# GLIPv2: Unifying Localization and VL Understanding

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

We present GLIPv2, a grounded VL understanding model, that serves both localization tasks (e.g., object detection, instance segmentation) and Vision-Language (VL) understanding tasks (e.g., VQA, image captioning). GLIPv2 elegantly unifies localization pre-training and Vision-Language Pre-training (VLP) with three pre-training tasks: phrase grounding as a VL reformulation of the detection task, region-word contrastive learning as a novel region-word level contrastive learning task and the masked language modeling. This unification not only simplifies the previous multi-stage VLP procedure, but also achieves mutual benefits between localization and understanding tasks. Experimental results show that a single GLIPv2 model (all model weights are shared) achieves near SoTA performance on various localization and understanding tasks. The model also shows (1) strong zero-shot and few-shot adaption performance on open-vocabulary object detection tasks and (2) superior grounding capability on VL understanding tasks.

## 1 Introduction

Recently, a general interest arises in building general-purpose vision systems (21; 24; 55), also called vision foundation models (6; 56), that solve various vision tasks simultaneously, such as image classification (30), object detection (39), and Visual-Language (VL) understanding (3; 11; 27). Of particular interest, is the unification between *localization* tasks (e.g., object detection (39) and segmentation (8; 20)) and VL *understanding* tasks (e.g., VQA (3) and image captioning (11)). Localization pre-training benefits VL tasks (1; 58), and the "localization->VLP" two-stage pre-training procedure (41; 48; 13; 47; 34; 32; 60; 37; 35) is the common practice in VL community. A long-standing challenge is unification of localization and understanding, which aims at *mutual* benefit between these two kinds of tasks, simplified pre-training procedure and reduced pre-training cost.

However, these two kinds of tasks appear to be dramatically different: localization tasks are vision-only and require fine-grained output (e.g., bounding boxes or pixel masks), while VL understanding tasks emphasize fusion between two modalities and require high-level semantic outputs (e.g., answers or captions). (21; 24; 55) have made early attempts at unifying these tasks in a straightforward multi-task manner, where a low-level visual encoder is shared across tasks and two separate high-level branches are designed for localization and VL understanding, respectively. The localization tasks are still vision-only and does not benefit from the rich semantics in vision-language data. As a result, such unified models see marginal mutual benefit or even performance degradation (24) when compared with task-specific models.

In this paper, we identify "VL grounding" as a "meta"-capability for localization and understanding capabilities. VL grounding involves not only *understanding* an input sentence but also *localizing* the mentioned entities in the image (see an example in Figure 1). We build a **grounded VL understanding** model (GLIPv2) as a unified model for localization and VL understanding tasks.

**Localization + VL understanding = grounded VL understanding**. Localization tasks involve both localization and semantic classification, where classification can be cast as a VL understanding prob-

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

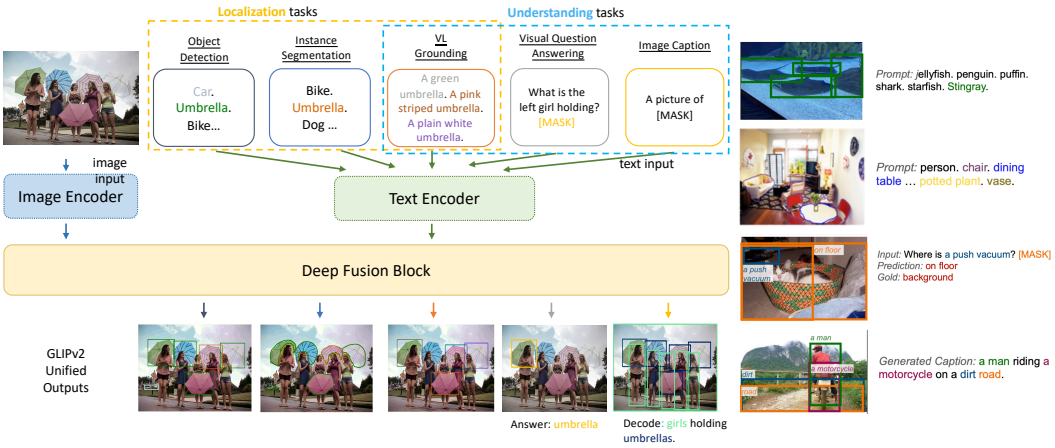


Figure 1: Left: GLIPv2, a pre-trained grounded VL understanding model, unifies various localization and VL understanding tasks. These two kinds of tasks mutually benefit each other, and enables new capabilities such as language-guided detection/segmentation and grounded VQA/captioning. Right: Additional examples from ODinW (detection), LVIS (segmentation), VQA, COCO Captioning.

lem using the *classification-to-matching* trick (Section 3.1). Therefore, we reformulate localization tasks as VL grounding tasks, in which the language input is a synthesized sentence as concatenation of category names (36). Localization data are turned into VL grounding data, accordingly. The massive VL understanding data (image-text pairs) can be easily turned into VL grounding data in a self-training manner (36). Therefore, GLIPv2 has a unified pre-training process: all task data are turned into grounding data and GLIPv2 is pre-trained to perform grounded VL understanding.

**A stronger VL grounding task: inter-image region-word contrastive learning**. GLIP (36) proposes the phrase grounding task as its pre-training task, which we argue is an easy task and does not fully utilize data information. For example, in the VL grounding task in Figure 1, the phrase grounding task only requires the model to match a given image region to one of the three phrases in the text input, i.e., "green, pink striped or plain white umbrella?". This 1-in-3 choice is very easy, only requires color understanding, but loses lots of information in this grounding data: the umbrellas are not any other colors, like black, yellow, etc; objects in those regions are umbrellas but not any other categories, like car, bike, etc. From a contrastive learning view, this phrase grounding task only has 2 negatives. More negatives can be created from this annotation, and thus enable stronger contrastive learning. In GLIPv2, we introduce the novel inter-image region-word contrastive learning task, which leverages phrases from other sentences in the same batch as potential negatives, as another much stronger VL grounding task. This new region-world contrastive loss enables GLIPv2 to learn more discriminative region-word features, and demonstrates improvements over all downstream tasks.

**GLIPv2 achieves mutual benefit between localization and VL understanding**. 1) Experimental results (Table 2) show that a single GLIPv2 model (all model weights are shared) achieves near SoTA performance on various localization and understanding tasks. 2) Thanks to semantic-rich annotations from the image-text data, GLIPv2 shows superior zero-shot and few-shot transfer learning ability to open-world object detection and instance segmentation tasks, evaluated on the LVIS dataset and the "Object Detection in the Wild (ODinW)" benchmark. 3) GLIPv2 enables language-guided detection and segmentation ability, and achieves new SoTA performance on the Flick30K-entities phrase grounding and PhraseCut referring image segmentation tasks. 4) Inherently a grounding model, GLIPv2 leads to VL understanding models with strong grounding ability, which are self-explainable and easy to debug. For example, GLIPv2, when GLIPv2 is finetuned on VQA, it can answer questions while localizing mentioned entities (see Figure 1 and Section 4.4).

## 2 Related Work

**Localization models.** Traditionally, localization tasks such as object detection and segmentation are single-modality and output bounding boxes or pixel masks (44; 38; 23; 14; 45; 10; 9). One

challenge of these single-modality models lies in generalization to rare and novel concepts: it is hard to collect localization data that cover many rare categories (20). A long line of research focuses on this generalization problem, under the name of zero-shot (4; 61; 7; 62), weakly-supervised (18; 5; 51), or open-vocabulary (57; 19) localization. Built upon MDETR (25) and GLIP (36), GLIPv2 converts localization tasks into a grounded vision-language task using the classification-to-matching trick (Section 3). Thus GLIPv2 can learn from the semantic-rich vision-language data and shows strong performance on open-vocabulary localization tasks. **Vision-language understanding models.** Vision-language (VL) understanding tasks such as VQA (3), image captioning (11), and image-text retrieval (26) involve understanding visual semantics and how they are expressed in natural language. Many VL models (e.g., BUTD) (2; 58) rely on a pre-trained localization model as their visual encoder; the downside is the pro-longed "localization->VLP" pre-training pipeline (41; 48; 13; 47; 34; 32; 60; 37; 35). In contrast, GLIPv2 simplifies the pre-training pipeline and enables *grounded* VL understanding for better interpretability (Section 4.4).

**Unifying localization and understanding.** (21; 24; 55) made pioneering efforts in unifying localization and understanding. However, localization tasks are still treated as a single-modality task while VL tasks involve two modalities. The unification is achieved via straightfoward multi-tasking: a low-level visual encoder is shared across tasks and two separate branches are designed for localization and VL understanding. Such unified models do not bring evident mutual benefit and often underperform task-specific models. In contrast, GLIPv2 identifies grounded VL understanding as a meta-task for localization and understanding. The task unification brings architecture unification: the unified grounded VL understanding model empowers a localization branch with VL capacity, arriving a unified branch that excels at both tasks. **GLIPv2 vs GLIP.** 1) GLIP shows grounded pre-training improves localization. GLIPv2 further shows grounded pre-training improves VL understanding, and thus leads to a unified model for localization and VL understanding. 2) GLIPv2 introduces the inter-image region-word contrastive loss, which is another and stronger grounding task than the pre-training task in GLIP. The proposed loss can be viewed as a region-word level generalization of the prevalent image-level contrastive learning (33; 43; 54). 3) GLIPv2 outperforms GLIP on all benchmarks with the same pre-training data.

# 3 GLIPv2: Unifying Localization and VL Understanding

Based on the reformulation of object detection as a generalized phrase grounding task in GLIP (36), we unify both localization tasks and VL understanding tasks as grounded vision-language tasks. A grounded vision-language task takes both image and text as input, and output region-level understanding results (e.g., detection, segmentation) and/or image-level understanding results with associated grounding/localization information (e.g., VQA, image captioning). We will present the unified grounded VL formulation and architecture in Section 3.1, the pre-training losses in Section 3.2, and transfer to downstream tasks in Section 3.3.

## 3.1 A Unified VL Formulation and Architecture

At the center of GLIPv2's unified formulation is the *classification-to-matching* trick, which reformulates any *task-specific fixed-vocab classification problem as an task-agnostic open-vocabulary vision-language matching* problem. The best example is the reformulation of image classification as image-text matching in CLIP (43), which enables the model to directly learn from raw image-text data, and achieves strong zero-shot results on open-vocabulary classification tasks. In GLIPv2, we replace every semantic classification linear layer in traditional single-modality vision models with a vision-language matching dot-product layer.

As illustrated in Figure 1, GLIPv2's unified VL architecture is based on the generic architecture which we term Architecture $\mathbf{\Pi}$. It consists of a dual encoder, denoted as $\text{Enc}_V$ and $\text{Enc}_L$, and a fusion encoder, denoted as $\text{Enc}_{VL}$. The model takes an image-text pair $(\text{Img}, \text{Text})$ as input, and extract visual and text features as below:

$$\mathring{O} = \text{Enc}_V(\text{Img}), \quad \mathring{P} = \text{Enc}_L(\text{Text}), \quad O, P = \text{Enc}_{VL}(\mathring{O}, \mathring{P}), \tag{1}$$

where $(\mathring{O}, \mathring{P})$ and $(O, P)$ denote the image/text features *before* and *after* VL fusion, respectively.

**Vision-Language understanding tasks.** Arch $\mathbf{\Pi}$ is the most popular model architecture for VL understanding tasks. Given the cross-modality fused representations $O$ and $P$, it is straightforward to

add light-weight task-specific heads for various VL tasks. For example, GLIPv2 adds a two-layer MLP on top of text features $P$ as the masked language modeling (MLM) head, to perform the MLM pre-training. We provide model details of VQA and image captioning in Section 3.3.

**(Language-guided) object detection and phrase grounding.** Following GLIP (36), GLIPv2 uses the classification-to-matching trick to unify detection and grounding. More specifically, for detection, we simply replace the class logits $S_{\text{cls}} = OW^T$, where $W$ is the weight matrix of the box classier, with a task-agnostic region-word similarity logits $S_{\text{ground}} = OP^T$, where text features $P$ are label embeddings from a task-agnostic language encoder. As shown in Figure 1, object detection and phrase grounding share the same input/output format and model architecture. See GLIP (36) for more details. Their only difference is the input text format: (1) for object detection, the text input is a string of concatenated candidate object labels ; (2) for phrase grounding, the text input is a natural language sentence. We refer to GLIP (36) for more details.

**(Language-guided) instance segmentation and referring image segmentation.** Given the object detection results, an instance segmentation head is added to classify each pixel within the box into a semantic class. Again, GLIPv2 uses the classification-to-matching trick, to produce a unified instance segmentation head for the standard instance segmentation tasks and the referring image segmentation tasks, and leverage both tyeps of data for its pre-training. This classification-to-matching trick can be also applicable to many other semantic classification heads in single modality CV models (e.g., semantic segmentation), and thus transfers them to language-guided CV models.

### 3.2 GLIPv2 Pre-training

The GLIPv2 is pre-trained with three pre-training losses: phrase grounding loss $\mathcal{L}_{\text{ground}}$ from a vision-language reformulation of the object detection task, region-word contrastive loss $\mathcal{L}_{\text{inter}}$ from a novel region-word level contrastive learning task, and the standard masked language modeling loss $\mathcal{L}_{\text{mlm}}$ proposed in BERT (16).

$$\mathcal{L}_{\text{GLIPv2}} = \underbrace{\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}}}_{\mathcal{L}_{\text{ground}}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{mlm}} \tag{2}$$

Similar to losses in detection tasks, the grounding loss $\mathcal{L}_{\text{ground}}$ has two parts: the localization loss $\mathcal{L}_{\text{loc}}$ trains localization heads with bounding-box supervision, e.g., RPN loss, box regression loss and/or centerness loss (49); the intra-image region-word alignment loss $\mathcal{L}_{\text{intra}}$ is essentially the semantic classification/retrieval loss for each region.

**Intra-image region-word alignment loss.** Given one image-text pair $(\text{Img}, \text{Text})$, we obtain the image and text features *after* cross-modality fusion $O$ and $P$. The Intra-image region-word alignment loss is computed by

$$\mathcal{L}_{\text{intra}} = loss(OP^T; T), \tag{3}$$

where $OP^T$ is the similarity score between image regions and word tokens, and $T$ is the target affinity matrix determined by the ground-truth annotations. The loss function $loss$ is typically a cross-entropy loss for two-stage detectors (45) and a focal loss (38) for one-stage detectors.

However, as discussed in Section 1, this intra-image region-word contrastive learning is rather weak in the sense of contrastive learning, due to limited number of phrases that can one caption can contain. GLIP (36) alleviates this problem by appending a few negative sentences to form a longer text input with more (negative) phrases. However, constrained by the maximal length of text tokens (256 in GLIP and GLIPv2), only a few negative sentences can be added and the number of negative phrases remains at the order of 10's. This small-negative-example problem also exists in detection data (36) when the input text cannot include all class names in a detection dataset, e.g., Objects365.

**Inter-image region-word contrastive loss.** In GLIPv2, we propose to use the phrases from other image-text pairs in the same batch as negative examples, which effectively increase the number of negative examples to the order of 1000's, with nearly negligible additional computational cost.

As in (1), given a batch of image-text pairs $(\text{Img}^i, \text{Text}^i)_{i=1}^B$ and their ground-truth annotations $(T^i)_{i=1}^B$, the model produces the image and text features *before* and *after* VL fusion, denoted as $(\mathring{O}^i, \mathring{P}^i)_{i=1}^B$ and $(O^i, P^i)_{i=1}^B$, respectively. Then as illustrated in Figure 2 (Left), a batch-wise similarity matrix $S_{\text{ground}}^{\text{batch}}$ and a batch-wise target affinity matrix $T^{\text{batch}}$ are constructed by considering

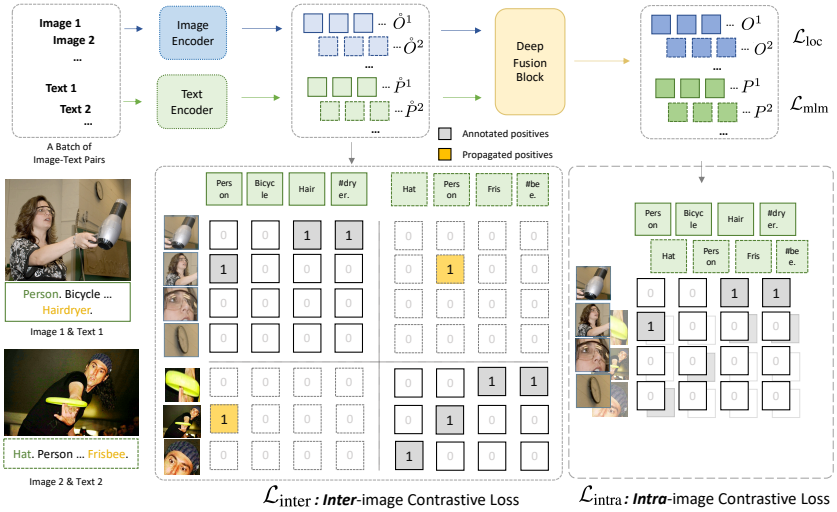


Figure 2: GLIPv2 pre-training losses: the intra-image alignment loss $\mathcal{L}_{\text{intra}}$ (right) takes features after VL fusion and compute loss over region-word pairs within each image-text pair; the inter-image contrastive loss (left) $\mathcal{L}_{\text{inter}}$ takes features before VL fusion and compute loss over all region-word pairs across a batch of image-text pairs. Label propagation is used to determine the off-diagonal blocks of the $\mathcal{L}_{\text{inter}}$ target matrix (4).

all the image regions and text phrases across this batch. Their $(i, j)$'th blocks are obtained as below:

$$S^{\text{batch}}_{\text{ground}}[i,j] = \mathring{O}^i (\mathring{P}^j)^T, \quad T^{\text{batch}}[i,j] = \begin{cases} T^i, & \text{if } i = j \\ \text{obtained by label propagation}, & \text{otherwise.} \end{cases} \tag{4}$$

The inter-image region-word contrastive loss is then defined as the standard bi-directional contrastive loss applied on all image regions and phrases in this batch:

$$\mathcal{L}_{\text{inter}} = \text{cross_entropy_loss}(S^{\text{batch}}_{\text{ground}}, T^{\text{batch}}, \text{axis} = 0) + \text{cross_entropy_loss}(S^{\text{batch}}_{\text{ground}}, T^{\text{batch}}, \text{axis} = 1). \tag{5}$$

Compared with that in the inter-image contrastive loss (3), the number of negatives are multiplied by batch size $B$ in this inter-image contrastive loss (5). We elaborate two important details in (4). (1) GLIPv2 uses the image text features $(\mathring{O}^i, \mathring{P}^i)_{i=1}^B$ before VL fusion, *not* $(O^i, P^i)_{i=1}^B$ after VL fusion, to compute the batch-wise similarity matrix in the inter-image contrastive loss (4). Otherwise, the image and text features after VL fusion would have seen the paired information (1), and thus the model can easy rule out the negatives from mis-paired images/texts. (2) We cannot simply assign all regions and texts coming from unpaired image-text as negative pair, as done in standard contrastive loss in CLIP (43). Instead, we determine the off-diagonal blocks in the target affinity matrix $T^{\text{batch}}$ by *label propagation*. For example, as illustrated in Figure 2 (Left), if a region is annotated as "person", it should be a positive pair with all "person" phrases in detection-type texts. We do not propagate positives to grounding-type texts (natural sentences) because phrases in sentences carry contexts that are unique to that image-sentence pair.

**Pre-training with both detection and paired-image-text data.** GLIPv2 pre-training data is in the image-text-target triplet format $(\text{Img}, \text{Text}, T)$, where the target affinity matrix $T$ contains the box-label localization annotations. We also use massive image-text pair data $(\text{Img}, \text{Text})$ to pre-train GLIPv2, by generating grounding boxes $\hat{T}$ for phrases in the text with the GLIP pre-trained model from (36). The human-annotated OD/grounding data provides high-fidelity localization supervision, while the massive image-text data greatly improves the concept diversity for GLIPv2.

**Second-stage pre-training of the segmentation head.** GLIPv2 performs a second-stage pre-training of the language-guided segmentation head on both instance segmentation and image referring segmentation data, while fixing all other parts of the model.

### 3.3 Transfer GLIPv2 to Localization and VL Tasks

We introduce two ways to easily transfer GLIPv2 to various downstream tasks. In addition, GLIPv2 can perform conventional VL tasks (e.g., VQA) along with localization, effectively making every task we consider a "grounded VL understanding" task.

**One model architecture for all.** GLIPv2 can be transferred to downstream tasks by fine-tuning the model with an (optional) task-specific head. 1) For *detection and segmentation* tasks, no task-specific head is needed as the pre-training architecture can inherently perform detection and segmentation. 2) For *VL* tasks: for VQA, a classification head is added on top of the hidden representation of the start-of-sequence token; for caption generation, we train with a unidirectional language modeling loss, which maximizes the likelihood of the next word given context. We use a unidirectional attention mask and prevent the image part from attending to the text in the fusion layers.

**One set of weight for all.** There is a growing interest in developing models that can be transferred to various task while only changing the least amount of parameters to save training time and storage cost (46; 31). Following GLIP, GLIPv2 can be transferred to localization tasks in a *zero-shot* or a *prompt-tuning* setting ( Section 4.2). One single GLIPv2 model can serve various tasks, where each task only keep few or no parameters. Of particular interest is the prompt tuning setting. For a certain localization task, the text prompt is the same for all input images; thus we could directly tune $\mathring{P}$, a small prompt embedding matrix, to adapt GLIPv2 to new tasks. Prompt tuning in a deep-fused model such as GLIPv2 is different from the conventional linear probing/prompt tuning setting (52; 43; 59) in shallow-interacting vision models such as CLIP. The latter can also be viewed as only tuning a small prompt/softmax embedding $P$; however, tuning $P$ only affects the very last layer of the model while the visual representation is still frozen. In contrast, GLIP/GLIPv2's visual representation is conditioned on the prompt embedding $\mathring{P}$; tuning $\mathring{P}$ changes the text, visual, as well as the fused embeddings. As a result, prompt tuning in GLIPv2 is highly effective, often matching the performance of fine-tuning (see Table 2). This is in contrast to the common observation in CV that linear probing lags behind fine-tuning by a large gap (22).

**Grounded VL understanding.** GLIPv2 also enables grounded VL understanding, where we retain the ability to perform grounding when fine-tuning the model to a downstream VL task. This increases the interpretability of the model. Specifically, we first turn the VL data of the downstream task into grounded VL data using a pre-trained GLIP model. Then we train the model with both downstream task head and grounding head. For VQA, the model is trained to predict the answer and ground entities in the question as well as the implied entity in the answer; for captioning, the model is trained to predict the next word given the context and ground the current decoded word. By tuning localization tasks into a grounded VL task and augmenting VL tasks with grounding ability, we effectively turn every task into a grounded VL understanding task (see examples in Figure **??**).

## 4 Experiments

In this section, we show that GLIPv2 serves as a performant and easy-to-deploy general-purpose vision system. 1) **One Model Architecture for All** (Section 4.1). GLIPv2 can be directly fine-tuned to both localization and VL understanding tasks with minimal architecture change. It achieves performance on par with SOTA models with specialized architectures. 2) **One Model Weight for All** (Section 4.2). GLIPv2 can be transferred to localization tasks in a zero-shot manner with zero parameter update; with prompt tuning, a single GLIPv2 model can achieve comparable performance with fully fine-tuned settings on both localization and understanding tasks.

Following GLIP (36), we adopt Swin Transformer (40) as the image encoder $\text{Enc}_V$, text transformers (50; 43) as the text encoder $\text{Enc}_L$, Dynamic Head (15) with language-aware deep fusion (36) as the fusion encoder $\text{Enc}_{VL}$, and Hourglass network (42) as instance segmentation head feature extractor. We train GLIPv2 at three scales: GLIPv2-T, GLIPv2-B, and GLIPv2-H.

**GLIPv2-T** has the same model config and initialization as GLIP-T: Swin-Tiny and BERT-Base as the dual encoder. The model is pre-trained on the following data: 1) O365, 2) GoldG as in GLIP-T (C), and 3) Cap4M, 4M image-text pairs collected from the web with boxes generated by GLIP-T (36). **GLIPv2-B/GLIPv2-H** are based on Swin-Base/Swin-Huge and the pre-layernorm text transformer (17) as dual encoder, and are initialized from the UniCL (54) checkpoints. We observe much stabler training with GPT-type pre-layernorm transformer (17) than BERT-type post-layernorm

| Model | Model Type | COCO-Det (test-dev) | ODinW (test) | LVIS (minival) | COCO-Mask (test-dev) | Flickr30K (test) | PhraseCut (test) | VQA (test-dev / test-std) | Captioning (Karpathy-test) |
|---|---|---|---|---|---|---|---|---|---|
| Mask R-CNN (23) | Localization | 39.8 | - | 33.3 / - | - / 37.1 | - | - | - | - |
| DETR (9) | | 42.0 | - | 17.8 / - | - | - | - | - | - |
| DyHead-T (15) | | 49.7 | 60.8 | - | - | - | - | - | - |
| DyHead-L (15) | | 60.3* | - | - | - | - | - | - | - |
| VisualBERT (34) | Understanding | - | - | - | - | 71.33 | - | 70.8 / 71.0 | - |
| UNITER (12) | | - | - | - | - | - | - | 73.8 / 74.0 | - |
| VinVL (58) | | - | - | - | - | - | - | **76.5 / 76.6** | 130.8 |
| GPV (21) | Localization & Understanding | - | - | - | - | - | - | 62.5 / - | 102.3 |
| UniT (24) | | 42.3 | - | - | - | - | - | 67.6 / - | - |
| MDETR (25) | | - | - | 24.2 / - | - | 84.3 | 53.7 | 70.6 / 70.6 | - |
| Unicorn (55) | | - | - | - | - | 80.4 | - | 69.2 / 69.4 | 119.1 |
| GLIP-T (36) | Localization & Understanding | 55.2 | 64.9 | - | - | 85.7 | - | - | - |
| GLIP-L (36) | | 61.5* | 68.9 | - | - | 87.1 | - | - | - |
| GLIPv2-T (**Ours**) | Localization & Understanding | 55.5 | 66.5 | 50.6 / 41.4 | 53.5 / 42.0 | 86.5 | 59.4 | 71.6 / 71.8 | 122.1 |
| GLIPv2-B (**Ours**) | | 58.8 | 69.4 | 57.3 / 46.2 | 59.0 / 45.8 | 87.5 | **61.3** | 73.1 / 73.3 | 128.5 |
| GLIPv2-H (**Ours**) | | **60.6 (62.4*)** | **70.4** | **59.8 / 48.8** | **59.8 / 48.9** | **87.7** | **61.3** | 74.6 / 74.8 | **131.0** |

Table 1: One model architecture results. For COCO-Det test-dev, * indicates multi-scale evaluation. For LVIS, we report the numbers for both `bbox` and `segm` on minival to avoid data contamination due to the pre-training. For Flickr30K test, we report the metric under `R@1`. For COCO-Mask, we also report both `bbox` and `segm` on test-dev.

transformer. The training data contain: 1) FiveODs (2.78M data) [1]; 2) GoldG as in MDETR (25); and 3) CC15M+SBU, 16M public image-text data with generated boxes by GLIP-L (36). **Segmentation heads** of GLIPv2 models are pre-trained on COCO, LVIS (20) and PhraseCut(53), with all other model parameters are frozen.

**Note** All datasets above were collected by the creators (cited) and consent for any personally identifiable information (PII) was ascertained by the authors where necessary. Due to limited space, we refer to supplementary for details of training recipes and hyper-parameters.

### 4.1 One Model Architecture for All

We compare GLIPv2 to existing object detection and vision-language pre-training methods on a wide range of tasks. We fine-tune the model on 8 different downstream tasks and report the performance in Table 1. We make the following observations.

**GLIPv2 v.s. specialized Localization methods.** GLIPv2 outperforms previous localization models on generalization to both common and rare classes and domains *with a single model architecture and pre-training stage*. *1) OD on common categories (COCO-Det)*, GLIPv2-T achieves 5.8 improvement compared to the standard DyHead-T trained on O365 (55.5 v.s. 49.7). GLIPv2-H reaches 62.4 AP on test-dev, and surpass the performance of the previous SoTA model GLIP-L. *2) OD on rare / unseen categories (LVIS)*, GLIPv2-T outperforms a supervised MDETR on the `bbox` by a great margin (59.8 v.s. 24.2). *3) Generalization to diverse real-word tasks (ODinw)*, GLIPv2-T (55.5) performs better than original GLIP-T (64.9) on the average of 13 public datasets; GLIPv2-B outperforms GLIP-L by 0.5 AP. *4) Instance segmentation (COCO-Mask & PhraseCut)*, for traditional instance segmentation (i.e., COCO-Mask), GLIPv2-H outperforms the well-known Mask R-CNN by a great margin on `segm`. For language-guided segmentation (i.e., PhraseCut), compared to MDETR, GLIPv2-T achieves an improvement of 5.7 mask AP.

**GLIPv2 v.s. specialized VL Understanding methods.** GLIPv2 rivals with SoTA specialized models for VL tasks. *1) For VQA*, GLIPv2 outperforms VisualBERT and UNITER and approaches the previous SoTA model VinVL. *2) For Captioning*, the best GLIPv2 even surpasses VinVL (VinVL and GLIPv2 are not trained with CIDEr optimization).

**GLIPv2 v.s. localization and VL models.** Prior works such GPV, UniT and Unicorn have also explored unifying localization and VL models (see a discussion in Section 2). GLIPv2 outperforms all previous systems on both localization and VL tasks. For the best GLIPv2-H, it outperforms the UniT by a great margin (18.3 AP) on COCO object detection tasks. Meanwhile, it also surpasses UniT's performance on VQA by 6.9 points and GPV's peformance on Image Captioning as well.

[1] Besides O365, it combines with 4 additional OD datasets including COCO (39), OpenImages (28), Visual Genome (29), and ImageNetBoxes (30)

| Model | Direct Evaluation | | | | Prompt Tuning | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | COCO-Mask (minival) | ODinW (test) | LVIS-Det (minival) | Flickr30K (minival) | COCO-Det (test-dev) | ODinW (test) | LVIS (minival) | COCO-Mask (test-dev) | PhraseCut (test) |
| GLIP-T | 46.6/– | 46.5 | 26.0 | 85.7 | – | 46.5 | - | - | - |
| GLIP-L | 49.8/– | 52.1 | 37.3 | 87.1 | 58.8 | 67.9 | - | - | - |
| GLIPv2-T | **47.3**/35.7 | 48.5 | **29.0** | 86.0 | 53.4 (-2.1) | 64.8 (-1.7) | 49.3 / 34.8 (-1.3 / -6.6) | 53.2 / 41.2 (-0.3 / -0.8) | 49.4 |
| GLIPv2-B | 61.9†/43.4 | 54.2 | 48.5 | 87.2 | 59.0 (+0.2) | 67.3 (-2.1) | 56.8 / 41.7 (-0.5 / -4.5) | 58.8 / 44.9 (-0.2 / -0.9) | 55.9 |
| GLIPv2-H | 64.1†/47.4 | **55.5** | 50.1 | **87.7** | **60.2 / 61.9*** (-0.4 / -0.5) | **69.1** (-1.3) | **59.2 / 43.2** (-0.6 / -5.7) | **59.8 / 47.2** (-0.0 / -1.7) | **56.1** |

Table 2: One set of weights results v.s. Original GLIP. * indicates multi-scale evaluation. Numbers in red clearly points out the difference between the prompt tuning and full fine-tuning results (see Table 1). Numbers in gray mean that they are not in *zero-shot* manner. †: these two numbers are artificially high due to some overlap between COCO-minival and VisualGenome-train.

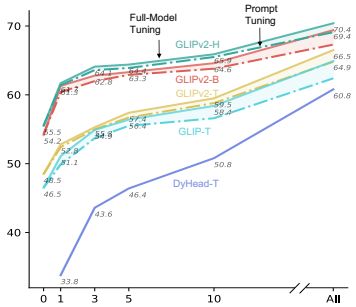


Figure 3: Data efficiency of GLIPv2 on ODinW. X-axis is the amount of task-specific data, from zero-shot to all data. Y-axis is the average AP across 13 datasets.

| Model | Zero-Shot | Prompt Tuning / Fine Tuning | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 3 | 5 | 10 | All |
| DyHead-T O365 (36) | - | - / 33.8 | - / 43.6 | - / 46.4 | - / 50.8 | - / 60.8 |
| $\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}}$ (GLIP-T) | 46.5 | 49.9 / 51.3 | 53.7 / 54.9 | 55.5 / 56.4 | 56.6 / 58.4 | 62.4 / 64.9 |
| $\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}$ | 48.4 | 52.1 / 51.4 | 55.6 / 55.3 | 56.7 / 56.6 | 58.3 / 59.5 | 62.9 / 66.3 |
| $\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{mlm}}$ | 48.5 | 52.4 / 52.8 | 55.6 / 55.6 | 57.4 / 57.4 | 58.8 / 59.7 | 64.8 / 66.5 |

Table 3: Zero-shot, prompt tuning and full fine-tuning performance on ODinW. GLIPv2 models exhibit superior data efficiency.

**Takeaway.** Most notably, GLIPv2 outperforms previous "unified" models (GPV, UniT, MDETR, Unicorn) by a large margin. This is the first time that a single model architecture could achieve near SoTA performance on both localization and understanding. In contrast, in prior work, there exists certain trade-off between localization and understanding: models that aim to achieve high understanding performance tend to have lower localization performance (e.g., UNiT's detection performance is limited to the DETR(9) architecture), as it is not trivial to merge a SoTA localization branch and a SoTA VL branch into a single model.

### 4.2 One Set of Model Parameters for All

GLIPv2 is pre-trained to perform grounding; thus it can be transferred to various localization tasks with changing zero or few parameters. We evaluate GLIPv2 under two such settings: 1) direct evaluation, where we transfer the model "as is" without any parameter change, and 2) prompt tuning, where only the prompt embedding is tuned for specific tasks (Section 3.3).

**Direct evaluation.** The pre-trained GLIPv2 can be directly evaluated on any object detection task (by concatenating the object categories into a text prompt) and visual grounding task, without any further tuning. We evaluate the models on 4 localization tasks: COCO, ODinW, LVIS, and Flickr30 and their results are presented in Table 2. Note that for GLIPv2-B and GLIPv2-H, the training set of Flick30K and LVIS are present in the pre-training data. Thus, reported numbers on these metrics are not *zero-shot* evaluation (we have marked them gray). For all other evaluation results, the models are evaluated in *zero-shot* settings without any further tuning.

*GLIPv2 can be effortlessly transferred to different localization tasks without further tuning.* 1) For *COCO*, GLIPv2-T achieves a zero-shot performance of 47.3 without seeing any COCO training images. This surpasses well-established supervised systems (e.g., Mask R-CNN) and also outperforms GLIP-T by 0.7 AP. 2) For *ODinW*, GLIPv2 also shows strong zero-shot performance. GLIPv2-T (48.5) surpasses the GLIP-T (46.5). Meanwhile, the zero-shot performance of GLIPv2-B and GLIPv2-H even surpasses the 10-shot tuning performance of DyHead-T (to be introduced in Figure 3). 3)

| Row | Model | COCO | ODinW | LVIS | Flickr30K | VQA | Captioning |
|---|---|---|---|---|---|---|---|
| 1 | No pre-train | –/50.6 | –/60.8 | – | – | 64.6 | 111.5 |
| 2 | + $\mathcal{L}_{\text{mlm}}$ | –/48.5 | –/37.4 | – | – | 64.6 | 110.9 |
| 3 | + $\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}}$ | 46.6/55.2 | 46.5/64.9 | 26.0 | 85.7 | 69.4 | 119.7 |
| 4 | + $\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{mlm}}$ | 47.0/55.2 | 47.6/66.2 | 28.5 | 86.5 | 69.8 | 120.7 |
| 5 | + $\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}$ | 47.1/55.4 | 48.4/66.3 | 28.6 | 85.8 | 68.7 | 120.4 |
| 6 | + $\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{mlm}}$ | 47.3/55.5 | 48.5/66.5 | 29.0 | 86.3 | 70.7 | 122.1 |

Table 4: Pre-training losses on Tiny-scale model. Involving intra-image region-word alignment loss $\mathcal{L}_{\text{intra}}$, inter-image region-word contrastive loss $\mathcal{L}_{\text{inter}}$ and MLM loss $\mathcal{L}_{\text{mlm}}$ will benefit both localization and understanding tasks.

| $\mathcal{L}_{\text{inter}}$ | Pre-train Data | COCO | ODinW | LVIS | Flick30K |
|---|---|---|---|---|---|
| ✗ | O365, GoldG | 48.06 | 43.14 | 25.6 | 84.36 |
| ✓ | O365, GoldG | 48.59 | 42.64 | 26.9 | 83.90 |
| ✗ | O365, GoldG, Cap4M | 48.21 | 51.35 | 34.2 | 85.56 |
| ✓ | O365, GoldG, Cap4M | 48.79 | 52.70 | 35.0 | 85.50 |
| ✗ | O365, GoldG, Cap12M | 48.50 | 49.32 | 35.5 | 85.79 |
| ✓ | O365, GoldG, Cap12M | 49.26 | 53.15 | 36.6 | 85.84 |

Table 5: Pre-train data scale up on Base-scale model. Results are reported at the last checkpoint. See supplementary for results at all checkpoints.

| Model | COCO Caption | | | Flickr30K Grounding | | |
|---|---|---|---|---|---|---|
| | B4 | CIDEr | SPICE | R@1 | R@5 | R@10 |
| GLIPv2-T | 36.5 | 119.8 | 21.6 | 80.8 | 94.4 | 96.5 |
| GLIPv2-B | 37.4 | 123.0 | 21.9 | 81.0 | 94.5 | 96.5 |

Table 6: GLIPv2 can perform captioning and grounding at the same time (a.k.a., grounded VL understanding).

For *LVIS*, GLIPv2-T achieves a 3 AP improvement performance compared to the GLIP-T. 4) For *Flickr30K*, GLIPv2-B achieves even higher number (87.2) compared to original GLIP-L (87.1).

**Prompt Tuning.** *Prompt tuning GLIPv2 achieves similar performance as full fine-tuning.* When comparing the performance of each task in Table 1 and 2 at the same time, for GLIPv2, prompt tuning performance almost matches the one model architecture results on localization tasks, without changing any of the grounding model parameters.

### 4.3 GLIPv2 as a Strong Few-Shot Learner

We demonstrate GLIPv2's performance on ODinW datasets with respect to different amounts of training data in Figure 3. The performance improvement between GLIPv2-T and GLIP-T exhibits more superior data efficiency for prompt tuning. We compare with the SoTA detector DyHead-T, pre-trained on Objects365 in Table 3.

### 4.4 Analysis

**Pre-training losses and pretraining data.** We evaluate the effectiveness about the pre-training tasks (i.e., intra-image region-word alignment loss, inter-image region-word contrastive loss, and MLM loss) and pre-trainin data in Table 4 and Table 5. See detailed analysis in supplementary.

**Grounded Vision-Language Understanding** GLIPv2 can be trained to perform a VL task and grounding at the same time (Section 3.3). We denote such an ability as grounded VL understanding. In Figure 1, we showcase grounded predictions of GLIPv2 on VQA and COCO captions. We also conduct quantitative evaluations (Table 6). The model achieves strong performance for both VL understanding (on COCO Caption) and localization (on Flickr30K Grounding). Such an ability to produce high-level semantic outputs (i.e., answers and captions) and supporting localization results is another appealing trait of GLIPv2, as potential users can have a better understanding of the model behaviour. See more examples in the supplementary.

## 5 Conclusion and Social Impacts

This paper proposes GLIPv2, an unified framework for VL representation learning that serves both localization tasks and VL understanding tasks. We experimentally verify the effectiveness of the unified model and the novel region-word contrastive learning. Compared to existing methods, GLIPv2 achieves competitive near SoTA performance on various localization and understanding tasks.

# References


[1] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6077–6086. IEEE (2018)

[2] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)

[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: International Conference on Computer Vision (ICCV) (2015)

[4] Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 384–400 (2018)

[5] Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2846–2854 (2016)

[6] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)

[7] Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. Advances in Neural Information Processing Systems **32** (2019)

[8] Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)

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

[10] Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4983 (2019)

[11] Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

[12] Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Universal image-text representation learning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 104–120. Springer (2020)

[13] Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740 (2019)

[14] Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems. pp. 379–387 (2016)

[15] Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., Zhang, L.: Dynamic head: Unifying object detection heads with attentions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7373–7382 (2021)

[16] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[17] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544 (2021)

[18] Gokberk Cinbis, R., Verbeek, J., Schmid, C.: Multi-fold mil training for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2409–2416 (2014)

[19] Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Zero-shot detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)

[20] Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019)

[21] Gupta, T., Kamath, A., Kembhavi, A., Hoiem, D.: Towards general purpose vision systems. arXiv preprint arXiv:2104.00743 (2021)

[22] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)

[23] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)

[24] Hu, R., Singh, A.: Unit: Multimodal multitask learning with a unified transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1439–1449 (2021)

[25] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., 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)

[26] Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems **27** (2014)

[27] Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)

[28] Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github. com/openimages **2**(3), 18 (2017)

[29] 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 (IJCV) **123**(1), 32–73 (2017)

[30] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)

[31] Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)

[32] Li, G., Duan, N., Fang, Y., Jiang, D., Zhou, M.: Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. arXiv preprint arXiv:1908.06066 (2019)

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

[34] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)

[35] Li, L.H., You, H., Wang, Z., Zareian, A., Chang, S.F., Chang, K.W.: Unsupervised vision-and-language pre-training without parallel images and captions. arXiv preprint arXiv:2010.12831 (2020)

[36] Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. arXiv preprint arXiv:2112.03857 (2021)

[37] 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: Proceedings of the European Conference on Computer Vision (ECCV). pp. 121–137. Springer (2020)

[38] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

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

[40] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)

[41] Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 13–23 (2019)

[42] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)

[43] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)

[44] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)

[45] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)

[46] Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020)

[47] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)

[48] Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)

[49] Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)

[50] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

[51] Wang, P., Cai, Z., Yang, H., Swaminathan, G., Vasconcelos, N., Schiele, B., Soatto, S.: Omni-detr: Omni-supervised object detection with transformers. arXiv preprint arXiv:2203.16089 (2022)

[52] Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957 (2020)

[53] Wu, C., Lin, Z., Cohen, S., Bui, T., Maji, S.: Phrasecut: Language-based image segmentation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10216–10225 (2020)

[54] Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J.: Unified contrastive learning in image-text-label space. arXiv preprint arXiv:2204.03610 (2022)

[55] Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Crossing the format boundary of text and boxes: Towards unified vision-language modeling. arXiv preprint arXiv:2111.12085 (2021)

[56] Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)

[57] Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021)

[58] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., 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)

[59] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021)

[60] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified vision-language pre-training for image captioning and VQA. AAAI (2020)

[61] Zhu, P., Wang, H., Saligrama, V.: Zero shot detection. IEEE Transactions on Circuits and Systems for Video Technology **30**(4), 998–1010 (2019)

[62] Zhu, P., Wang, H., Saligrama, V.: Don't even look once: Synthesizing features for zero-shot detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11693–11702 (2020)

## Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]