# The ART of Composition: Attention-Regularized Training for Compositional Visual Grounding

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Vision–Language Models (VLMs) have achieved strong performance on (implicit and explicit) visual grounding and related tasks. However, such abilities are generally tested on simple, single-object phrases. We find that grounding performance degrades for complex, multi-object references. These limitations largely arise from training objectives that leverage image-caption alignment, where direct multi-object references are rare, number of possible such references is theoretically large (exponential in the number of objects) and attribution is difficult. To address this, without requiring any additional annotations, we propose Compositional Attention-Regularized Training (CompART) training, which decomposes captions into object-centric phrases and constructs composite phrases by pairing them with conjunctions. We then introduce a *composition loss* that encourages the attention induced by a composite phrase to equal the sum of the attentions of its constituent phrases, promoting balanced multi-object localization. We evaluate CompART across four VLM architectures, spanning both contrastive-based and generative-based models, on four benchmarks for multi-object grounding and two VQA benchmarks for general visual understanding. CompART consistently achieves improved grounding for both single- and multi-object references across diverse VLM architectures and datasets, and further demonstrates enhanced visual understanding, as evidenced by gains on VQA, despite not being explicitly trained for the task.[1]

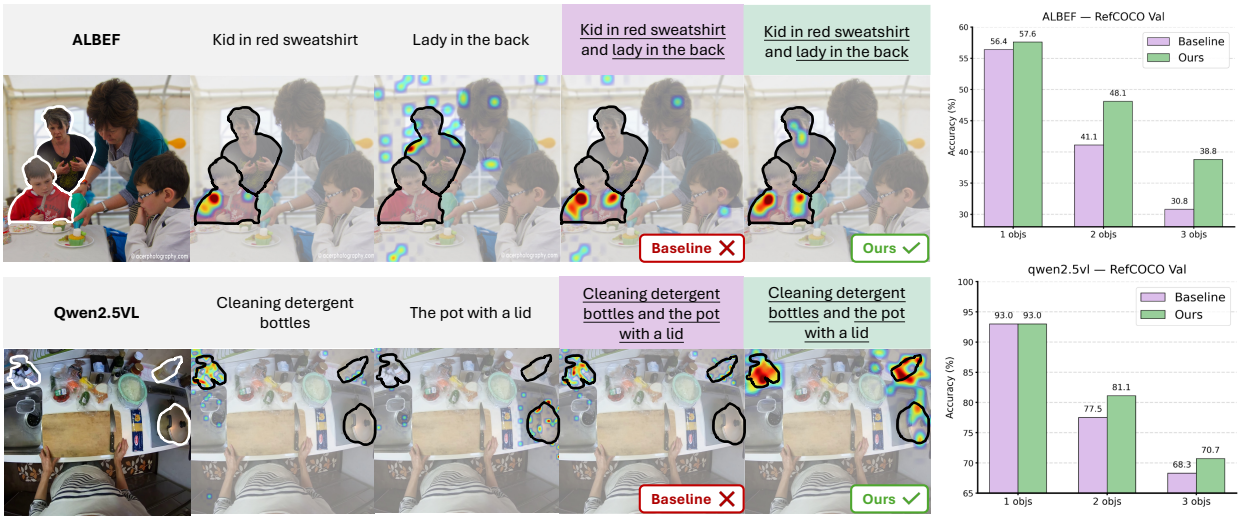


Figure 1: **Motivation.** We present qualitative and quantitative results of ALBEF and Qwen2.5VL on single- and multi-object grounding. While current VLMs perform well on simple single-object grounding tasks, their performance degrades substantially when localizing the same set of objects jointly using composite phrases ( purple ). Proposed CompART significantly improves performance ( green ), without requiring added annotations.

[1] We will release all models, data and code upon acceptance

Vision–Language Models (VLMs) (Singh et al., 2022; Li et al., 2020; 2022b; 2023c; Xu et al., 2023; Alayrac et al., 2022; Kim et al., 2021; Li et al., 2022a; Liu et al., 2023c; Radford et al., 2021; Jia et al., 2021; Zhou et al., 2024; Bai et al., 2025) have demonstrated remarkable success in a wide range of vision tasks including, visual question answering (VQA) (Goyal et al., 2017), as well as both implicit/emergent (Jiayun et al., 2024; He et al., 2024; Lan et al., 2024; Zeng et al., 2024), where the grounding capability emerges from coarse image-caption supervision, and explicit visual grounding (Chen et al., 2023; You et al., 2023; Liu et al., 2024b; Rasheed et al., 2024), where the models are explicitly trained to do grounding. However, current VLMs continue to struggle with grounding complex queries involving multiple objects. As shown in Figure 1, when VLMs are prompted with a multi-object query such as `'kid in red sweatshirt and lady in the back'`, attention often collapses onto a single salient region, `'kid'`. As a result, grounding performance drastically degrades as the number of objects in the query increases (see Figure 1, purple bar). This gap likely stems from the current training paradigm, which relies on unstructured image–caption pairs for vision-language alignment. In such data, direct multi-object references are relatively rare, the number of possible combinations grows exponentially with the number of objects, and attribution across objects is inherently ambiguous. Consequently, this limits VLMs' ability to attend to or localize multi-object phrases which can adversely affect model performance in tasks that require multi-object reasoning, such as grounding (Liu et al., 2023a; 2024b; Hu et al., 2023), instance counting (Li et al., 2023b), and spatial reasoning (Kamath et al., 2023; Chen et al., 2024).

One solution for this limitation would be to collect dedicated data with composite references and corresponding images, however, this would require annotation of a potentially exponential[2] set of phrases and also would rely on VLMs implicitly learning compositional behavior. Instead, we explore whether the compositional structure inherent in captions can provide useful supervisory signals. Prior work in natural language processing has shown that leveraging syntactic and compositional structure improves semantic modeling across tasks such as translation (Andreas et al., 2013; Eriguchi et al., 2017) and sentence classification (Komninos & Manandhar, 2016; Wu et al., 2020), and earlier vision–language systems similarly exploited structured language representations for retrieval (Vendrov et al., 2015) and single object grounding (Xiao et al., 2017). Related work in vision has also explored how structuring interactions between regions can improve the compositionality of visual representations (Hossain et al., 2024; Chen et al., 2019). However, such structured supervision has yet to be integrated into modern large-scale VLM training pipelines. Motivated by this, we propose Compositional Attention Regularization training (CompART), a simple method that incorporates attention regularization, based on compositional text structure, into VLM optimization. CompART improves model performance on multi-object grounding (Figure 1, green bar), instance counting and spatial relation understanding, without requiring added dense annotations.

Specifically, we first decompose each image caption into *object-centric phrase-chunks*, each corresponding to a distinct visual entity. We then construct *composite phrases* by pairing two such phrase chunks using a simple conjunction ("and"), which is applicable regardless of the specific relationship between the objects. Motivated by the observation that VLMs can reliably localize objects when queried individually, but often exhibit attention collapse to a single object region under multi-object queries, we introduce *composition loss* that exploits the emergent localization behavior of VLMs on single-object queries (Jiayun et al., 2024; He et al., 2024; Kang et al., 2025). Concretely, our *composition loss* encourages the attention map induced by a composite phrase to approximate the sum of the attention maps of its constituent phrases, thereby explicitly regularizing attention. As a result, *composition loss* mitigates single-object dominance and promotes simultaneous multi-object grounding without added dense annotation. Although our *composition loss* is applied only to object pairs during training due to computational constraints, the resulting model generalizes to grounding scenarios involving more than two-objects (see Figure 1) and further demonstrates improvements across diverse VQA tasks.

We evaluate CompART across four representative VLM architectures, spanning both contrastive models (ALBEF (Li et al., 2021) and BLIP (Li et al., 2022b)) and generative models (TinyLLaVA (Zhou et al., 2024) and Qwen2.5VL (Bai et al., 2025)). We include contrastive VLMs alongside more recent generative models because, despite having significantly fewer parameters, they often exhibit comparable grounding performance (Kang et al., 2025; Bhatia et al., 2024; Chen et al., 2023), making them particularly attractive in

[2]Exponential in the number of entities involved.

low-compute settings. We assess the impact of CompART on six benchmarks, comprising four grounding benchmarks and two VQA benchmarks. Across all settings, CompART consistently improves over the corresponding base models. For example, on average, CompART improves two-object grounding performance of TinyLLaVA-0.5B by 4.9% points and Qwen2.5VL by 2.6% points across all four grounding benchmarks. Importantly, our method introduces no architectural modifications to the base models. Instead, we propose a compositional attention regularization objective that is architecture-agnostic and integrates seamlessly into standard training pipelines, requiring no additional dense or bounding-box annotations. Despite its simplicity, CompART enhances multi-object grounding and improves broader visual understanding, as reflected by consistent gains across benchmarks.

**Contributions.** In summary, we make the following contributions:

**(1)** identify a consistent performance gap when transitioning from single- to multi-object grounding in both contrastive and generative VLMs;

**(2)** introduce Compositional Attention Regularization Training (CompART), a simple dense-annotation-free solution for improving multi-object grounding;

**(3)** demonstrate that proposed framework leads to consistent improvements across diverse VLMs on multi-object grounding and VQA benchmarks.

## 1 Related Work

**Vision–Language Models.** Vision–language pretraining underpins modern multimodal learning and achieves strong performance across diverse tasks. Existing VLMs broadly fall into two categories: (1) *contrastive-based models*, which learn aligned image–text representations via large-scale contrastive objectives, including dual-encoder approaches such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), as well as fusion-based variants such as ALBEF (Li et al., 2021) and BLIP (Li et al., 2022b) that enable explicit cross-modal interaction; and (2) *generative-based* VLMs that integrate visual encoders with large language models (e.g., LLaVA (Liu et al., 2023c), Phi3.5V (Abdin et al., 2024), DeepseekVL (Lu et al., 2024), Molmo (Deitke et al., 2025), InternVL (Zhu et al., 2025), TinyLLaVA (Zhou et al., 2024), Qwen2.5-VL (Bai et al., 2025)), enabling open-ended multimodal reasoning and generation.

**Visual Grounding.** Visual grounding methods can broadly be categorized into implicit and explicit grounding approaches. Implicit grounding (He et al., 2024; Jin et al., 2023; Zeng et al., 2024; Hossain et al., 2026) refers to methods where phrase–region correspondences emerge from a model's internal representations, such as cross-modal attention maps, without directly predicting spatial outputs. In contrast, explicit grounding (Rohrbach et al., 2016; Yu et al., 2018; Lu et al., 2019; Lai et al., 2024; Rasheed et al., 2024) approaches train models to localize objects by predicting spatial outputs such as bounding boxes, coordinates, or segmentation masks directly.

Several implicit grounding approaches are trained in a weakly supervised manner using only paired image–text data without relying on region proposals or detector supervision. These methods encourage phrase–region alignment through training strategies applied to vision–language models, such as progressive learning (Yang et al., 2024b), consistency across paraphrases (Lee et al., 2023; He et al., 2024), and object disentanglement constraints (Zeng et al., 2024). Another line of work exploits pseudo ground-truth signals such as region proposals generated by object detectors, where textual phrases are aligned with candidate regions extracted from the image (Rohrbach et al., 2016; Yu et al., 2018; Lu et al., 2019; Chou et al., 2022; Luo et al., 2024; Jin et al., 2023). These methods learn phrase–region correspondences by matching referring expressions to candidate region proposals. Further improvements have been explored through strategies such as knowledge distillation from teacher models (Wang et al., 2021), prompt-based optimization (Luo et al., 2024), and curriculum learning (Dai & Yang, 2024). While effective, these approaches primarily enhance contrastive-based models for single-object grounding and often rely on task-specific architectural modifications.

More recent work, such as GroudingDINO (Liu et al., 2024b), GRES (Liu et al., 2023a) and Beyond-one-to-one(Hu et al., 2023), leverages dense annotations to extend contrastive models to more challenging

multi-object grounding scenarios. ReMeREC (Hu et al., 2025) further introduces relation-aware multi-entity referring comprehension to explicitly model inter-object relationships. In parallel, several grounding-oriented multimodal large language models extend instruction-tuned LLMs with explicit spatial grounding capabilities, enabling direct coordinate prediction or mask generation via segmentation decoders (Lai et al., 2024; Rasheed et al., 2024; Zhang et al., 2024a;b; Chen et al., 2023; You et al., 2023; Yin et al., 2025; Jiang et al., 2025). However, these methods largely depend on dense annotations to support multi-object grounding.

Different from the above approaches, CompART leverages caption compositionality and attention regularization to improve multi-object grounding in both contrastive-based and generative-based VLMs. Our method is simple, architecture-agnostic, and does not require dense annotations.

**The Role of Grounding in Visual Understanding.** Accurate visual grounding has been shown to improve broader visual understanding tasks such as VQA (Reich & Schultz, 2024). Several studies demonstrate that encouraging models to attend to question-relevant image regions leads to improved reasoning accuracy and robustness (Selvaraju et al., 2019; Wu & Mooney, 2019; Le et al., 2023). Other works explicitly learn grounded representations or analyze the role of grounding in VQA, showing that aligning language queries with relevant visual regions improves interpretability and model performance (Urooj et al., 2021; Reich & Schultz, 2024).

## 2 Compositional Attention Regularization Training (CompART)

In Figure 2, we present a high-level overview of CompART. We first extract *object-centric phrase-chunks* from the original image captions and form *composite phrases* by pairing them (Sec. 2.1). We then compute localization maps for both the individual phrases and their composite counterparts (Sec. 2.2), which are used to derive the *composite loss* for attention-regularized training (Sec. 2.3). The inference pipeline is discussed in Sec. 2.5. We apply CompART to both contrastive-based (*i.e.*, ALBEF (Li et al., 2021), BLIP (Li et al., 2022b)) and generative-based models (*i.e.*, TinyLLaVA (Zhou et al., 2024), Qwen2.5VL (Bai et al., 2025)). In addition to TinyLLaVA and Qwen2.5VL, which inherently compute text-to-visual attention by virtue of their decoder-only LLM architecture, we also consider contrastive-based models such as ALBEF and BLIP, as they provide explicit text-conditioned visual attention that enables phrase-level localization. However, due to their architectural differences, we adopt slightly different strategies as described below.

### 2.1 Phrase Chunking

**Contrastive-based Models.** Given an image $v$ and its corresponding caption $c$, we employ an LLM to extract object-centric phrase chunks $c_i$ and $c_j$, where $i$ and $j$ refer to the $i$-th and $j$-th phrases in $c$. We then construct a composite phrase chunk $c_{ij}$ by combining $c_i$ and $c_j$. Subsequently, the phrase chunks are paired with $v$, which serves as the model input. See examples in Figure 2(a).

**Generative-based Models.** Similar to the aforementioned procedure, for an image $v$ and its corresponding caption $c$, we extract object-centric phrase chunks $c_i$ and $c_j$. Since generative-based VLMs produce outputs autoregressively conditioned on the input image and query, we construct a query $q_i$ by replacing $c_i$ in $c$ with a special token `[blank]` and appending the instruction "Please fill in the `[blank]`." The removed phrase $c_i$ serves as the ground-truth target. Similarly, we obtain $q_j$ by replacing $c_j$ in $c$ with `[blank]`. To form the query $q_{ij}$ corresponding to the composite phrase $c_{ij}$, we replace both $c_i$ and $c_j$ in $c$ with `[blank]` and use $c_{ij}$ as the target output. See examples in Figure 2(b).

### 2.2 Phrase Localization

**Contrastive-based Models:** For a given image–phrase pair $\{v, c_i\}$, we obtain the text-to-image cross-attention map $\mathbf{A}_i \in \mathbb{R}^{K \times P}$, where $P$ denotes the total number of visual patches and $K$ denotes the number of tokens corresponding to $c_i$. However, $\mathbf{A}_i$ is often noisy and leads to over-segmentation (Jiayun et al., 2024). To address that, following Jiayun et al. (2024) and He et al. (2024), we adopt a GradCAM-based refinement strategy (Selvaraju et al., 2017) which suppresses spurious activations and emphasizes regions most relevant

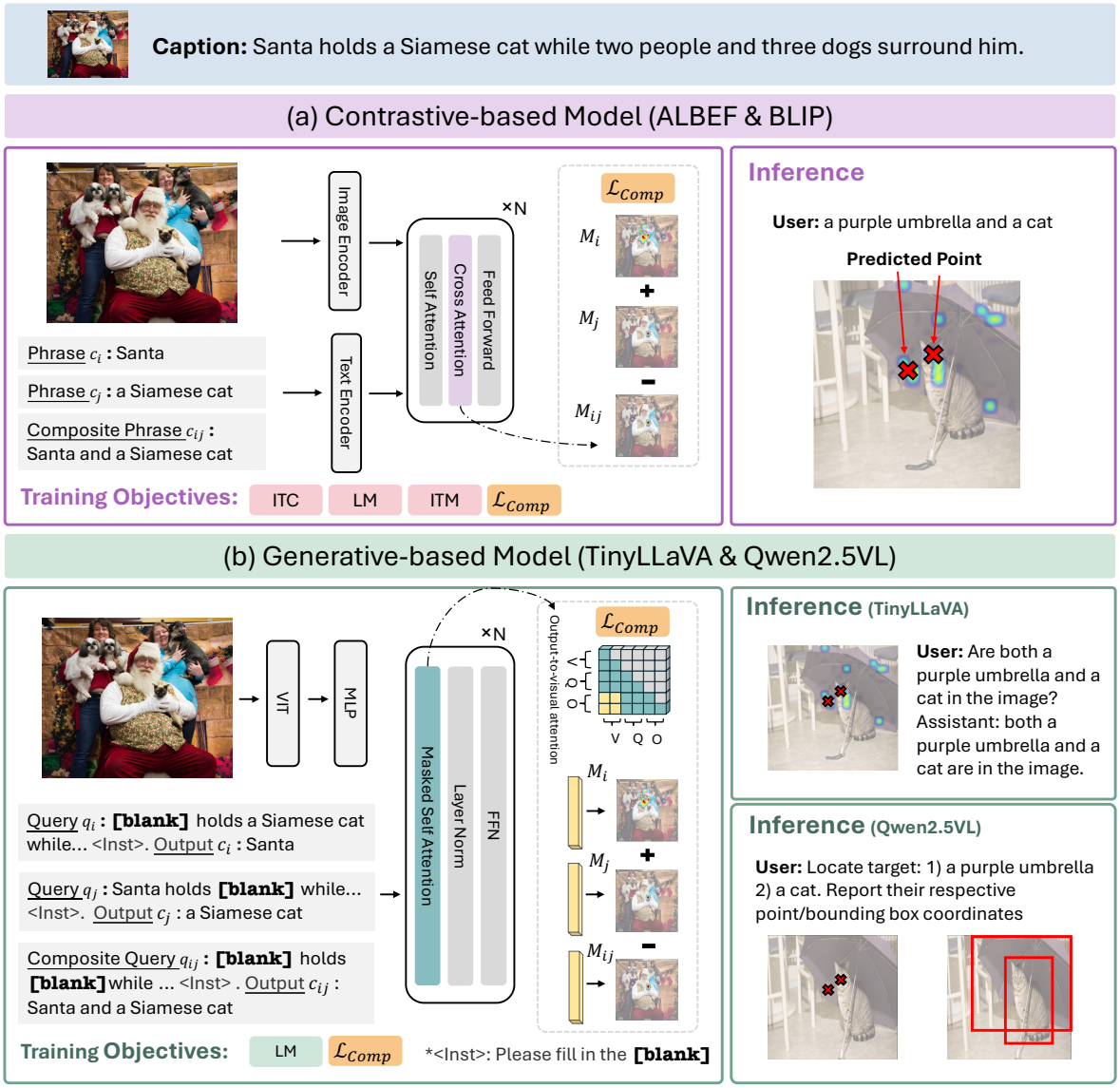


Figure 2: **Overview of CompART.** During training, we first extract *object-centric phrase-chunks* from the image caption and form *composite phrases* by pairing them. Then, we extract localization maps for the phrases. Since contrastive and generative VLMs are trained differently, the composite queries are formulated slightly differently for each VLM. Our proposed *composition loss* ($\mathcal{L}_{\text{Comp}}$) encourages the localization map of a composite phrase to correspond to the union of the object regions associated with the phrase chunks. During inference, for models without explicit grounding capability (ALBEF, BLIP, and TinyLLaVA), we use the peak of the attention maps corresponding to the query objects as the predicted points. Since Qwen2.5VL supports explicit grounding capabilities, we directly obtain output points or bounding box coordinates.

to the phrase. Specifically, the refined localization map is obtained as the element-wise product between $\mathbf{A}_i$ and the gradient of the image–text matching (ITM) loss $\mathcal{L}_{\text{ITM},i}$ with respect to $\mathbf{A}_i$[3]. Subsequently, we obtain the phrase-specific localization map $\mathbf{M}_i \in \mathbb{R}^P$ by averaging over $K$ tokens as:

$$\mathbf{M}_i = \frac{1}{K} \sum_{k=1}^{K} \texttt{ReLU}\left( \mathbf{A}_i^{(k)} \cdot \frac{\partial \mathcal{L}_{\text{ITM},i}}{\partial \mathbf{A}_i^{(k)}} \right). \tag{1}$$

[3]While ALBEF and BLIP are also trained with image-text-contrastive (ITC) and language-modelling (LM) objectives, we compute gradients specifically with respect to ITM, as it directly enforces matching between the image and phrase *within* a sample. In contrast, ITC is a contrastive objective applied *across* samples, and LM operates purely on language.

**Generative-based Models:** For a given image–phrase pair $\{v, c_i\}$, we construct the model input as $\{v, q_i\}$ and set the target output to $c_i$, as defined in Sec. 2.1. For phrase localization, we extract the output-to-visual ($c_i$ to $v$) attention as $\mathbf{A}_i \in R^{K \times P}$, where $K$ represents the number of tokens in $c_i$ and $P$ represent the total number of patches in $v$. We define our phrase-specific localization map $\mathbf{M}_i \in \mathbb{R}^P$ as:

$$\mathbf{M}_i = \frac{1}{K} \sum_{k=1}^{K} \left( \mathbf{A}_i^{(k)} \right). \tag{2}$$

### 2.3 Composition Loss

We next introduce *composition loss* ($\mathcal{L}_{\text{Comp}}$) for attention-regularized training to improve visual grounding. $\mathcal{L}_{\text{Comp}}$ encourages the model to jointly attend to multiple object regions associated with each individual phrase chunks by minimizing the Euclidean distance between the composite phrase attention map ($\mathbf{M}_{ij}$) and the sum of the individual phrase attention maps ($\mathbf{M}_i, \mathbf{M}_j$), defined as:

$$\mathcal{L}_{\text{Comp}} = \|\mathbf{M}_i + \mathbf{M}_j - \mathbf{M}_{ij}\|_2^2 . \tag{3}$$

$\mathcal{L}_{\text{Comp}}$ directly addresses the lack of multi-object grounding by explicitly regularizing consistency between the attention maps of individual phrase chunks and their composite form.

### 2.4 Final Objective

Finally, we train the models using a weighted sum of their respective base training objectives ($\mathcal{L}_{\text{base}}$) and our *composition loss* ($\mathcal{L}_{\text{Comp}}$). ALBEF and BLIP are trained with their base objectives, such as ITC, ITM, and LM losses, whereas TinyLLaVA and Qwen2.5VL only use LM loss. The final training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda \times \mathcal{L}_{\text{Comp}}, \tag{4}$$

where $\lambda$ is a loss coefficient to scale the $\mathcal{L}_{\text{Comp}}$ to be the same range as $\mathcal{L}_{\text{base}}$. The combination of base and *composition loss* encourages the models to retain their general vision-language capabilities while improving visual grounding.

### 2.5 Inference

The goal of CompART is to enhance the overall visual grounding capabilities of VLMs, rather than explicitly optimizing them for specific VQA tasks. Nevertheless, stronger visual grounding is expected to translate into improved performance on general VQA benchmarks. In the following, we describe our inference strategies for both visual grounding and VQA tasks.

**Visual Grounding:** Ability and protocol for obtaining visual grounding differs among VLM families. For example, Qwen2.5-VL (Bai et al., 2025), which is instruction-tuned for coordinate prediction, has an *explicit* ability to localize objects, and one can directly instruct the model to output object points or bounding boxes coordinates. ALBEF (Li et al., 2021), BLIP (Li et al., 2022b) and TinyLLaVA (Zhou et al., 2024), are not equipped with such abilities and only posses *implicit* grounding typically exhibited by their attention maps (He et al., 2024; Jiayun et al., 2024; Lan et al., 2024; Zeng et al., 2024). Prior works either (i) post process such maps to produce segmentations (Jiayun et al., 2024; Hossain et al., 2026; Cha et al., 2023) or (ii) evaluate such models using a pointing metric (He et al., 2024; Yang et al., 2024b; Liu et al., 2023b; Shaharabany et al., 2022; Shaharabany & Wolf, 2023; Lin et al., 2024), which measures ability of the model to produce a coordinate (typically peak of the attention map) within the referred object's bounding box. However, the first approach is sensitive to the choice of post-processing methods (Krähenbühl & Koltun, 2011; Araslanov & Roth, 2020), potentially obfuscating grounding abilities. We adopt the second mechanism for such models and derive localization map from the text-to-image attention maps following prior work (He et al., 2024; Kang et al., 2025) and as described in Sec. 2.2 – treating the peak of the map as the prediction of the object location.

**VQA:** We follow the default inference setup of the base models for VQA tasks.

# 3 Experiments & Results

## 3.1 Training Details

We use ALBEF, BLIP, TinyLLaVA-0.5B, and Qwen2.5VL-3B as base VLMs to implement CompART. Our training data is sourced from two popular image-captioning datasets: Pixmo Caption (Deitke et al., 2025) and MSCOCO (Lin et al., 2014). Specifically, we obtain 8.5K random image–caption pairs from Pixmo Caption, followed by using Qwen3-30B (Yang et al., 2025) to prepare the object-centric phrase chunks. We construct an average of around 10 composite phrases per image, resulting in a total of 90K training samples. In the case of MSCOCO, we directly use object-centric phrase chunks from He et al. (2024), which contains 123K images and an average of 11 composite phrases per image, resulting in a total of 1.4M training samples. Both Pixmo and MSCOCO are used to train ALBEF, BLIP, and TinyLLaVA; however, due to computational constraints, we only use Pixmo to train Qwen2.5VL.

We calculate object localization maps from the grounding layers of the VLMs. Specifically, for ALBEF and BLIP, we follow He et al. (2024) and Jiayun et al. (2024) and extract text-to-image attention maps from the $3^{\text{rd}}$ and $9^{\text{th}}$ cross-attention layers, respectively. For TinyLLaVA-0.5B and Qwen2.5VL-3B, we extract output-to-visual attention maps from the $23^{\text{rd}}$ layer and the $21^{\text{st}}$ layers of the VLMs. These layers are chosen empirically, as they assign the highest attention to visual tokens. All experiments are conducted on 8 NVIDIA L40 GPUs (48GB each). Full training details and hyperparameters are provided in the supplementary.

## 3.2 Evaluation Setup

**Visual Grounding.** We evaluate CompART on four visual grounding benchmarks: Flickr30K (Plummer et al., 2015), along with all validation and test splits of RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016). We assess performance across single- to five-object grounding settings. For single-object grounding, we use the full evaluation splits. For multi-object grounding, we consider images that contain ground-truth annotations for multiple objects and evaluate all possible object combinations. Specifically, for a $k-$object grounding setting, we use images that contain at least $k$ annotated objects and evaluate all possible $k-$object combinations within each image. Detailed evaluation set statistics are provided in the supplementary material; data splits will be released upon publication.

It should be noted that for multi-object grounding, predicting a single point is insufficient: a prediction is considered correct if it falls within any one of the target objects, which does not reflect whether the model attends to all referred objects. Similarly, when predicting a single bounding box and evaluating it using Acc@0.5, the metric does not capture the extent to which the model simultaneously localizes multiple objects. Therefore, multi-object grounding evaluation requires predicting distinct points or bounding boxes for each referred object and evaluating each prediction against its corresponding ground-truth annotation. Accordingly, we evaluate the localization of each referred object in such multi-object grounding setting.

We assess grounding performance in two setups: point estimation and bounding box-based estimation. As ALBEF, BLIP, and TinyLLaVA are not instruction-tuned to directly generate bounding boxes, we evaluate them using point estimation only, whereas Qwen2.5VL is evaluated using both point- and bounding-box estimation. For point estimation, we adopt Pointing Game Accuracy (Zhang et al., 2016), which measures the percentage of predicted points that fall within the ground-truth bounding box. For bounding-box estimation, we report Acc@0.5.

For contrastive VLMs such as ALBEF and BLIP, we directly use the object query phrases as input. However, for the generative model TinyLLaVA, we construct input-output sequences for each query phrase as: `"Are there {query phrase} in the image? There are {query phrase} in the image."` to obtain output-to-visual localization map. In all cases, localization maps for all queried objects are obtained in single forward pass. Object-specific point predictions are then computed by taking the peak attention from the localization map corresponding to each respective object phrase as detailed in Sec. 2.2. For Qwen2.5VL, we instruct the model to directly output either object point coordinates or bounding-box coordinates for all objects in the given query.

Table 1: **Single- and two-object grounding results** based on Point Game Accuracy and Bounding-Box Acc@0.5. The numbers in **bold** indicate the best results in each group.

| | Average | | Flickr30K | | RefCOCO | | RefCOCOg | | RefCOCO+ | |
|---|---|---|---|---|---|---|---|---|---|---|
| **No. of objects** | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| TAS (Lin et al., 2024) | - | - | 77.9 | - | - | - | - | - | - | - |
| G++ (Shaharabany & Wolf, 2023) | - | - | 78.1 | - | - | - | - | - | - | - |
| WWbL (Shaharabany et al., 2022) | - | - | - | - | 31.1 | - | 31.8 | - | 34.6 | - |
| TRIS (Liu et al., 2023b) | - | - | - | - | 51.9 | - | 52.6 | - | 40.9 | - |
| PCNet (Yang et al., 2024b) | - | - | - | - | 60.6 | - | 57.7 | - | 56.5 | - |
| *The above numbers are for point acc. and are for reference only; direct comparison may not be appropriate.* | | | | | | | | | | |
| Training data: MSCOCO | | | | | | | | | | |
| ALBEF (Li et al., 2021) | 66.6 | 55.4 | 79.4 | 78.8 | 56.9 | 41.5 | **68.1** | 52.7 | 62.0 | 48.8 |
| + SelfEQ (He et al., 2024) | 68.0 | 58.1 | 84.1 | 83.8 | 57.6 | 42.9 | 67.6 | 54.8 | 62.5 | 51.1 |
| + CompART (Ours) | **68.1**$_{\uparrow 1.5}$ | **61.7**$_{\uparrow 6.3}$ | **85.5** | **86.7** | **58.2** | **49.1** | 66.1 | **58.0** | **62.7** | **53.2** |
| BLIP (Li et al., 2022b) | 66.8 | 62.5 | 81.3 | 84.9 | 58.1 | 51.6 | 65.9 | 59.6 | 61.8 | 53.7 |
| + CompART (Ours) | **68.8**$_{\uparrow 2.1}$ | **63.7**$_{\uparrow 1.3}$ | **84.7** | **86.8** | **59.8** | **52.2** | **67.5** | **61.1** | **63.4** | **54.8** |
| TinyLLaVA-0.5B (Zhou et al., 2024) | 19.8 | 24.4 | 33.4 | 42.8 | 12.7 | 17.0 | 19.5 | 20.3 | 13.6 | 17.6 |
| + CompART (Ours) | **28.8**$_{\uparrow 9.0}$ | **27.1**$_{\uparrow 2.7}$ | **44.3** | **44.2** | **21.3** | **20.1** | **26.1** | **22.8** | **23.4** | **21.2** |
| Training data: Pixmo | | | | | | | | | | |
| ALBEF (Li et al., 2021) | 67.1 | 55.9 | 80.3 | 79.3 | 57.3 | 42.0 | **68.4** | 53.1 | 62.5 | 49.1 |
| + CompART (Ours) | 67.1 | **57.0**$_{\uparrow 1.2}$ | **81.4** | **80.8** | **57.7** | **42.4** | 66.6 | **55.0** | **62.7** | **49.9** |
| BLIP (Li et al., 2022b) | 65.4 | 61.6 | 81.1 | **86.3** | 56.2 | 50.1 | 65.0 | 57.9 | 59.2 | 52.1 |
| + CompART (Ours) | **66.8**$_{\uparrow 1.5}$ | **62.5**$_{\uparrow 0.8}$ | **81.8** | 84.3 | **58.4** | **51.7** | **65.4** | **59.7** | **61.7** | **54.1** |
| TinyLLaVA-0.5B (Zhou et al., 2024) | 20.8 | 25.1 | 34.7 | 43.1 | 13.5 | 17.7 | 20.7 | 21.2 | 14.6 | 18.4 |
| + CompART (Ours) | **28.4**$_{\uparrow 7.5}$ | **30.0**$_{\uparrow 4.9}$ | **45.3** | **48.3** | **19.3** | **21.7** | **28.0** | **26.5** | **20.9** | **23.6** |
| Qwen2.5VL-3B (Bai et al., 2025) | 89.3 | 72.3 | **88.6** | 82.9 | **92.9** | 77.1 | 90.2 | 59.9 | **85.5** | 69.2 |
| + CompART (Ours) | 89.2$_{\downarrow 0.1}$ | **74.9**$_{\uparrow 2.6}$ | 88.5 | **83.6** | 92.8 | **80.6** | 90.2 | **64.4** | 85.3 | **71.1** |
| Qwen2.5VL-3B$^{\text{BBox}}$ (Bai et al., 2025) | 64.5 | 58.7 | 73.4 | 73.8 | 87.4 | 76.9 | 83.3 | 74.6 | **78.4** | 68.4 |
| + CompART (Ours) | 64.5 | **59.1**$_{\uparrow 0.4}$ | **73.5** | **74.2** | 87.4 | **77.5** | 83.3 | **75.2** | 78.2 | **68.7** |
| *These numbers are for bbox acc@0.5 and are for reference only; direct comparison may not be appropriate.* | | | | | | | | | | |
| WeakMCN (Cheng et al., 2025) | - | - | - | - | 67.2 | - | - | - | 44.4 | - |
| ReCLIP (Subramanian et al., 2022) | - | - | - | - | 46.3 | - | 59.2 | - | 47.7 | - |
| DeepSeek-VL-1.3B (Kang et al., 2025) | - | - | - | - | 73.9 | - | 67.3 | - | 61.9 | - |
| DeepSeek-VL-7B (Kang et al., 2025) | - | - | - | - | 84.5 | - | 82.0 | - | 78.4 | - |
| Shikra-7B (Chen et al., 2023) | - | - | - | - | 85.9 | - | 82.3 | - | 80.4 | - |
| Ferret-7B (You et al., 2023) | - | - | - | - | 87.1 | - | 84.4 | - | 80.4 | - |

**VQA.** Since generative-based VLMs are also instruction-tuned for VQA tasks, we evaluate their performance on two popular general VQA benchmarks SeedBench (Li et al., 2023b) and SeedBench2 (Li et al., 2023a) using their default evaluation protocols.

### 3.3 Results

**Single- and Two-Object Grounding.** In Table 1, we present the results of single- and two-object grounding performance based on point game accuracy. To strictly evaluate the impact of our proposed CompART in improving the object grounding performance, we train the base VLMs using their original training objectives and compare them against models trained with CompART.

As shown in Table 1, CompART consistently improves grounding performance across all four benchmarks and model families. For example, ALBEF improves single-object and two-object grounding performance by 1.5% and 6.7% points, repectively, when trained with MSCOCO. Additionally, in single-object setting on TinyLLaVA-0.5B, CompART improves average performance by 9.0% and 7.5% points when trained with MSCOCO and PixMo, respectively. In the two-object setting, the corresponding improvements are 2.7% and 4.9% points. Since Qwen2.5VL-3B already achieves close to 90% accuracy in the single-object setting,

Table 2: **Stress test of CompART** in three-, four-, and five-object grounding setups based on Point Game Accuracy and Bounding-Box Acc@0.5. The numbers in **bold** indicate the best results.

| | **Flickr30K** | | | **RefCOCO** | | | **RefCOCOg** | | | **RefCOCO+** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No. of objects** | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 |
| TinyLLaVA-0.5B | 41.3 | 39.1 | 36.5 | 13.7 | 11.5 | 10.1 | 19.7 | 18.0 | 16.9 | 15.0 | 12.4 | 11.7 |
| +CompART | **45.3** | **41.9** | **37.9** | **16.1** | **13.1** | **11.6** | **22.2** | **19.1** | **17.5** | **18.4** | **14.8** | **13.5** |
| Δ | +4.0 | +2.8 | +1.4 | +2.3 | +1.6 | +1.5 | +2.5 | +1.1 | +0.6 | +3.4 | +2.4 | +1.7 |
| Qwen2.5VL-3B | 72.5 | 63.9 | 63.4 | 67.1 | 64.0 | 59.8 | 58.3 | 64.2 | 61.7 | 57.9 | 52.1 | 43.1 |
| | **77.5** | **72.4** | **68.9** | **69.5** | **64.4** | **60.0** | **60.2** | **65.2** | **62.3** | **59.9** | **52.5** | **43.3** |
| +CompART | +5.0 | +8.5 | +5.5 | +2.4 | +0.4 | +0.3 | +1.9 | +1.0 | +0.6 | +2.0 | +0.5 | +0.2 |
| Qwen2.5VL-3B$^{\text{BBox}}$ | 68.0 | 61.8 | 56.3 | 66.2 | 57.0 | 51.3 | 65.6 | 55.8 | 49.0 | 56.5 | 44.9 | 37.1 |
| +CompART | **68.3** | **62.3** | **57.0** | **67.4** | **58.6** | **53.8** | **66.4** | **57.7** | **50.6** | **57.0** | 44.9 | **37.5** |
| Δ | +0.3 | +0.5 | +0.7 | +1.1 | +1.6 | +2.4 | +0.9 | +1.9 | +1.6 | +0.6 | 0.0 | +0.4 |

leaving limited room for improvement. In two-object setting, CompART improves grounding performance of Qwen2.5VL-3B by an average of 2.6% points. As Qwen2.5VL-3B can directly produce object bounding boxes, we also evaluate their object grounding performance based on bounding box estimation. As shown in Table 1, despite not explicitly regularizing object boundaries, CompART consistently improves two-object bounding-box grounding performance across all four benchmarks.

Although CompART is designed to improve multi-object grounding, we observe significant improvements in single-object grounding as well. For example, in the case of TinyLLaVA, we notice a gain of up to 9%. We believe these gains mainly stem from two factors. First, for generative models such as TinyLLaVA, our "fill-in-the-[blank]" LM objective uses object-centric phrase chunks as targets, providing more direct object-level supervision than standard caption-level LM training. Second, the MSE-based composition loss suppresses attention noise. If single-object attention maps contain spurious activations, these activations are unlikely to be learned in multi-object maps due to the constraints imposed by $\mathcal{L}_{\text{base}}$ during training, and are therefore penalized. Thus, $\mathcal{L}_{\text{Comp}}$ encourages cleaner single-object attention maps and further improves single-object grounding.

Additionally, we note that most existing grounding methods do not address multi-object grounding. Therefore, for reference, we report their single-object grounding results using both Pointing Game Accuracy and bounding-box Acc@0.5. We also include a recent method, SelfEQ (He et al., 2024), that targets implicit grounding in contrastive VLMs through a similar lightweight training as ours. For a fair comparison, we recompute its performance across all benchmarks for both single- and two-object grounding. As shown in Table 1, our Qwen2.5VL-3B trained with CompART achieves comparable or superior performance, even when compared to larger VLMs such as DeepSeek-VL-7B, Shikra-7B, and Ferret-7B.

**Generalization beyond Two-Objects.** While CompART is trained to regularize attention over two objects, we stress-test its capabilities in grounding upto five-objects. The results presented in Table 2 show that CompART effectively generalizes well in all three-, four-, and five-object grounding settings across all four benchmarks. Specifically, for TinyLLaVA-0.5B, CompART yields large gains on Flickr30K, with 4.0%, 2.8%, and 1.4% points for three-, four-, and five-object grounding, respectively. For Qwen2.5VL-3B point estimation, we also observe significant gains on Flickr30K, with 5.0%, 8.5%, and 5.5% points for three-, four-, and five-object grounding, respectively. In the case of Qwen2.5VL-3B bounding box-based estimation, we observe increasing gains as the number of objects increases. For example, improvements of 1.1%, 1.6%, and 2.4% points are observed for three-, four-, and five-object grounding on RefCOCO.

**General VQA.** To evaluate the impact of improved multi-object grounding on VQA performance, we evaluate Qwen2.5VL-3B on SeedBench (Li et al., 2023b) and SeedBench2 (Li et al., 2023a). These benchmarks cover a diverse set of vision tasks, including single- and multi-object understanding, spatial relation, and chart understanding, among others. As shown in Tables 3, CompART improves overall performance by 2.2%

Table 3: **VQA Result on SeedBench and SeedBench2.** We report accuracy, and the numbers in **bold** indicate the best results.

| | SeedBench | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Avg.** | Ident. | Attr. | Loc. | Count | Spatial | Interact | Visual | Text | Scene |
| Qwen2.5VL-3B | 62.8 | 64.9 | 69.5 | 52.3 | 67.4 | 44.1 | **59.8** | 62.8 | 78.3 | 66.2 |
| +CompART | **65.0**$_{+2.2}$ | **66.8** | **74.6** | **56.9** | **68.7** | **47.2** | 58.8 | **64.7** | 78.3 | **69.2** |
| | **SeedBench2** | | | | | | | | | |
| | **Avg.** | Celeb. | Land. | Chart. | RefExp. | Sci. | Emotion | Math. | | |
| Qwen2.5VL-3B | 58.5 | 74.2 | 75.2 | 81.8 | 43.7 | 74.4 | 13.0 | 47.0 | | |
| +CompART | **60.5**$_{\uparrow 2.0}$ | **75.8** | **76.8** | **82.4** | **46.2** | 74.4 | **18.8** | **49.2** | | |

Table 4: **Ablation on Loss.** We present grounding performance for one- to three-object settings using Pointing Game Accuracy. † refers to the VLM variant trained with original image captions instead of phrase chunks. Pixmo is used for training. The numbers in **bold** indicate the best results in each group.

| | **RefCOCO** | | | **RefCOCOg** | | | **RefCOCO+** | | |
|---|---|---|---|---|---|---|---|---|---|
| **No. of objects** | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| TinyLLaVA-0.5B | 12.7 | 17.0 | 13.3 | 19.5 | 20.3 | 18.9 | 13.6 | 17.6 | 14.1 |
| + $\mathcal{L}_{\text{base}}^{\dagger}$ | 13.5 | 17.7 | 13.7 | 20.7 | 21.2 | 19.7 | 14.6 | 18.4 | 15.0 |
| + $\mathcal{L}_{\text{base}}$ | 13.7 | 16.9 | 13.7 | 20.4 | 22.4 | 21.0 | 15.3 | 18.6 | 15.3 |
| + $\mathcal{L}_{\text{base}}$ + $\mathcal{L}_{\text{Comp}}$ | **19.3** | **21.7** | **16.1** | **28.0** | **26.5** | **22.2** | **20.9** | **23.6** | **21.7** |

and 2.0% on SeedBench and SeedBench2, respectively. Notably, CompART yields large gains on single-object tasks, improving performance by 5.1% and 4.6% points on instance attribute and instance location tasks. We also observe substantial improvements of 3.1% and 3.0% points on spatial relation and scene understanding tasks. Finally, on SeedBench2, CompART improves performance by 2.5% points on both expression and emotion recognition tasks, and by 2.2% points on math-related VQA tasks.

### 3.4 Ablation Study & Analysis

**Ablation on Loss**: We conduct a thorough ablation exploring the impact of individual components of CompART. The results presented in Table 4 show that simply training the base VLM with the image-caption pairs ($\mathcal{L}_{\text{base}}{}^{\dagger}$) or image-phrase chunks ($\mathcal{L}_{\text{base}}$) does not lead to any noticeable improvements over the base model. However, training with the $\mathcal{L}_{\text{Comp}}$ loss significantly improves grounding performance, with average gains of 6.3%, 4.6%, and 3.3% in the one-, two-, and three-object settings, respectively.

**Layer Section.** To compute object-specific phrase-localization maps, we use the output-to-visual attention from the VLM layers exhibiting the highest visual attention. To estimate, we randomly sample 500 images from MSCOCO and compute the output-to-visual attention across all the layers for both TinyLLaVA-0.5B

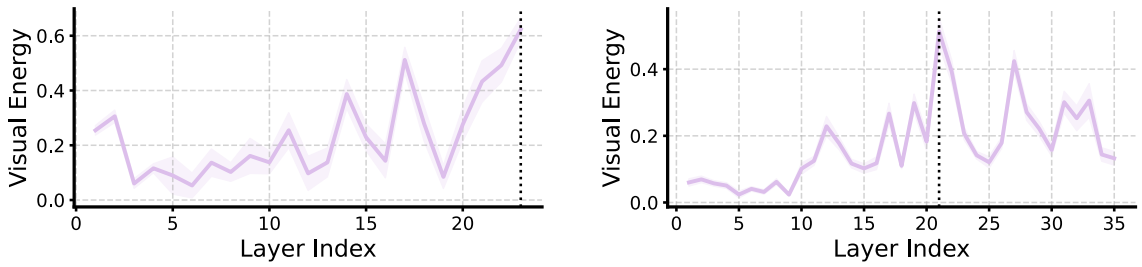


Figure 3: We present the total output-to-visual attention across all layers for TinyLLaVA-0.5B (left) and Qwen2.5VL-3B (right). The layers exhibiting the highest visual attention are indicated by vertical dashed lines; i.e., the 23$^{\text{rd}}$ layer in TinyLLaVA-0.5B and the 21$^{\text{st}}$ layer in Qwen2.5VL-3B.

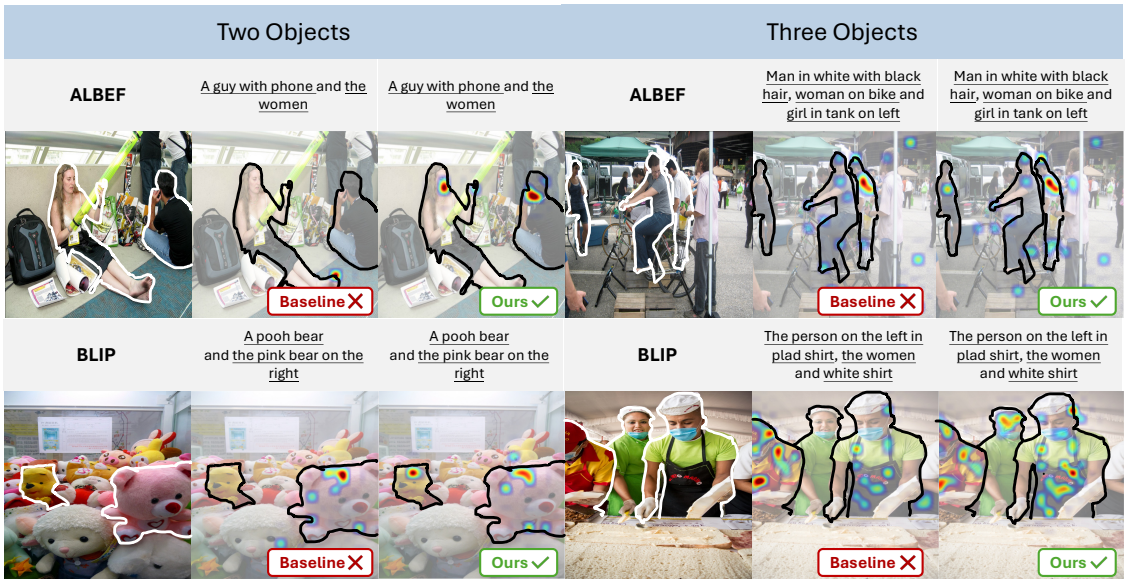


Figure 4: **Qualitative results for contrastive models.** White/black outlines indicate the ground-truth referred objects. With CompART, the models produce improved attention maps that ground multiple referred objects simultaneously.

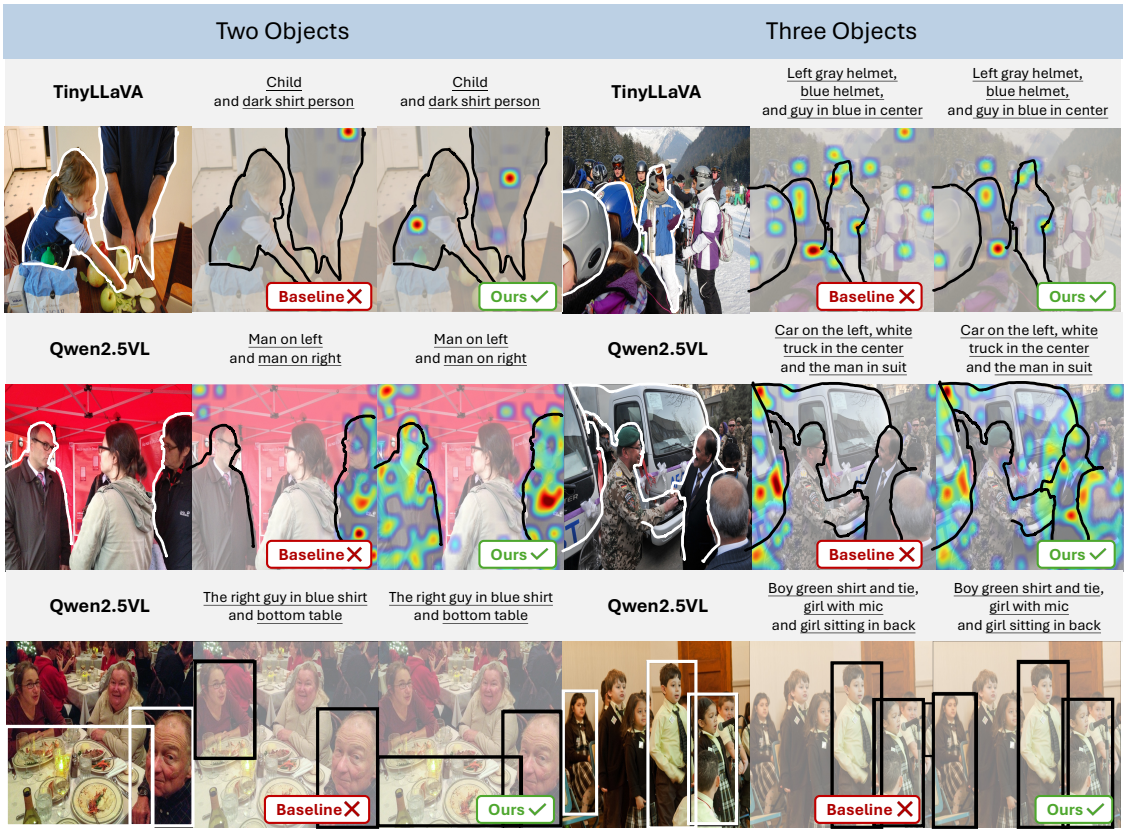


Figure 5: **Qualitative results for generative models.** White/black outlines indicate the ground-truth referred objects. With CompART, the models produce improved attention maps that ground multiple referred objects simultaneously. Unlike TinyLLaVA, Qwen2.5VL can directly produce bounding boxes for the queried objects; therefore, we present qualitative analysis for both point estimation and bounding box estimation.

and Qwen2.5VL-3B. As shown in Figure 3, the 23$^{rd}$ layer in TinyLLaVA-0.5B exhibits the highest attention, while the 21$^{st}$ layers in Qwen2.5VL-3B show the highest attention energy. Accordingly, we use the attention from the respective layers for TinyLLaVA-0.5B and Qwen2.5VL-3B to compute the corresponding phrase-localization maps.

### 3.5 Qualitative Results

Figure 4 and Figure 5 present qualitative analyses for both contrastive and generative VLMs. The results show that base VLMs fail to correctly localize all regions corresponding to multiple objects in a query, and often attention collapses to a single region. In contrast, CompART substantially improves object grounding performance with improved attention, not only in the two-object setting—on which the model is trained—but also generalizes beyond this setting to more complex scenarios involving three-object grounding.

### 3.6 Failure Analysis

Figure 6 presents failure cases of CompART. We notice CompART can still fail to rectify base models' capabilities in complex scenes with many visually similar objects, where grounding requires fine-grained discrimination or counting. As shown in Figure 6, ALBEF and Qwen2.5-VL fail to identify targets such as the "third toilet from the bottom on the right", "the velvet devil" or the "fourth cow from the left." These examples suggest that remaining errors are mainly due to fine-grained disambiguation and counting, rather than a general inability to attend to multiple objects, which we plan to explore in future.

Figure 6: **Examples of failure cases.** CompART can fail to rectify base models' capabilities in complex scenes with many visually similar objects, where grounding requires fine-grained discrimination or counting.

## 4 Conclusion

In this paper, we identify a consistent performance gap where both contrastive and generative VLMs, despite succeeding at single-object localization, struggle to attend to multiple entities simultaneously when presented with composite queries. This attention collapse limits the utility of VLMs in complex, real-world vision tasks. To mitigate this, we propose CompART, a lightweight and architecture-agnostic method that leverages the inherent compositionality of natural language as supervision. By decomposing captions into object-centric phrases and regularizing the model's attention maps through our proposed *composition loss*, we encourage models to treat composite references as the sum of their constitutes. Our results across four diverse architectures and four benchmarks demonstrate that CompART significantly improves grounding accuracy for multi-object queries without requiring additional dense annotations. Furthermore, the improved grounding leads to better performance on diverse downstream VQA tasks. We hope this work encourages further exploration into improving multi-object visual grounding.

## References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, 2022.

Jacob Andreas, Andreas Vlachos, and Stephen Clark. Semantic parsing as machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 47–52, 2013.

Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, pp. 4253–4262, 2020.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv e-prints*, pp. arXiv–2502, 2025.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6763–6782, 2024.

Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11165–11174, 2023.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, pp. 14455–14465, 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 433–442, 2019.

Silin Cheng, Yang Liu, Xinwei He, Sebastien Ourselin, Lei Tan, and Gen Luo. Weakmcn: Multi-task collaborative network for weakly supervised referring expression comprehension and segmentation. In *CVPR*, pp. 9175–9185, 2025.

Shih-Han Chou, Zicong Fan, James J Little, and Leonid Sigal. Semi-supervised grounding alignment for multi-modal feature learning. In *2022 19th Conference on Robots and Vision (CRV)*, pp. 48–57. IEEE, 2022.

Qiyuan Dai and Sibei Yang. Curriculum point prompting for weakly-supervised referring image segmentation. In *CVPR*, pp. 13711–13722, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, pp. 91–104, 2025.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. Learning to parse and translate improves neural machine translation. *arXiv preprint arXiv:1702.03525*, 2017.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6904–6913, 2017.

Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C Berg, and Vicente Ordonez. Improved visual grounding through self-consistent explanations. In *CVPR*, pp. 13095–13105, 2024.

Mir Rayat Imtiaz Hossain, Leonid Sigal, and James J Little. Framework-agnostic semantically-aware global reasoning for segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 999–1009, 2024.

Mir Rayat Imtiaz Hossain, Mennatullah Siam, Leonid Sigal, and James J. Little. Segmentation from attention: Training-free layer selection and one-shot tuning for segmentation in VLMs. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=a5lAwubXro.

Yizhi Hu, Zezhao Tian, Xingqun Qi, Chen Su, Bingkun Yang, Junhui Yin, Muyi Sun, Man Zhang, and Zhenan Sun. Remerec: Relation-aware and multi-entity referring expression comprehension. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4010–4019, 2025.

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *ICCV*, pp. 4067–4077, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Qing Jiang, Xingyu Chen, Zhaoyang Zeng, Junzhi Yu, and Lei Zhang. Rex-thinker: Grounded object referring via chain-of-thought reasoning. *arXiv preprint arXiv:2506.04034*, 2025.

Luo Jiayun, Siddhesh Khandelwal, Leonid Sigal, and Boyang Li. Plug-and-play, dense-label-free extraction of open-vocabulary semantic segmentation from vision-language models. *CVPR*, 2024.

Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *CVPR*, pp. 2681–2690, 2023.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *CVPR*, pp. 9339–9350, 2025.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/kim21k.html.

Alexandros Komninos and Suresh Manandhar. Dependency based embeddings for sentence classification tasks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1490–1500, 2016.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24, 2011.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, pp. 9579–9589, 2024.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, pp. 70–88. Springer, 2024.

Thao Minh Le, Vuong Le, Sunil Gupta, Svetha Venkatesh, and Truyen Tran. Guiding visual question answering with attention priors. In *CVPR*, pp. 4381–4390, 2023.

Jungbeom Lee, Sungjin Lee, Jinseok Nam, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In *ICCV*, pp. 21870–21881, 2023.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models, 2023a. URL https://arxiv.org/abs/2311.17092.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023b.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022a.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900. PMLR, 2022b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023c.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *ECCV*, 2020.

Pengyue Lin, Ruifan Li, Yuzhe Ji, Zhihan Yu, Fangxiang Feng, Zhanyu Ma, and Xiaojie Wang. Triple alignment strategies for zero-shot phrase grounding under weak supervision. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4312–4321, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pp. 23592–23601, 2023a.

Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *ICCV*, pp. 22124–22134, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023c.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024a. URL https://arxiv.org/abs/2310.03744.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pp. 38–55. Springer, 2024b.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019.

Yaxin Luo, Jiayi Ji, Xiaofu Chen, Yuxin Zhang, Tianhe Ren, and Gen Luo. Apl: Anchor-based prompt learning for one-stage weakly supervised referring expression comprehension. In *ECCV*, pp. 198–215. Springer, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions, 2016. URL https://arxiv.org/abs/1511.02283.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025. URL https://arxiv.org/abs/2504.05299.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, pp. 13009–13018, 2024.

Daniel Reich and Tanja Schultz. On the role of visual grounding in vqa. *arXiv preprint arXiv:2406.18253*, 2024.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, pp. 817–834. Springer, 2016.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *ICCV*, pp. 2591–2600, 2019.

Tal Shaharabany and Lior Wolf. Similarity maps for self-training weakly-supervised phrase grounding. In *CVPR*, pp. 6925–6934, 2023.

Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *NeurIPS*, 35:28222–28237, 2022.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, pp. 15638–15650, June 2022.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *CVPR*, pp. 8465–8474, 2021.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *CVPR*, pp. 14090–14100, 2021.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. *NeurIPS*, 32, 2019.

Yongliang Wu, Shuliang Zhao, and Wenbin Li. Phrase2vec: phrase embedding based on parsing. *Information Sciences*, 517:100–127, 2020.

Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, pp. 5945–5954, 2017.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *AAAI*, volume 37, pp. 10637–10647, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024a. URL https://arxiv.org/abs/2407.10671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zaiquan Yang, Yuhao Liu, Jiaying Lin, Gerhard Hancke, and Rynson Lau. Boosting weakly supervised referring image segmentation via progressive comprehension. *Advances in Neural Information Processing Systems*, 37:93213–93239, 2024b.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *CVPR*, pp. 14358–14368, 2025.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pp. 69–85. Springer, 2016.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pp. 1307–1315, 2018.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *CVPR*, pp. 14141–14151, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*, pp. 19–35. Springer, 2024a.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*, pp. 14227–14238, 2024b.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.