-grained alignments between vision and language in pre-training, which are not restricted to object-level or image-level, and leverage the learned alignments to downstream V+L tasks. Unfortunately, existing methods cannot satisfactorily handle multi-

[1]ByteDance AI Lab. Correspondence to: Yan Zeng <zengyan.yanne@bytedance.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1]The code and pre-trained models are available at https://github.com/zengyan-97/X-VLM.

grained alignments between vision and language.

In this paper, we propose performing multi-grained vision language pre-training by aligning text descriptions with the corresponding visual concepts in images. Taking Figure 1 as an example, we have the following data for training: 1) the image caption describing the whole image; 2) region annotations such as "man wearing backpack" each of which has been related to a region in the image, while previous approaches roughly align the region descriptions with the whole image; 3) object labels such as "backpack" which are utilized by previous methods to train object detectors. We re-formulate the data, so that an image may have multiple bounding boxes, and a text [2] is directly associated with the visual concept in each box. The 'visual concept' (Krishna et al., 2017; Zhang et al., 2021; Changpinyo et al., 2021) may be an object, a region, or the image itself, as the example in Figure 1 (c). By doing so, our approach learns unlimited visual concepts associated with diverse text descriptions, which are also not restricted to object-level or image-level.

Our multi-grained model, denoted as X-VLM, consists of an image encoder that produces representations of visual concepts (including the image itself) in an image, a text encoder, and a cross-modal encoder that conducts cross-attention between the vision features and language features to learn vision language alignments. The key to learning multi-grained alignments is to optimize X-VLM by: 1) locating visual concepts in the image given associated texts by a combination of box regression loss and intersection over union loss; 2) in the meantime aligning the texts with the visual concepts, e.g. by a contrastive loss, a matching loss, and a masked language modeling loss, where the alignments are in multi-granularity, as illustrated in Figure 1 (c). In fine-tuning and inference, X-VLM can leverage the learned multi-grained alignments to perform the downstream V+L tasks without bounding box annotations in the input images.

We demonstrate the effectiveness of our approach on various downstream tasks. On image-text retrieval, X-VLM learning multi-grained vision language alignments outperforms VinVL (Zhang et al., 2021) which is based on object-centric features, achieving an absolute gain of 4.65% in terms of R@1 score on MSCOCO. X-VLM also outperforms ALIGN (Jia et al., 2021), ALBEF (Li et al., 2021a), and METER (Dou et al., 2021) by a large margin even though they are pre-trained on more data or have more parameters. On visual reasoning tasks, X-VLM achieves absolute improvements of 0.79% on VQA and 1.06% on NLVR2 compared to VinVL (Zhang et al., 2021), with a much faster inference speed. X-VLM also outperforms SimVLM$_{\mathrm{base}}$ (Wang et al., 2021) pre-trained with 1.8B in-house data, especially on NLVR2 by 2.4%. On visual grounding (RefCOCO+), X-VLM achieves absolute improvements of 4.5% compared to UNITER (Chen et al., 2020) and 1.1% compared to MDETR (Kamath et al., 2021) which is specialized for grounding tasks. X-VLM also has comparable performance with SimVLM$_{\mathrm{base}}$ in the image caption generation task.

The contributions of this work are as follows:

- We propose performing multi-grained vision language pre-training to handle the alignments between texts and visual concepts.
- We propose to optimize the model (X-VLM) by locating visual concepts in the image given the associated texts and in the meantime aligning the texts with the visual concepts, where the alignments are in multi-granularity.
- We empirically verify that our approach effectively leverages the learned multi-grained alignments in fine-tuning. X-VLM consistently outperforms existing state-of-the-art methods on many downstream V+L tasks.

2. Related Work

The existing work on vision language pre-training typically falls into two categories: fine-grained and coarse-grained.

Most existing methods belong to the fine-grained approach, which relies on object detection (Tan & Bansal, 2019; Lu et al., 2019; Li et al., 2019; 2020a; Chen et al., 2020; Li et al., 2020b; Gan et al., 2020; Li et al., 2021b). An object detector first identifies all regions that probably contain an object, then conducts object classification on each region. An image is then represented by dozens of object-centric features of the identified regions. Object detectors, such as Faster R-CNN (Ren et al., 2015), Bottom-Up and Top-Down Attention (BUTD) (Anderson et al., 2018), are trained on image annotations of common objects, e.g. COCO (Lin et al., 2014) (110K images) and Visual Genome (Krishna et al., 2017) (100K), and can be utilized. VinVL (Zhang et al., 2021) has, for example, achieved SoTA performances on many V+L tasks by utilizing a powerful object detector pre-trained with a large collection of image annotations (2.5M images). The challenge with the approach is that object-centric features cannot represent relations among multiple objects in multiple regions. Furthermore, it is not easy to define the categories of objects in advance that are useful for downstream V+L tasks.

The coarse-grained approach builds VLMs by extracting and encoding overall image features with convolutional network (Jiang et al., 2020; Huang et al., 2020; 2021) or vision

[2] We take the object labels as text descriptions of objects.

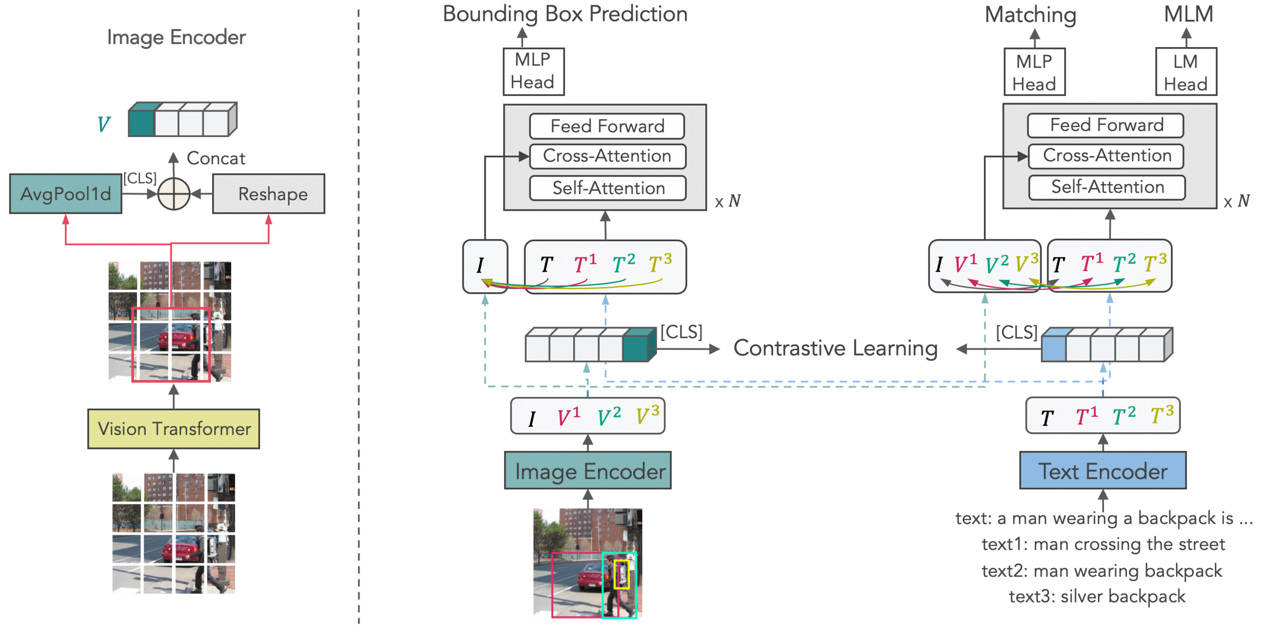

Figure 2. Pre-training model architecture and objectives of X-VLM. As shown on the left side, we extract features from the subset of patches from the vision transformer to represent images/regions/objects (I and V^{1-3}), which are then paired with corresponding text features (T and T^{1-3}) for contrastive learning, matching, and MLM. Meanwhile, the image (I) is paired with different textual descriptions (T and T^{1-3}) for bounding box prediction to locate visual concepts in the image.

transformer (Kim et al., 2021; Li et al., 2021a). The performances are usually not as good as the fine-grained approach. Though object-centric features are only related to certain objects, learning fine-grained alignments, e.g. object-level, has shown to be critical for some downstream tasks such as visual reasoning and visual grounding. To cope with the problem, SOHO (Huang et al., 2021) employs online clustering on image features to obtain more comprehensive representations, ViLT (Kim et al., 2021) uses a more advanced vision transformer, i.e. Swin-Transformer (Liu et al., 2021b) for image encoding, and ALBEF (Li et al., 2021a) exploits contrastive learning and momentum distillation in learning of image-text alignments. However, the improvements still cannot close the gap with the fine-grained approach.

Recently, there emerge some methods managing to learn both object-level and image-level alignments. However, these approaches still rely on object detectors and thus suffer from the aforementioned problems. For example, VL-BERT (Su et al., 2020) incorporates Faster R-CNN into pre-training. E2E-VLP (Xu et al., 2021a) adds an end-to-end object detection module (i.e. DETR (Carion et al., 2020)). Uni-EDEN (Li et al., 2022) uses Faster R-CNN as the vision backbone. KD-VLP (Liu et al., 2021a) relies on external object detectors to perform object knowledge distillation. In contrast, X-VLM does not rely on object detection. Besides, X-VLM learns multi-grained vision language alignments, which are not restricted to object-level or image-level. Also, unlike Uni-EDEN, which aligns objects to language by object classification and aligns images to language by caption generation, X-VLM learns visual concepts in different granularities in a unified way. We will show the effectiveness of X-VLM in the experiments.

3. Method

3.1. Overview

X-VLM consists of an image encoder (I_{trans}), a text encoder (T_{trans}), and a cross-modal encoder (X_{trans}). All encoders are based on Transformer (Vaswani et al., 2017). The cross-modal encoder fuses the vision features with the language features through cross-attention at each layer.

We re-formulate the widely used pre-training datasets (see Section 4.1) so that an image may have multiple bounding boxes, and each of them is associated with a text that describes an object or a region, denoted as $(I, T, \{(V^j, T^j)\}^N)$. Note that some images do not have associated texts, i.e., T is NaN, and some images do not have bounding boxes, i.e., $N = 0$. Here, V^j is an object or region in the bounding box $\boldsymbol{b}^j = (cx, cy, w, h)$ represented by the normalized center coordinates, width, and height of the box. When the image itself represents a visual concept, $\boldsymbol{b} = (0.5, 0.5, 1, 1)$. Figure 2 illustrates the architecture and pre-training objectives of X-VLM.

3.2. Vision Encoding

The image encoder efficiently produces multi-grained visual concept representations in an image. The encoder is based on vision transformer (Dosovitskiy et al., 2020). It first splits an image into non-overlapping patches and linearly embeds all patches. Then, these patches are passed into the transformer layers, yielding $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_{N^I}\}$. For an image of resolution of 224x224 and patch size of 32x32, we have $N^I = 49$.

We assume that $\boldsymbol{v}_{p_i}$ encodes the information of the corresponding patch p_i. Therefore, we represent a visual concept V^j (object, region, or the image) that corresponds to a set of patches by aggregating information among the patches as shown in Figure 2. Specifically, we reshape the patch features while keeping their position information, denoted as $\{\boldsymbol{v}_{p_1^j}, ..., \boldsymbol{v}_{p_M^j}\}$. $\{p_1^j, ..., p_M^j\}$ are patches of V^j. We also calculate the average of the features to represent the whole visual concept, denoted as $\boldsymbol{v}_{\text{cls}}^j$, and prepend it.

The image encoder then creates $N+1$ concept representations in different granularities, represented as $I_{\text{trans}}(V^j) = \{\boldsymbol{v}_{\text{cls}}^j, \boldsymbol{v}_{p_1^j}, ..., \boldsymbol{v}_{p_M^j}\}, j \in [0, N]$. We let $I_{\text{trans}}(V^0)$ denote the image representation in which all patch features are utilized. In the following section, we will describe how the representations are utilized in the learning of multi-grained alignments.

3.3. Cross-Modal Modeling

As shown in Figure 2, we optimize X-VLM by locating visual concepts in the image given the corresponding texts and in the meantime aligning the texts and visual concepts, where the alignments are in multi-granularity.

Bounding Box Prediction We let the model predict the bounding box $\boldsymbol{b}^j$ of visual concept V^j given the image representation and the text representation, where $\boldsymbol{b}^j = (cx, cy, w, h)$. By locating different visual concepts in the same image, we expect that the model better learns fine-grained vision language alignments. The bounding box is predicted by:

$$\hat{\boldsymbol{b}}^j(I, T^j) = \text{Sigmoid}(\text{MLP}(\boldsymbol{x}_{\text{cls}}^j)), \qquad (1)$$

where Sigmoid is for normalization, MLP denotes multi-layer perceptron, and $\boldsymbol{x}_{\text{cls}}^j$ is the output `[CLS]` embedding of the cross-modal encoder given I and T^j.

For bounding box prediction, ℓ_1 is the most commonly-used loss. However, it has different scales for small and large boxes, even if their relative errors are similar. To mitigate this issue, we use a linear combination of the ℓ_1 loss and the generalized Intersection over Union (IoU) loss (Rezatofighi et al., 2019), which is scale-invariant. The overall loss is defined as:

$$\mathcal{L}_{\text{bbox}} = \mathbb{E}_{(V^j,T^j)\sim I; I\sim D}[\mathcal{L}_{\text{iou}}(\boldsymbol{b}_j, \hat{\boldsymbol{b}}_j) + ||\boldsymbol{b}_j - \hat{\boldsymbol{b}}_j||_1] \qquad (2)$$

Meanwhile, we align texts and visual concepts by three objectives which are widely used in vision language pre-training (Chen et al., 2020; Radford et al., 2021; Li et al., 2021a). We extend the objectives to incorporate multi-grained visual concepts in the images.

Contrastive Learning We predict (visual concept, text) pairs, denoted (V, T), from in-batch negatives. Note that visual concepts include objects, regions, and images. Similar to Radford et al. (2021), we randomly sample a mini-batch of N pairs, and calculate the in-batch vision-to-text similarity and text-to-vision similarity.

Given a pair (V, T), T is the positive example for V, and we treat the other $(N-1)$ texts within the mini-batch as negative examples. We define cosine similarity $s(V, T) = g_v(\boldsymbol{v}_{\text{cls}})^\top g_w(\boldsymbol{w}_{\text{cls}})$. $\boldsymbol{w}_{\text{cls}}$ is the output `[CLS]` embedding of the text encoder. g_v and g_w are transformations that map the `[CLS]` embeddings to normalized lower-dimensional representations. Then, we calculate the in-batch vision-to-text similarity as:

$$p^{\text{v2t}}(V) = \frac{\exp(s(V, T)/\tau)}{\sum_{i=1}^N \exp(s(V, T^i)/\tau)}, \qquad (3)$$

Similarly, the text-to-vision similarity is:

$$p^{\text{t2v}}(T) = \frac{\exp(s(V, T)/\tau)}{\sum_{i=1}^N \exp(s(V^i, T)/\tau)}, \qquad (4)$$

where τ is a learnable temperature parameter. Let $\boldsymbol{y}^{\text{v2t}}(V)$ and $\boldsymbol{y}^{\text{t2v}}(T)$ denote the ground-truth one-hot similarity, in which only the positive pair has the probability of one. The contrastive loss is defined as the cross-entropy H between $\boldsymbol{p}$ and $\boldsymbol{y}$:

$$\begin{aligned}\mathcal{L}_{\text{cl}} = \frac{1}{2}\mathbb{E}_{V,T\sim D}\big[&\text{H}(\boldsymbol{y}^{\text{v2t}}(V), \boldsymbol{p}^{\text{v2t}}(V)) \\ &+ \text{H}(\boldsymbol{y}^{\text{t2v}}(T), \boldsymbol{p}^{\text{t2v}}(T))\big] \qquad (5)\end{aligned}$$

Matching Prediction We determine whether a pair of visual concept and text is matched. For each visual concept in a mini-batch, we sample an in-batch hard negative text by following $p^{\text{v2t}}(V)$ in Equation 3. Texts that are more relevant to the concept are more likely to be sampled. We also sample one hard negative visual concept for each text. We use $\boldsymbol{x}_{\text{cls}}$, the output `[CLS]` embedding of the cross-modal encoder, to predict the matching probability p^{match}, and the loss is:

$$\mathcal{L}_{\text{match}} = \mathbb{E}_{V,T\sim D}\text{H}(\boldsymbol{y}^{\text{match}}, \boldsymbol{p}^{\text{match}}(V, T)) \qquad (6)$$

where $\boldsymbol{y}^{\text{match}}$ is a 2-dimensional one-hot vector representing the ground-truth label.

Masked Language Modeling We predict the masked words in the text based on the visual concept. We randomly mask out the input tokens with a probability of 25%, and the replacements are 10% random tokens, 10% unchanged, and 80% `[MASK]`. We use the cross-modal encoder's outputs, and append a linear layer followed by softmax for prediction. Let $\hat{T}$ denote a masked text, and $\boldsymbol{p}^j(V, \hat{T})$ denote the predicted probability of the masked token t_j. We minimize the cross-entropy loss:

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{t_j \sim \hat{T}; (V, \hat{T}) \sim D} \text{H}(\boldsymbol{y}^j, \boldsymbol{p}^j(V, \hat{T})) \tag{7}$$

where $\boldsymbol{y}^j$ is a one-hot distribution in which the ground-truth token t_j has the probability of one.

Finally, the pre-training objective of X-VLM is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{mlm}} \tag{8}$$

4. Experiment

4.1. Pre-training Datasets

Table 1. Statistics of the pre-training datasets. See Appendix A.1 for detailed statistics of object and region annotations.

	Dataset	# Images	# Captions	# Ann
4M	COCO	0.11M	0.55M	0.45M
	VG	0.10M	-	5.7M
	SBU	0.86M	0.86M	-
	CC-3M	2.9M	2.9M	-
16M	4M	4.0M	5.1M	6.2M
	Objects365	0.58M	-	2.0M
	OpenImages	1.7M	-	4.2M
	CC-12M	11.1M	11.1M	-

We compare X-VLM with existing approaches at two settings, as listed in Table 1. We refer to them as the 4M setting and 16M setting respectively. Following UNITER (Chen et al., 2020) and other existing work, we prepare our pre-training data using two in-domain datasets, COCO (Lin et al., 2014) and Visual Genome (VG) (Krishna et al., 2017), and two out-of-domain datasets, SBU Captions (Ordonez et al., 2011) and Conceptual Captions (CC) (Sharma et al., 2018).

In the 4M setting, we utilize image annotations only from COCO and VG, which contain 2.5M object annotations and 3.7M region annotations. Note that BUTD, the most widely used object detector, is trained on the same set of object annotations. The existing methods of only learning image-text alignments also utilize the region annotations of VG under the assumption that region descriptions can describe the whole images. In contrast, we take the object labels as text descriptions of objects, and re-formulate the image annotations so that an image has multiple boxes and each box is associated with a text. The text describes the visual concept in the box, which can be an object, a region, or the image itself.

In the 16M setting, we exploit a much noisier Conceptual 12M dataset (CC-12M) (Changpinyo et al., 2021) following ALBEF (Li et al., 2021a). We additionally exploit Objects365 (Shao et al., 2019) and OpenImages (Kuznetsova et al., 2018) following VinVL (Zhang et al., 2021).

Since most downstream V+L tasks are built on top of COCO and VG, we exclude all images that also appear in the validation and test sets of downstream tasks to avoid information leak. We also exclude all co-occurring Flickr30K (Plummer et al., 2015) images via URL matching, because COCO and VG are from Flickr, and there are some overlaps.

4.2. Implementation Details

The image encoder of X-VLM is vision transformer (Dosovitskiy et al., 2020), which is initialized with Swin Transformer$_{\text{base}}$ (Liu et al., 2021b). The text encoder and the cross-modal encoder consist of six transformer layers respectively. The text encoder is initialized using the first six layers of BERT$_{\text{base}}$ (Devlin et al., 2019), and the cross-modal encoder is initialized using the last six layers. In total, X-VLM has 215.6M parameters for pre-training.

X-VLM takes images of resolution of 224×224 as input. For text input, we set the maximum number of tokens to 30. During fine-tuning, we increase the image resolution to 384×384 and interpolate the positional embeddings of image patches following Dosovitskiy et al. (2020).

We apply mixed precision for pre-training. In the 4M setting, we train the model for 200K steps on 8 NVIDIA A100 GPUs and the batch size is set to 1024, which tasks ~ 3.5 days. In the 16M setting, we train the model on 24 GPUs with a batch size of 3072. We sample the data by making half of the images in a batch containing bounding box annotations. We use the AdamW (Loshchilov & Hutter, 2019) optimizer with a weight decay of 0.02. The learning rate is warmed-up to $1e^{-4}$ from $1e^{-5}$ in the first 2500 steps and decayed to $1e^{-5}$ following a linear schedule.

4.3. Downstream Tasks

We adapt X-VLM to five downstream V+L tasks. We follow the settings in the previous work on fine-tuning (see Appendix A.2). Note that we have cleaned the pre-training datasets to avoid data leaks since downstream V+L tasks have overlaps in images with COCO and Visual Genome.

Image-Text Retrieval There are two subtasks: text retrieval (TR) and image retrieval (IR). We evaluate X-VLM on

Table 2. Image-text retrieval results on MSCOCO and Flickr30K datasets. IR: Image Retrieval and TR: Text Retrieval. We compute Recall@K with K = 1, 5, 10, as the evaluation metric. Zero-shot retrieval results are given in Appendix A.3.

Method	# Params	# Pre-train Images	MSCOCO (5K test set) TR	MSCOCO (5K test set) IR	Flickr30K (1K test set) TR	Flickr30K (1K test set) IR
			R@1/R@5/R@10	R@1/R@5/R@10	R@1/R@5/R@10	R@1/R@5/R@10
UNITER$_{\text{large}}$	300M	4M	65.7 / 88.6 / 93.8	52.9 / 79.9 / 88.0	87.3 / 98.0 / 99.2	75.6 / 94.1 / 96.8
METER-Swin	380M	4M	73.0 / 92.0 / 96.3	54.9 / 81.4 / 89.3	92.4 / 99.0 / 99.5	79.0 / 95.6 / 98.0
ALBEF	210M	4M	73.1 / 91.4 / 96.0	56.8 / 81.5 / 89.2	94.3 / 99.4 / 99.8	82.8 / 96.7 / 98.4
METER-CLIP	380M	4M	76.2 / 93.2 / 96.8	57.1 / 82.7 / 90.1	94.3 / 99.6 / 99.9	82.2 / 96.3 / 98.4
VinVL$_{\text{large}}$	550M	5.6M	75.4 / 92.9 / 96.2	58.8 / 83.5 / 90.3	-	-
ALIGN	490M	1.8B	77.0 / 93.5 / 96.9	59.9 / 83.3 / 89.8	95.3 / 99.8 / 100.0	84.9 / 97.4 / 98.6
ALBEF	210M	14M	77.6 / 94.3 / 97.2	60.7 / 84.3 / 90.5	95.9 / 99.8 / 100.0	85.6 / **97.5 / 98.9**
X-VLM	216M	4M	80.4 / 95.5 / **98.2**	63.1 / 85.7 / **91.6**	96.8 / 99.8 / 100.0	86.1 / 97.4 / 98.7
X-VLM	216M	16M	**81.2 / 95.6 / 98.2**	**63.4 / 85.8** / 91.5	**97.1 / 100.0** / 100.0	**86.9** / 97.3 / 98.7

Table 3. Results on downstream V+L tasks, including visual reasoning (VQA and NLVR2), visual grounding (RefCOCO+), and image caption generation (COCO Caption). RefCOCO+ scores with * are evaluated in the weakly-supervised setting. COCO Captioning scores with $^+$ are models optimized with CIDEr for the second stage of fine-tuning.

Method	VQA		NLVR2		RefCOCO+			COCO Caption	
	test-dev	test-std	dev	test-P	val^d	testA^d	testB^d	BLEU@4	CIDEr
ViLBERT	70.55	70.92	-	-	72.34	78.52	62.61	-	-
VL-BERT	71.16	-	-	-	72.59	78.57	62.30	-	-
VILLA	73.59	73.67	78.39	79.30	76.05	81.65	65.70	-	-
SOHO	73.25	73.47	76.37	77.32	-	-	-	-	-
E2E-VLP	73.25	73.67	77.25	77.96	-	-	-	36.2	117.3
KD-VLP	74.20	74.31	77.36	77.78	-	-	-	-	-
UNITER$_{\text{large}}$	73.82	74.02	79.12	79.98	75.90	81.45	66.70	-	-
ALBEF(4M)	74.54	74.70	80.24	80.50	-	-	-	-	-
ALBEF(14M)	75.84	76.04	82.55	83.14	58.46*	65.89*	46.25*	-	-
METER-Swin	76.43	76.42	82.23	82.47	-	-	-	-	-
VinVL$_{\text{large}}$(5.6M)	76.52	76.60	82.67	83.98	-	-	-	41.0$^+$	**140.9**$^+$
METER-CLIP	77.68	77.64	82.33	83.05	-	-	-	-	-
SimVLM$_{\text{base}}$(1.8B)	77.87	78.14	81.72	81.77	-	-	-	39.0	**134.8**
X-VLM(4M)	78.07	78.09	84.16	84.21	80.17	86.36	71.00	39.8 / **41.3**$^+$	133.1 / 140.8$^+$
X-VLM(16M)	**78.22**	**78.37**	**84.41**	**84.76**	**84.51**	**89.00**	**76.91**	**39.9** / 41.0$^+$	134.0 / 140.3$^+$

MSCOCO and Flickr30K (Plummer et al., 2015) datasets. We adopt the widely used Karpathy split (Karpathy & Li, 2015) for both datasets. We optimize $\mathcal{L}_{\text{cl}}$ and $\mathcal{L}_{\text{match}}$ and fine-tune the model for 10 epochs. In inference, we first compute $s(I, T)$ for all images and texts, and then take the top-k candidates and calculate $\boldsymbol{p}^{\text{match}}(I, T)$ for ranking. Following ALBEF, k is set to 256 for MSCOCO and 128 for Flickr30K.

Visual Question Answering (Goyal et al., 2017) It requires the model to predict an answer given an image and a question. Following the previous work (Cho et al., 2021; Li et al., 2021a), we use a six-layer Transformer decoder to generate answers based on the outputs of the cross-modal encoder. We fine-tune the model for 10 epochs. During inference, we constrain the decoder to only generate from the 3,129 candidate answers to make a fair comparison with existing methods.

Natural Language for Visual Reasoning (NLVR2 (Suhr et al., 2019)) The task lets the model determine whether a text describes the relations between two images. Following ALBEF, we extend the cross-modal encoder to enable reasoning over two images, and perform an additional pre-training step for one epoch using the 4M images: given two images and a text, the model assigns the text to either the first image, the second image, or none of them. Then, we fine-tune the model for 10 epochs.

Visual Grounding The task (RefCOCO+ (Yu et al., 2016)) aims to locate the region in an image that corresponds to a specific text description. Previous approaches formulate grounding as a ranking task by relying on the region proposals provided by pre-trained object detectors (Lu et al., 2019; Su et al., 2020; Chen et al., 2020; Gan et al., 2020). In contrast, X-VLM is able to directly predict the bounding boxes of the target regions given images and text descriptions. We also evaluate X-VLM on a weakly-supervised setting, proposed by ALBEF, in which case only image-text pairs are available, and thus we fine-tune X-VLM using $\mathcal{L}_{\text{cl}}$ and $\mathcal{L}_{\text{match}}$;

Image Captioning The task requires a model to generate textual descriptions of input images. We evaluate X-VLM on the COCO Captioning dataset (Chen et al., 2015). We report BLEU-4 and CIDEr scores on the Karparthy test split. To apply X-VLM for captioning, we do not need to add a decoder. Instead, we simply adapt X-VLM to a multi-modal decoder. Specifically, we train X-VLM with language modeling loss for one epoch on 4M data. Then, we fine-tune it on the COCO Captioning dataset. Additionally, following VinVL, we also report the results after applying CIDEr optimization (Rennie et al., 2017) for the second stage of fine-tuning, which are denoted with $^{+}$.

4.4. Results on Image-Text Retrieval

Table 2 compares X-VLM with SoTA approaches on MSCOCO and Flickr30K, which are based on either object-centric features (i.e. UNITER and VinVL) or overall image features (i.e. ALIGN, METER, and ALBEF). ALIGN (Jia et al., 2021) is a dual-encoder model similar to CLIP (Radford et al., 2021) specially for image-text retrieval tasks, which is trained on in-house 1.8B image-text pairs. Other VLMs, including our approach, for more general purposes, have a cross-modal encoder and thus use the output of the cross-modal encoder for ranking.

Even though existing approaches either have more parameters or more training data, X-VLM under the 4M setting outperforms all the previous methods by a large margin, achieving new SoTA results. Specifically, X-VLM(4M) which learns multi-grained vision language alignments outperforms VinVL which is based on object-centric features. In contrast, ALBEF which learns only image-text alignments outperforms VinVL only when increasing the training data to 14M. Compared to METER-Swin (Dou et al., 2021) which also uses Swin Transformer as the image encoder, X-VLM has better performance. Furthermore, even though X-VLM(4M) has already achieved very high performance on the image-text retrieval tasks, we still obtain improvements on R@1 when increasing the training instances to 16M. Additionally, Appendix A.3 shows that when increasing the training data to 16M, X-VLM obtains substantial improvements on zero-shot image-text retrieval. Moreover, X-VLM also outperforms ALIGN on zero-shot MSCOCO by a large margin.

Additionally, METER provides an empirical study of VLMs and shows that the vision backbone (or parameter initialization) is important for the model performance. From Swin Transformer to CLIP-ViT, METER improves significantly on both retrieval and VQA (Table 2 and 3). We also have some preliminary observations and leave detailed studies of different backbones of X-VLM for future work.

4.5. Results on Visual Reasoning

Table 3 shows experimental results on visual reasoning (VQA and NLVR^{2}). First, though ALBEF(14M) outperforms VinVL on image-text retrieval, the coarse-grained approaches such as SOHO, METER-Swin, and ALBEF, all have worse performances than VinVL in visual reasoning tasks, except that METER-CLIP and SimVLM outperform VinVL on VQA. Besides, VinVL also substantially outperforms previous methods that rely on object detectors to learn both object-level and image-level alignments, such as E2E-VLP and KD-VLP.

Nevertheless, X-VLM(4M) with moderate model size and pre-trained on fewer instances outperforms VinVL. Specifically, X-VLM(4M) achieves absolute improvements of 1.52% on VQA and 0.86% on NLVR2 (average on metrics) over VinVL. Meanwhile, as reported in Li et al. (2021a), X-VLM, which encodes images without an object detection process, enjoys ~ 10 times faster inference speed than VinVL. The results indicate that our approach of X-VLM is both effective and efficient. X-VLM also outperforms SimVLM$_{\text{base}}$ which is pre-trained on in-house 1.8B data, especially on NLVR2.

4.6. Results on Visual Grounding

Table 3 reports the performance of X-VLM on RefCOCO+. X-VLM(4M) achieves absolute improvements of 4.5% compared to UNITER. As aforementioned, previous approaches formulate grounding as a ranking task by relying on the region proposals provided by object detectors. In contrast, X-VLM is able to directly predict the target boxes, which is much simpler and more efficient. Furthermore, X-VLM for general V+L purposes outperforms MDETR (Kamath et al., 2021) specialized for visual grounding tasks. X-VLM(4M) using the same set of image annotations achieves absolute improvements of 1.1% (average on metrics), compared to MDETR.

We also evaluate X-VLM in the weakly-supervised setting, proposed by ALBEF. X-VLM(4M) obtains 68.46/76.53/57.09 for val^{d}/testA^{d}/testB^{d} respectively, achieving an absolute improvement of 10.5% (average on metrics) compared to ALBEF(14M). When increasing pre-training images to 16M, X-VLM obtains 77.26/84.11/67.13.

Figure 3 provides a few examples of images from the test set of RefCOCO+. For the supervised setting, we show the bounding boxes predicted by X-VLM given the text descriptions. For the weakly-supervised setting, following ALBEF, we provide the Grad-CAM visualization, which uses the cross-attention maps in the fourth layer of the cross-modal encoder. The visualization examples show that X-

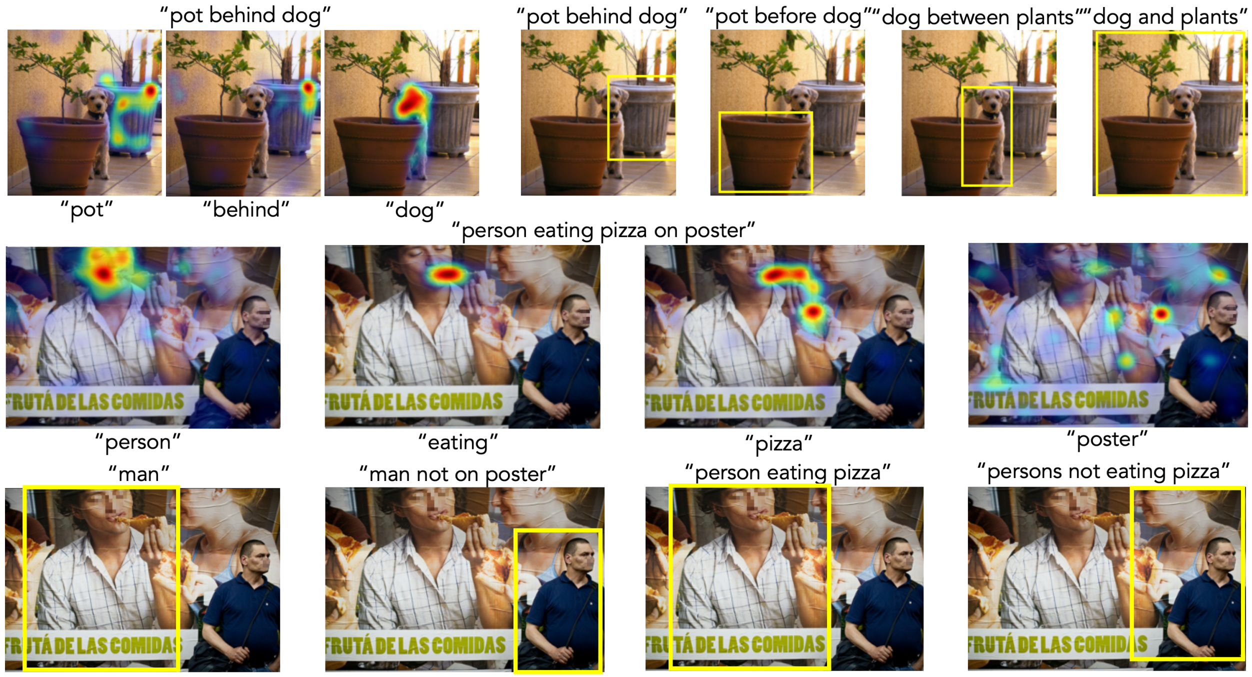

Figure 3. Grad-CAM visualization and bounding box prediction on unseen images. X-VLM predicts correct regions even though the textual descriptions only differ in a single word. X-VLM can also align each word in the text to the corresponding image region. Appendix A.4 gives more examples, showing X-VLM's superior ability of multi-grained vision language alignments.

Table 4. Ablation study results. Models w/o object and w/o region are ablated variants where the model is training without concepts of object and region respectively. Model w/o bbox loss is the variant where bounding box prediction is ablated. Model w/o all represents that all the above components are ablated.

	Meta-Sum	MSCOCO		Flickr30K		VQA	NLVR^2	RefCOCO+	
		TR	IR	TR	IR	test-dev	test-P	testA^d	testB^d
X-VLM	**605.0**	**78.8**	**60.6**	**96.0**	**84.1**	76.20	82.42	72.07	54.84
w/o object	603.5	77.4	60.4	95.0	83.7	75.87	82.10	**73.37**	**55.69**
w/o region	596.0	76.8	60.2	**96.0**	83.6	75.84	82.20	70.73	50.60
w/o bbox loss	594.9	77.5	60.2	95.7	83.5	**76.77**	81.49	69.32	50.38
w/o all	580.6	74.5	57.9	95.6	82.8	74.90	80.70	67.79	46.43

VLM has a strong ability of cross-modal understanding. It successfully predicts the correct regions in images, even though the text descriptions only differ in a single word. Furthermore, X-VLM can align each word in the text to the corresponding image region. We provide more examples in Appendix A.4, showing X-VLM's superior performance in vision language alignment.

4.7. Results on Image Captioning

We show that X-VLM, usually considered as an "encoder-only" model, has comparable performance with SoTA generative methods on image caption generation, as indicated in Table 3. Specifically, X-VLM pre-trained on 16M instances performs similarly to SimVLM which uses not only 1.8B in-house image-text pairs but also a large-scale text corpus. Besides, we observe that CIDEr optimization largely boosts the CIDEr scores. X-VLM in moderate model size also has comparable performance to VinVL$_{\text{large}}$.

4.8. Ablation Study

We also conduct an in-depth ablation study to investigate the role of different components in the X-VLM, as shown in Table 4. All compared model variants are trained on 4M images for 80K steps with a batch size of 3072 to ensure a fair comparison. We use Recall@1 as an evaluation measure in the retrieval tasks and Meta-Sum as a general measure. We report RefCOCO+ evaluation results in the weakly-supervised setting.

First, we evaluate the effectiveness of visual concepts in different granularities, i.e. w/o object and w/o region. The results show that training without either of them hurts the performance, demonstrating the necessity of learning multi-grained alignments. Besides, we can observe that w/o region makes the performance drop more drastically than w/o object. Furthermore, the ablation study shows that bounding box prediction is a critical component of X-VLM, as w/o bbox loss leads to the lowest Meta-Sum. We also report the results of 'w/o all' where all the above components are ablated. Though in the 4M setting, only 210K images have dense annotations, X-VLM can leverage the data to learn multi-grained vision language alignment and substantially improve the performances in the downstream V+L tasks (Meta-Sum from 580.6 to 605.2).

5. Conclusion and Discussion

In this paper, we have proposed X-VLM, a strong and efficient approach to perform multi-grained vision language pre-training. Training of the model is driven by locating visual concepts in the image given the associated texts and aligning texts with relevant visual concepts, where the alignments are in multi-granularity. We have pre-trained X-VLM with 4M and 16M images, which are of moderate size. Also, X-VLM only consists of 216M parameters. These choices are made because we want to make our experiments as "green" (Schwartz et al., 2020; Xu et al., 2021b) as possible and be accessible to a larger group of people. Experiments on downstream V+L tasks, including image-text retrieval, visual reasoning, visual grounding, and image caption generation have shown that X-VLM outperforms the existing methods which could be larger and/or pre-trained on more data. As suggested by the comparison between X-VLM(4M) and X-VLM(16M), adding more pre-training datasets will probably lead to further performance improvements. As for applications, X-VLM has shown better performance in understanding multi-grained vision language alignments. For example, it can generate image captions probably having more object details, which makes it a better choice to help people with disability in vision to understand images. On the other hand, X-VLM in moderate model size is also easier to deploy.

Acknowledgements

We would like to thank Wangchunshu Zhou, Wenguan Huang, and Xiu-jun Li at ByteDance for their generous assistance in data collection and insightful comments in technical discussions.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6077–6086. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00636.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. URL https://arxiv.org/abs/1504.00325.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Cho, J., Lei, J., Tan, H., and Bansal, M. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pp. 1931–1942. PMLR, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Dou, Z.-Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Liu, Z., Zeng, M., et al. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*, 2021. URL https://arxiv.org/abs/2111.02387.

Gan, Z., Chen, Y., Li, L., Zhu, C., Cheng, Y., and Liu, J. Large-scale adversarial training for vision-and-language representation learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Huang, Z., Zeng, Z., Liu, B., Fu, D., and Fu, J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. URL https://arxiv.org/abs/2004.00849.

Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., and Fu, J. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12976–12985, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., and Chen, X. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10267–10276, 2020.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Karpathy, A. and Li, F. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298932. URL https://doi.org/10.1109/CVPR.2015.7298932.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. URL https://arxiv.org/abs/1811.00982.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. 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*, pp. 11336–11344. AAAI Press, 2020a. URL https://aaai.org/ojs/index.php/AAAI/article/view/6795.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021a.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. URL https://arxiv.org/abs/1908.03557.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2592–2607, Online, 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.202. URL https://aclanthology.org/2021.acl-long.202.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020b.

Li, Y., Fan, J., Pan, Y., Yao, T., Lin, W., and Mei, T. Uni-eden: Universal encoder-decoder network by multi-granular vision-language pre-training. *ACM Trans. Multim. Comput. Commun. Appl.*, 18(2):48:1–48:16, 2022. doi: 10.1145/3473140. URL https://doi.org/10.1145/3473140.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, Y., Wu, C., Tseng, S.-y., Lal, V., He, X., and Duan, N. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*, 2021a. URL https://arxiv.org/abs/2109.10504.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021b. URL https://arxiv.org/abs/2103.14030.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13–23, 2019.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 1143–1151, 2011.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303. URL https://doi.org/10.1109/ICCV.2015.303.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Ren, S., He, K., Girshick, R. B., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 91–99, 2015.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1179–1195. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.131. URL https://doi.org/10.1109/CVPR.2017.131.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I. D., and Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 658–666. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00075.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 618–626. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.74. URL https://doi.org/10.1109/ICCV.2017.74.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., and Sun, J. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8429–8438. IEEE, 2019. doi: 10.1109/ICCV.2019.00852. URL https://doi.org/10.1109/ICCV.2019.00852.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. 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)*, pp. 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. 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. URL https://openreview.net/forum?id=SygXPaEYvH.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644.

Tan, H. and Bansal, M. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. URL https://arxiv.org/abs/2108.10904.

Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., and Huang, F. E2E-VLP: End-to-end vision-language pretraining enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 503–513, Online, 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.42. URL https://aclanthology.org/2021.acl-long.42.

Xu, J., Zhou, W., Fu, Z., Zhou, H., and Li, L. A survey on green deep learning. *arXiv preprint arXiv:2111.05193*, 2021b.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85. Springer, 2016.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., and Berg, T. L. Mattnet: Modular attention network for referring expression comprehension. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1307–1315. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00142.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588, 2021.

A. Appendix

A.1. Statistics of Object and Region Annotations

Table 5. Statistics of annotations used in the pre-training.

Dataset	# Images	# Captions	# Objects	# Regions
COCO	0.11M	0.55M	0.45M	-
VG	0.10M	-	2.0M	3.7M
Objects365	0.58M	-	2.0M	-
OpenImages	1.7M	-	4.2M	-

Table 5 gives statistics of object and region annotations of each dataset. Only the Visual Genome dataset contains region annotations. Besides, the OpenImages dataset offers some relationship annotations, indicating pairs of objects in particular relations (e.g. ”woman playing guitar”, ”beer on table”), object properties (e.g. ”table is wooden”), and human actions (e.g. ”woman is jumping”), which can also be viewed as region annotations.

Note that we filtered out some samples because of: 1) invalid annotations (e.g. negative values for bounding boxes or boxes being outside of the images); 2) boxes being too small ($< 1\%$); 3) highly overlapped textual descriptions of regions ($> 75\%$), etc. After pre-processing, we keep: for example, COCO objects 446,873 (from 859,999), VG objects 2,043,927 (from 3,802,349), VG regions 3,699,598 (from 5,402,953).

A.2. Implementation Details of Downstream Tasks

We follow the settings in existing methods for fine-tuning. We describe how we implement fine-tuning on the downstream V+L tasks, and we also provide our fine-tuning scripts for more details. Note that we have cleaned our pre-training datasets to avoid data leaks since downstream V+L tasks have overlaps in images with COCO and Visual Genome.

Image-Text Retrieval We evaluate X-VLM on MSCOCO and Flickr30K (Plummer et al., 2015) benchmarks. We adopt the widely used Karpathy split (Karpathy & Li, 2015) for both datasets. We optimize $\mathcal{L}_{\text{cl}}$ and $\mathcal{L}_{\text{match}}$ for fine-tuning. Since there are multiple ground-truth texts associated with each image in the datasets, we change the ground-truth similarity of contrastive learning, $\boldsymbol{y}^{\text{v2t}}(I)$ and $\boldsymbol{y}^{\text{t2v}}(T)$, to consider multiple positives, where each positive example has a probability of $\frac{1}{\#\text{positives}}$. We fine-tune the model for 10 epochs. During inference, we first compute $s(I, T)$ for all images and texts. Then we take the top-k candidates and pass them into the cross-modal encoder to calculate $\boldsymbol{p}^{\text{match}}(I, T)$ for ranking. Following ALBEF, k is set to 256 for MSCOCO and 128 for Flickr30K.

Visual Question Answering (VQA (Goyal et al., 2017)) Following existing methods (Tan & Bansal, 2019; Chen et al., 2020; Li et al., 2021a), we use both train and validation sets for training, and include additional question-answer pairs from Visual Genome. The VQA model contains a 6-layer transformer-based decoder to generate answers based on the outputs of the cross-modal encoder following previous work (Cho et al., 2021; Li et al., 2021a). The decoder is initialized using the pre-trained weights from the cross-modal encoder. Then, the model is fine-tuned by optimizing the auto-regressive loss for 10 epochs. During inference, we constrain the decoder to only generate from the 3,129 candidate answers [3] to make a fair comparison with existing methods.

Natural Language for Visual Reasoning (NLVR2 (Suhr et al., 2019)) Since the task asks the model to distinguish whether a text describes a pair of images, we follow ALBEF to extend the cross-modal encoder to enable reasoning over two images. We also perform an additional pre-training step for 1 epoch using the 4M images: given a pair of images and a text, the model needs to assign the text to either the first image, the second image, or none of them. Then, we fine-tune the model for 10 epochs.

Visual Grounding The task aims to locate the region in an image that corresponds to a specific text description (RefCOCO+ (Yu et al., 2016)). We evaluate our approach in both supervised and weakly-supervised settings. The latter is proposed by ALBEF. In the supervised setting with bounding box annotations, we perform an additional pre-training step for one epoch using $\mathcal{L}_{\text{bbox}}$ only. Then, we fine-tune the model for 10 epochs. In the weakly-supervised setting where only image-text pairs are available, we fine-tune the model using $\mathcal{L}_{\text{cl}}$ and $\mathcal{L}_{\text{match}}$ for 5 epochs. During inference, following ALBEF, we apply Grad-CAM (Selvaraju et al., 2017) to acquire heatmaps and use them to rank the detected proposals

[3] There is a NULL answer. Thus, the actual number of candidate answers is 3,128.

provided by (Yu et al., 2018).

Image Captioning The task requires a model to generate textual descriptions of input images. We evaluate X-VLM on the COCO Captioning dataset (Chen et al., 2015). We report BLEU-4 and CIDEr scores on the Karparthy test split. To apply X-VLM for captioning, we do not need to add a decoder. Instead, we simply adapt X-VLM to a multi-modal decoder. Specifically, we train X-VLM with language modeling loss for one epoch on 4M data. Then, we fine-tune it on the COCO Captioning dataset with naive cross-entropy loss for five epochs. Additionally, following VinVL, we also report the results after applying CIDEr optimization (Rennie et al., 2017) for the second stage of fine-tuning which takes another five epochs.

A.3. Zero-Shot Image-Text Retrieval Results

Table 6. Zero-shot results on MSCOCO and Flickr30K datasets. IR: Image Retrieval and TR: Text Retrieval.

Method	# Params	# Pre-train Images	MSCOCO (5K test set)		Flickr30K (1K test set)	
			TR	IR	TR	IR
			R@1/R@5/R@10	R@1/R@5/R@10	R@1/R@5/R@10	R@1/R@5/R@10
CLIP	$\sim$100M	400M	58.4 / 81.5 / 88.1	37.8 / 62.4 / 72.2	88.0 / **98.7** / 99.4	68.7 / 90.6 / 95.2
ALIGN	490M	1.8B	58.6 / 83.0 / 89.7	45.6 / 69.8 / 78.6	**88.6 / 98.7 / 99.7**	**75.7** / 93.8 / 96.8
X-VLM	216M	4M	70.8 / 92.1 / 96.5	55.6 / 82.7 / 90.0	85.3 / 97.8 / 99.6	71.9 / 93.3 / 96.4
X-VLM	216M	16M	**71.6 / 93.1 / 97.0**	**56.1 / 83.0 / 89.8**	87.7 / 98.6 / 99.6	74.9 / **94.4 / 97.1**

Table 6 shows zero-shot image-text retrieval results and compares X-VLM with the dual encoder SoTAs (CLIP and ALIGN) which are pre-trained using only the retrieval objective. We can observe that though X-VLM is pre-trained using the combination of different objectives, it still has very competitive results on zero-shot retrieval tasks.

A.4. Case Study

Figure 4 and 5 provide visualizations of some images from the test set of RefCOCO+. We show the bounding boxes predicted by X-VLM given the text descriptions. For the weakly-supervised setting, we provide the Grad-CAM visualization which uses the cross-attention maps in the fourth layer of the cross-modal encoder. We can observe that in both settings X-VLM can predict correct regions even though the textual descriptions only differ in a single word. X-VLM can also align each word in the text to the corresponding image region, showing X-VLM's superior ability of multi-grained vision language alignments.

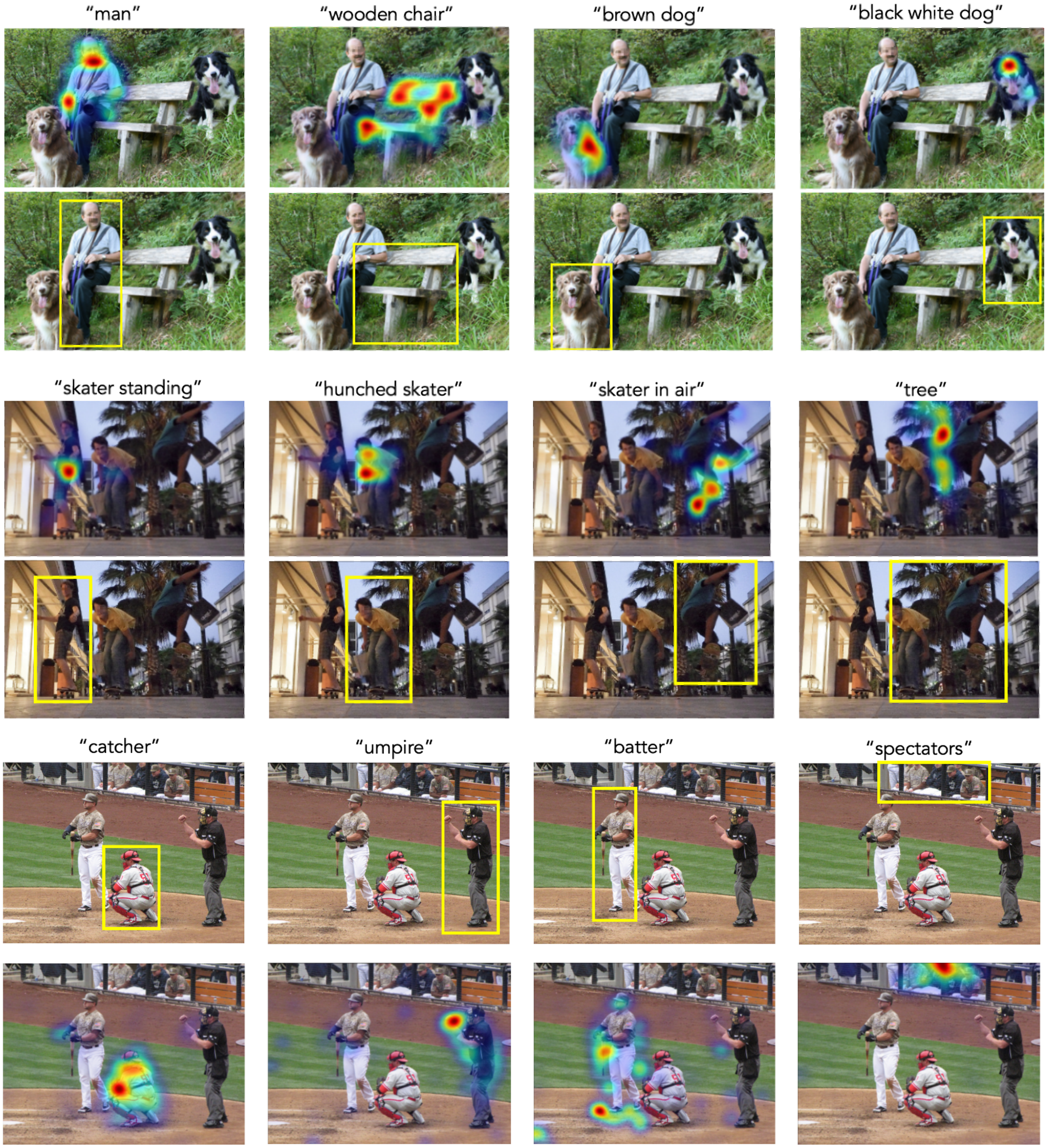

Figure 4. Locating visual concepts in unseen images given text descriptions. Since Grad-CAM gives visualizations each corresponds to an individual word, we only show the visualization of the subject word, e.g. ”dog” for ”brown dog”.

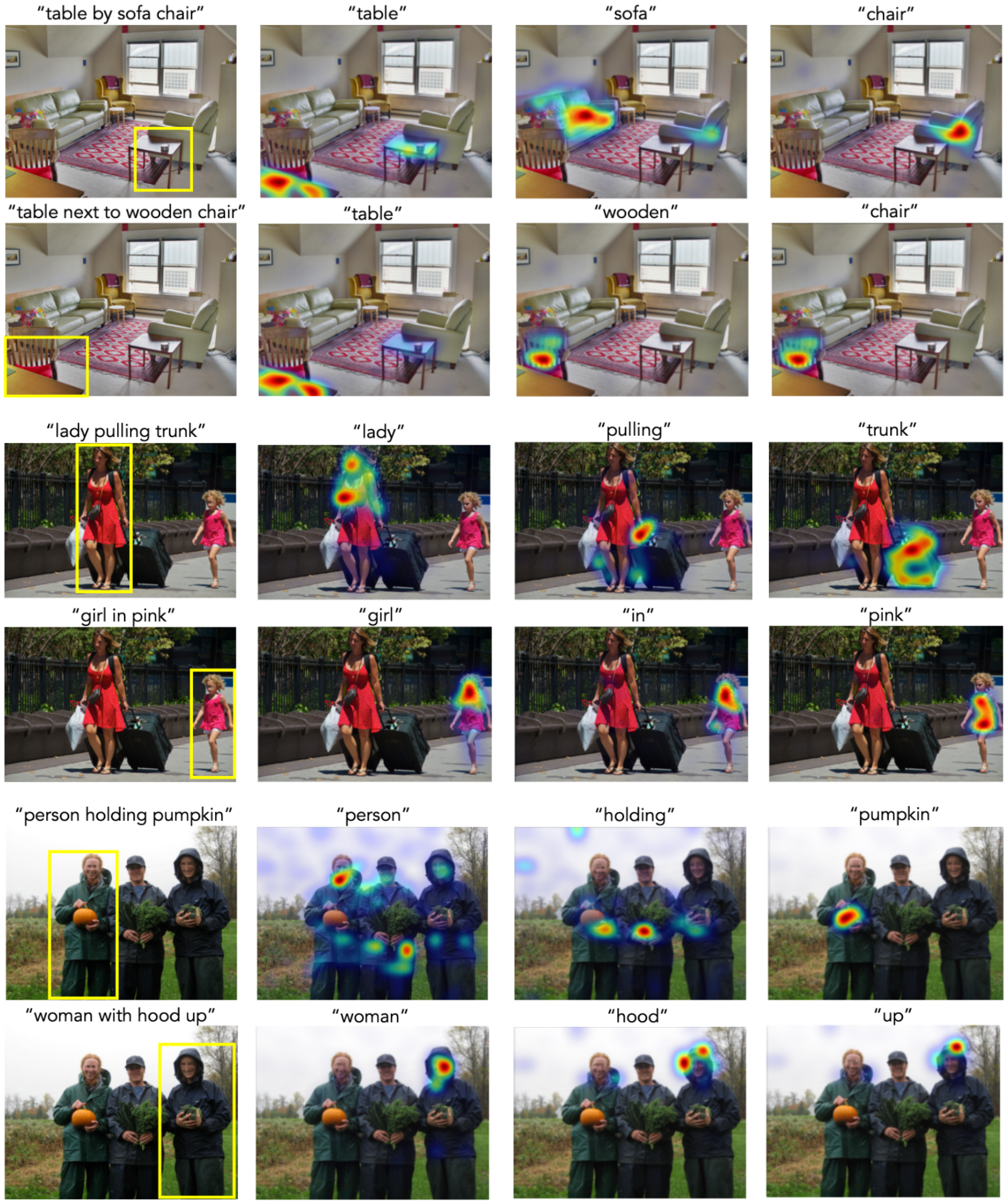

Figure 5. Bounding box prediction and per-word visualization on unseen images. It shows that X-VLM can also align concepts like "pulling" and "holding" to the corresponding regions in the images.