Learning Disentangled and Hierarchical Representation for Linguistic Compositionality in Language-based Object Detection

Anonymous ACL submission

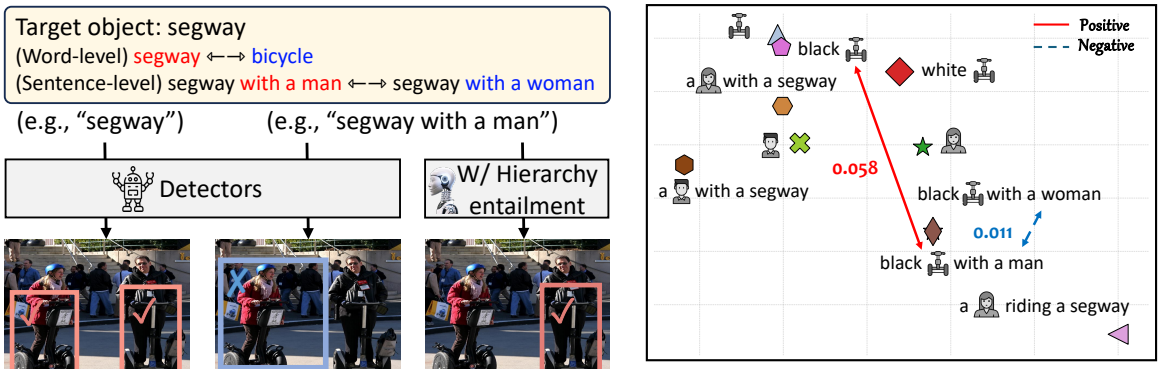
Abstract

Vision-language models (VLMs) have advanced multimodal perception, demonstrated by open-vocabulary object detection with simple language queries. State-of-the-art VLMs still struggle to handle complex queries involving descriptive attributes and relational clauses. To address this problem, we propose restructuring linguistic representations according to the hierarchical relations within sentences for language-based object detection. A key insight is the necessity of disentangling textual tokens into core components—objects, attributes, and relations—and subsequently aggregating them into hierarchically structured sentence-level representations. Building on this principle, we introduce the TaSe (**Ta**lk in Pieces, **Se**e in Whole) framework with three main contributions: (1) a hierarchical synthetic captioning dataset spanning three tiers from category names to descriptive sentences; (2) the three-component disentanglement module guided by a novel disentanglement loss function, transforms text embeddings into subspace compositions; and (3) aggregating disentangled components into hierarchically structured embeddings with the guide of proposed hierarchical objectives. Experimental results under the OmniLabel benchmark show a 24% performance improvement, demonstrating the importance of linguistic compositionality.

1 Introduction

Vision-language (VL) understanding is a long-standing and fundamental problem that requires learning modality-specific representations and cross-modal alignment. Foundational VLMs such as CLIP (Radford et al., 2021a) have leveraged web-scale image-text pairs to learn generic VL representations, achieving strong generalization performance on tasks such as image classification and image-text retrieval. Building upon these advances, recent studies have actively explored grounding language queries into specific image regions (e.g., open-vocabulary object detection (Liu et al., 2024a; Zhao et al., 2024; Yin et al., 2025a). Many approaches (Liu et al., 2024b; Li et al., 2022b) distill the general VL knowledge embedded in foundational models into object detectors and have demonstrated remarkable results in detecting previously unseen object categories—commonly referred to as open-vocabulary object detection (Gu et al., 2021). Current VL detectors succeed when input queries are short and consist of simple category names, while they struggle to comprehend complex language queries and accurately localize the corresponding objects. We further analyze this behavior using GLEE (Wu et al., 2024), a state-of-the-art visual grounding foundation model (see Fig. 1a). The model reliably detects objects given simple noun phrases (e.g., "segway"). VLMs fail when faced with more complex and specific queries (e.g., "segway with a man").

To investigate the underlying cause of this limitation, we visualize sentence-level text features using t-SNE (Van der Maaten and Hinton, 2008) in Fig. 1b. Interestingly, we observe that although some sentences ("a segway with a man" vs. "a segway with a woman") refer to different target objects ("with a man" vs. "with a woman"), their embeddings remain close in the feature space due to shared tokens ("segway"). The contrasts with the ideal cases, where embeddings of distinct target objects should be well-separated, while those of the same object with different attributes should be closer for object detection ("a segway" vs. "a black segway"). Such a mismatch is attributed to VL detector objectives (Liu et al., 2024b; Li et al., 2022b), which rely on augmented negative caption supervision and encourage coarse text-region discrimination. Augmented caption supervision (Li et al., 2023b; Yuksekgonul et al., 2022) is effective for coarse-grained concept analysis but provides limited guidance for modeling fine-grained ambi-



(a) Detection results from w/o and w/ our hierarchy entailment

(b) t-SNE of sentence features from text encoder of VLMs

Figure 1: (a) VL detectors struggle with attributes or relations due to limitations in capturing fine-grained semantics from image-text similarity. (b) Limitations of text encoders in VLMs for compositional understanding.

guities among concepts. Addressing these ambiguities requires sentence embeddings that capture compositional semantics by disentangling text into meaningful components.

In this paper, we propose a novel framework that disentangles component-wise text features ("Talk in Pieces") and explicitly learns hierarchical knowledge ("See in Whole") from these disentangled representations to construct sentence-level understanding for language-based object detection. We refer to our framework as **TaSe** (**Ta**lk in Pieces, **Se**e in Whole). We begin by formally defining a hierarchical structure tailored for language-based object detection. Specifically, we design our new HiVG dataset, a three-tiered hierarchy, object–attribute–relation, where the first tier represents object category names (i.e., segway), the second tier adds descriptive attributes (i.e., black), and the last tier includes relational phrases that describe interactions or contexts (i.e., with a woman). We also construct negative samples following the same three-tier structure. This mitigates bag-of-words behavior driven by lexical overlap and supports reliable semantic discrimination between positive and negative prompts. Our approach builds on phrase grounding datasets like Visual Genome (Krishna et al., 2017), which provide densely annotated phrases associated with images and object regions. Using a large language model (LLM) (Dubey et al., 2024; Yang et al., 2025; Achiam et al., 2023), we abstract these phrases into a three-tier hierarchy—object, attribute, relation—by sequentially removing relational and attribute information in phrases to obtain the final object categories.

To learn compositional sentence representations from the HiVG dataset, we disentangle it into the three components ("Talk in Pieces"): object, attribute, and relation. The disentangled design maps text representations into semantic subspaces and captures hierarchical structure through separated semantic units. We further design a lightweight learnable attention module for the TriDe (Three-component disentanglement). The key idea of TriDe is to leverage the hierarchical structure of the HiVG dataset to contrast component-wise tokens, allowing targeted tokens to be adjusted without loss of meaningful information. Then, we guide the model to learn linguistic representations that capture these levels of abstraction. This facilitates learning of sentence context enriched with descriptive attributes and relational clauses. We introduce a hierarchical aggregation method ("See in Whole") based on sentence-level hierarchy entailment.

To summarize, our main contributions are as follows: 1) We present an efficient hierarchical data generation pipeline that abstracts dense existing phrases into an explicit hierarchical structure of object–attribute–relation. 2) We introduce a framework for disentangling the three core components and employ the TriDe loss to guide this process. 3) We propose a method for learning disentangled and hierarchical representations that capture sentence-level inductive biases and generalize across VL detectors. With hierarchical learning on our generated dataset HiVG, our model significantly outperforms strong baselines, including state-of-the-art VL detectors, on challenging language-based object detection benchmarks such as OmniLabel (Schulter et al., 2023) and D3 (Xie et al., 2023).

2 Related Works

2.1 Language-based Object Detection

Language-based object detection aims to locate and identify objects in images using free-form text.

One of the leading approaches is to transfer the pre-trained model and align images and texts using contrastive learning (Gao et al., 2024; Li et al., 2022b, 2023b; Park et al., 2024). Contrastive learning enhances compositionality in VLMs by capturing relationships with contextual entities and improves the understanding of object relationships (Li et al., 2023b; Minderer et al., 2022a; Gu et al., 2021; Liu et al., 2024a). GLIP (Li et al., 2022b) proposes to add deep fusion layers between different modalities and learn a language-aware visual representation based on reformulated alignment scores.

Existing approaches overlook the need for contextualized sentence-level understanding of VL text embedding. For example, APE (Shen et al., 2024); GLEE (Wu et al., 2024); DINO-X (Ren et al., 2024); and Zeng et al. (2024) explore VLM alignment challenges and highlight the need to improve reasoning capabilities in multimodal large language models (MLLMs). These works investigate model capabilities from restricted VL perspectives, with a primary focus on fine-grained textual details and inter-object relationships. VL detectors still struggle to align images with syntactically intricate language queries (Wang et al., 2023), underscoring the need for a more grounded contextual understanding of text. While several works have explored disentanglement in visual components (Bengio et al., 2013; Wang et al., 2024; Qi et al., 2024), disentangled and hierarchical textual representations remain largely underexplored for VL models, particularly in the context of language-based object detection.

2.2 Hierarchical Entailment for Vision-Language Models

To better represent the embedding space of VLMs, hyperbolic learning has highlighted the need to capture hierarchical structures and relationships in multimodal data. Hyperbolic learning was formulated on the Poincaré ball by Ganea et al. (2018), learning entailment relations between embedded objects. The formulation now extends the more common Lorentz model as Lou et al. (2020) due to its computationally heavy Gyrovector operations. Hyperbolic learning maps the embedding into an entailment cone (EC) to represent hierarchical entailment in a continuous space. Recent studies investigated the use of the EC embedding for vision tasks (Atigh et al., 2022; Kong et al., 2024; Khrulkov et al., 2020), multimodal learning (Desai et al., 2023; Hong et al., 2023; Pal et al., 2024), and synthetic data generation (Kong et al., 2024). However, the hyperbolic manifold needs to transpose features from Euclidean to hyperbolic and requires additional hyperparameter configurations. Alper and Averbuch-Elor (2024) propose radial embedding (RE) optimization for learning hierarchical representations directly in Euclidean space. Inspired by this approach, we extend RE optimization to language-based object detection via sentence-level hierarchical representation learning.

3 TaSe: Disentangled and Hierarchical Text Representation Learning for Language-based Object Detection

This section introduces TaSe (**Ta**lk in Pieces, **Se**e in Whole), a framework for disentangling and hierarchical aggregation. Specifically, our approach comprises three components: 1) the HiVG dataset (Sec. 3.1), a synthetic dataset re-captioned from VG to capture hierarchical entailment relations; 2) disentangling text representations into object, attribute, and relation subspaces to align semantic pieces within sentences (Sec. 3.2); and 3) a hierarchical aggregation method to represent contextualized sentence embedding based on disentangled tokens (Sec. 3.3). Fig. 2 outlines how TaSe learns contextualized sentence representations for language-based object detection.

3.1 HiVG: Hierarchy Captioning Pipeline

Open-vocabulary detectors often rely on keywords and fail to separate target objects from their attributes and relations, owing to the absence of hard textual negatives that reflect linguistic hierarchy. To address this problem, we propose a **Hi**erarchical captioning pipeline that re-captions the **V**isual **G**enome dataset (**HiVG**) by leveraging LLMs (i.e., Llama-3-8B (Dubey et al., 2024), Qwen3-8B (Yang et al., 2025), and GPT-4-nano (Achiam et al., 2023)) and lexical databases (e.g., WordNet (Miller, 1995) and ConceptNet (Speer et al., 2017)). HiVG is a synthetic dataset constructed by spanning from category names to descriptive sentences and structuring hierarchical captions into three tiers: objects, attributes, and relations. Each caption in the Visual Genome (Krishna et al., 2017) annotation is transformed into three positive (e^+) and negative tiers (e^-) where e follows the notation introduced in Sec. 3.3.

- Tier 1. Category names (object): containing the class name (e.g., **jacket** (e_1^+) and **rainbow** (e_1^-)).
- Tier 2. Enriched descriptions (w/ attribute): adding an attribute to the object (e.g., **blue** (e_2^+)

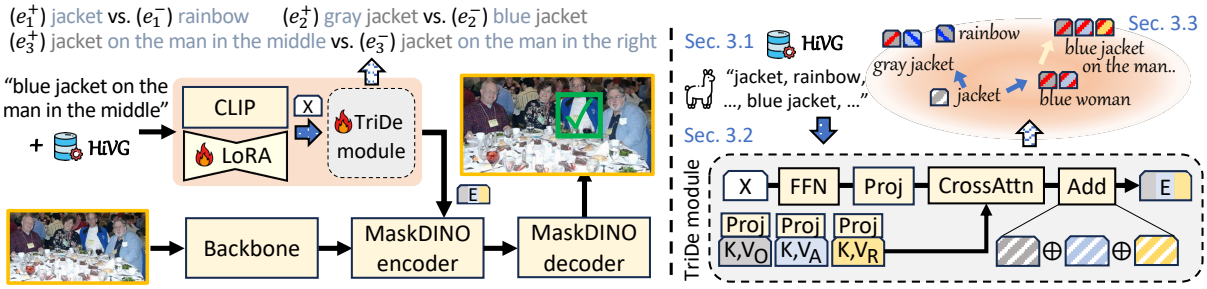


Figure 2: The overall framework of TaSe. (Left) The text encoder is fine-tuned with LoRA (Hu et al., 2021) and the TriDe module to restructure text representations. (Right) **Top:** Hierarchy aggregated embeddings with HiVG, where the recaptioned dataset passes through the TriDe module to learn linguistic hierarchy. **Bottom:** Architecture of the TriDe module.

and **red** (e_2^-)) for learning fine-grained linguistic compositionality.

- Tier 3. Contextual understanding (w/ attribute and relation): emphasizing the relationships between objects by injecting a relation into the second-tier caption (e.g., **on the man in the middle** (e_3^+) and **on the man in the right** (e_3^-)).

The core idea disentangles the category names from descriptive components (e.g., attributes and relations) in the text embedding space. The model focuses on the detection target rather than descriptive components to mitigate false positives. Further details are provided in the Appendix A.

3.2 Talk in Pieces: Component-Wise Text Disentanglement

Textual descriptions typically contain not only descriptive attributes but also complex relational structures, which cause false positives in language-based object detection. To address this, we propose the TriDe module to disentangle text embeddings into meaningful subspaces, which adaptively refines these components to enhance semantic representation.

Text embedding. We extract text features using CLIP text encoder with low-rank adaptation (LoRA) (Hu et al., 2021) for efficiently evolving text embedding from the text encoder. Let $\{v_i, t_i\}_{i=1}^{B}$ be a batch of image–text pairs. The text embedding $\mathbf{X} = \mathcal{T}_\theta(t_i)$, where $\mathrm{X} \in \mathbb{R}^{B\times T\times d_{\text{model}}}$, is obtained by the text encoder of CLIP. Here, B, T, and d_{model} denote the batch size, number of tokens, and embedding dimension, respectively. A text projection layer maps the input into $\mathrm{X} \in \mathbb{R}^{B\times T\times D}$, where D denotes the text embedding dimension.

Component-wise disentanglement. We disentangle text representations into three components—objects, attributes, and relations. This design mirrors the three-tier structure of HiVG and facilitates the learning of effective contextualized sentence embeddings. We adjust learnable vectors $\mathrm{V_O}, \mathrm{V_A}, \mathrm{V_R} \in \mathbb{R}^{B\times T\times D}$ to disentangle the text embedding into the three components. We employ a multi-head cross-attention layer between the learnable vectors and text embedding X. Let FFN, LN, and Proj denote the feed-forward network, layer normalization, and projection layer, respectively. The TriDe module is defined as follows:

$$\begin{aligned} \mathbf{X} &= \mathrm{LN}(\mathrm{Proj}(\mathrm{X} + \mathrm{FFN}(\mathrm{X}))), \\ [\mathrm{O}, \mathrm{A}, \mathrm{R}] &= \mathrm{CrossAttn}(\mathrm{X}, [\mathrm{V_O}, \mathrm{V_A}, \mathrm{V_R}]), \\ \mathbf{E} &= \mathrm{pooling}(\mathrm{FFN}(\mathrm{LN}(\mathrm{O} + \mathrm{A} + \mathrm{R}))), \end{aligned} \tag{1}$$

where O, A, R, and E represent the object, attribute, relation components, and restructured text embedding. Note that E is employed to learn hierarchical entailment for contextualized sentence representations in Sec. 3.3. The $\mathrm{CrossAttn}(\mathrm{Q}, \mathrm{V})$ with the scaling factor d_k is defined as follows:

$$\mathrm{CrossAttn}(\mathrm{Q}, \mathrm{V}) = \mathrm{Softmax}\Big(\frac{\mathrm{Proj}(\mathrm{Q})\mathrm{Proj}(\mathrm{V})^T}{\sqrt{d_k}}\Big)\mathrm{Proj}(\mathrm{V}). \tag{2}$$

Here, Q denotes the *query*. The *key* and *value* are identical in this formulation, so only the V term is used. E is mean-pooled to encode the sentence-level representation used for alignment with visual embeddings. The detailed algorithm is described in Appendix Alg. 1.

Objective for component disentanglement. We found that directly increasing the angular distance for negative captions amplifies the effect of shared lexical content, leading to unstable gradients and sharp loss escalation. The disentangle–aggregate design separates semantic components and recombines them, and yields stable optimization of angular relationships in the embedding space. Components are aligned with their positive counterparts, while negatives are enforced to remain distant according to their tier. Let t be the tier of HiVG, and

the disentanglement of text embedding is adjusted $\mathcal{L}_{\text{TriDe}}$ as follows:

$$\begin{aligned}\mathcal{L}_{\text{TriDe}} &= \lambda \textstyle\sum_{(i,j)\in\{(O,A),(O,R),(A,R)\}} |\mathbf{i}\cdot\mathbf{j}| \\ &+ \textstyle\sum_{X\in\{O,A\}} \sum_{t\in T_X} \Big(m + \cos(X_t^+, X_{t+1}^+) \\ &- \cos(X_t^+, X_t^-)\Big) + \Big(m - \cos(R_l^+, R_l^-)\Big).\end{aligned} \tag{3}$$

$\mathcal{L}_{\text{TriDe}}$ and m represent compositional loss for the TriDe module and margin, respectively. λ represents a hyperparameter for stability adjustment. Minimizing correlation among components promotes an inductive bias toward semantically grounded object representations.

3.3 See In Whole: Sentence-level Hierarchical Aggregation

We design a hierarchy aggregation method of disentangled features to serve fine-grained semantic distinction, which helps capture contextual meaning beyond simple word-level perception.

3.3.1 Background: hierarchical entailment in Euclidean space

The goal of hierarchical entailment is to learn general concepts by representing entailment relations via low-dimensional embeddings (Ganea et al., 2018). Conventional contrastive learning models learn relations between positive and negative samples, whereas the radial embedding (RE) objective (Alper and Averbuch-Elor, 2024) encodes hierarchical structure through exterior angles with respect to a reference point. Two key advantages of this representation learning with the RE are: 1) modeling sentence-level structure; and 2) learning compositional generalization without requiring transformation into sphere space. Let Ξ denote the exterior angle in radians and $r \in \mathbb{R}^d$ denote a root embedding, the exterior angle between embedding a and b in the RE objective is defined:

$$\begin{aligned}\Xi\langle a, b\rangle &= \cos^{-1}\left(\frac{\mathbf{a}'\cdot\mathbf{b}'}{\|\mathbf{a}'\|\|\mathbf{b}'\|}\right) \leq \pi. \\ \text{where } \mathbf{a}' &= \mathbf{a} - r, \quad \mathbf{b}' = \mathbf{b} - r - \mathbf{a}'.\end{aligned} \tag{4}$$

The value $\langle a, b\rangle_\Xi \in [0, \pi]$ is bounded. Given a root embedding r and a reference embedding of a text embedding e, the objective function of the RE is represented as follows:

$$\mathcal{L}_{RE} = \sum_i \left(\langle e_i^+, e_{>i}^+\rangle_\Xi - \langle e_i^+, e_i^-\rangle_\Xi\right). \tag{5}$$

Here, e_i^+ and $e_{>i}^+$ denote distinct positive embeddings, and e_i^- is a corresponding negative. The objective $\mathcal{L}_{RE}$ encourages smaller exterior angles between positive pairs while enforcing larger angles between positive and negative pairs. This deviation reflects a misalignment from the reference anchor r (frozen) and corresponds to a larger angular distance in the embedding space.

3.3.2 Local-reference hierarchy entailment

The proposed hierarchy entailment objective consists of two components: 1) Geometry-preserving entailment and 2) Local-reference entailment contrast (see Fig. 3).

Geometry-preserving entailment (Fig. 3 Left). We introduce a regularization term to preserve the inherent embedding space and define two objective functions for hierarchy entailment. Since angular-only supervision may induce directional bias and degrade representational fidelity, we regularize the embedding space to retain knowledge.

Local-reference entailment contrast (Fig. 3 Right). Existing contrastive learning is formulated using exterior angles, where sentence pairs are trained with respect to a fixed reference-conditioned point. A key consideration in sentence-level contrastive learning with a text encoder is that a negative e_t^- is also positive with respect to the previous tier e_{t-1}^+. This creates an inconsistent hierarchy in which the same sample can be treated as negative at the current tier but positive at the previous tier. To address this hierarchical inconsistency, we position samples in opposite directions with respect to a dynamic reference (denoted as Local ref.) in r, rather than a fixed root (denoted as Global ref.).

Hierarchical objectives. Let l denote the tier of HiVG in the object-attribute-relation hierarchy. Let e_t be an embedding at tier t of HiVG, e_t^+ its positive counterpart at tier l, and e_t^- a negative sample at the same tier. We use two complementary losses—(i) alignment across tiers and (ii) within-tier discrimination—defined compactly as

$$\begin{aligned}\mathcal{L}_{\text{H}^+} &= \textstyle\sum_{t=1}^{l} \Xi\langle e_t^+, e_{t+1}^+\rangle + \Xi\langle e_t^+, e_{t+1}^-\rangle, \\ \mathcal{L}_{\text{H}^-} &= \textstyle\sum_{t=1}^{l} \Xi\langle e_t^+, e_t^-\rangle, \\ b' = s(b-r) - a', &\quad s = 1\ (H^+),\ -1\ (H^-),\end{aligned} \tag{6}$$

with normalization $e = \frac{e-r}{\|e-r\|+\epsilon}, \quad e \in \{e^+, e^-\}$. Note that normalization maintains directional consistency independent of embedding scale. r is adapted to expose directional differences not captured by global alignment and to account for locally meaningful variation. The proposed formulation minimizes directional deviation from the reference point, which helps preserve the intrinsic structure of the embedding space while enforcing separation between negative pairs.

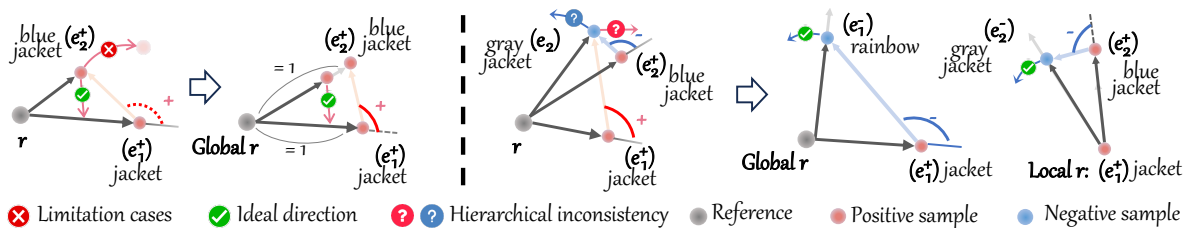


Figure 3: Limitations of RE and our solutions. (Left) Geometry-preserving entailment. Variations in vector magnitude can distort angular measurements, leading to inconsistent hierarchies. Normalization ensures the hierarchy depends on directional relationships. (Right) Local-reference entailment contrast. Under the global reference, higher tiers contain positive ("blue jacket,") and negative ("gray jacket,") samples that share semantic components ("jacket"), which leads to hierarchical inconsistency; a local reference better separates embeddings.

3.3.3 Training objectives

For learning contextualized features, the objective function of hierarchical and disentangled representation learning, $\mathcal{L}_{\text{TaSe}}$, is defined as:

$$\mathcal{L}_{\text{TaSe}} = \mathcal{L}_{\text{TriDe}} + \mathcal{L}_{\text{H}^+} + \mathcal{L}_{\text{H}^-}. \quad (7)$$

The loss function is formulated as a weighted sum of classification, localization, and hierarchy losses updated on the text encoder and VL fusion layers as follows:

$$\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_{bbox} + \mathcal{L}_{giou} + \mathcal{L}_{\text{TaSe}}, \quad (8)$$

where $\mathcal{L}_{class}$ represents Focal loss (Lin et al., 2017), $\mathcal{L}_{bbox}$ represents L1 loss, and $\mathcal{L}_{giou}$ represents generalized intersection over union (GIoU) loss (Rezatofighi et al., 2019).

4 Experiments

This section compares our method with baselines. The following sections provide the implementation details (Sec. 4.1), the main results for performance comparison (Sec. 4.2), and ablation studies conducted to analyze the results in three benchmark datasets (Sec. 4.3). Additional experimental details and results can be found in Appendix B.

4.1 Experimental settings

Implementation details. We build our method based on GLEE (Wu et al., 2024), a pre-trained foundation model composed of MaskDINO (Li et al., 2023a) and CLIP (Radford et al., 2021b) text-image encoders. GLEE was selected as a baseline because it is a powerful vision–language foundation model that performs well on many benchmarks (e.g., RefCOCO (Yu et al., 2016), yet still faces challenges in contextualizing text embeddings. In Sec. C.3, we demonstrate that TaSe can be applied to other diverse detector architectures.

Benchmarks and evaluation metrics. We evaluate language-based object detection on four benchmarks in a **zero-shot setting**. 1) D^3 (Xie et al., 2023) dataset is a widely used benchmark for visual grounding tasks. The dataset includes negative instances, multi-target scenarios, and long sentences. 2) OmniLabel (Schulter et al., 2023) dataset is an open-vocabulary detection dataset including both positive and negative descriptions of spatial relations, actions, and numeracy. Additionally, we evaluate on gRefCOCO (He et al., 2023) and OdinW (Li et al., 2022a) to assess performance on referring expression tasks. We use mean average precision (mAP) to validate the language-based object detection task.

4.2 Main Results

We investigate the impact of object detection on disentanglement and hierarchical representation learning through a set of analyses.

Key Findings. Hierarchical text representation learning mitigates sentence-level ambiguity and improves linguistic compositionality for vision-language alignment. **(1)** Hierarchical supervision improves compositionality by structuring sentence embeddings with entailment relations (Tab. 4); and **(2)** disentangling shared context and token-specific semantics supports hierarchy learning (Tab. 3).

Zero-shot generalization performance. The proposed model improves upon the baseline while adapting only the text embeddings (Tab. 1). Compared to GLEE, which served as the vision foundation model, we observe improvements of $+3.1$ in D^3 and $+5.2$ in OmniLabel AP scores. The AP scores in OmniLabel show that hierarchical learning improves performance in zero-shot evaluation, and the gains observed on open-vocabulary benchmarks further demonstrate its effectiveness. We also compare TaSe with baseline models on the gRefCOCO (He et al., 2023) and ODinW (Li et al., 2022b) in Tables 10 and 11. On ODinW, our model achieved an average improvement of $+4.9$ AP. On gRefCOCO, following (He et al., 2023), TaSe demonstrates improvements in object detec-

Model	D^3 (default)			OmniLabel (default)		
	FULL	PRES	ABS	AP	AP_c	AP_d
OFA-L (Wang et al., 2022)	4.2	4.1	4.6	2.7	2.7	2.6
MDETR (Kamath et al., 2021)	-	-	-	-	4.7	9.1
OWL (Minderer et al., 2022b)	9.6	10.7	6.4	8.0	15.6	5.4
UNINEXT (Lin et al., 2023)	21.6	23.7	15.4	22.2	27.2	18.8
G-DINO (Liu et al., 2024b)	20.7	20.1	22.5	19.3	23.6	16.4
GEN (Zhao et al., 2024)	21.4	20.6	23.7	22.2	27.2	18.8
GLIP (Li et al., 2022b)	19.1	18.3	21.5	19.3	23.6	16.4
GLEE-Lite* (Wu et al., 2024)	27.6	26.8	30.1	21.7	36.6	15.4
GLIP + DesCo (Li et al., 2023b)	24.2	22.9	27.8	23.8	27.4	21.0
GLEE-Lite + DesCo	28.3	27.6	30.3	24.6	**37.3**	18.3
GLEE-Lite + TaSe	**30.7**	**29.9**	**33.2**	**26.9**	36.8	**21.2**

Table 1: Evaluation on D^3 and OmniLabel. Detailed qualitative results are provided in Sec. C.1. Performance comparisons with MLLM-based baselines and latency results are reported in Tab. 9 and Tab. 8, respectively.

Methods	val	testA	testB	ODinW
MCN (Luo et al., 2020)	28.0	32.3	26.8	-
VLT (Ding et al., 2021)	36.6	40.2	30.2	-
MDETR	42.7	50.0	36.5	25.1
UNINEXT† (Lin et al., 2023)	58.2	46.4	42.9	-
GLEE-Lite (Wu et al., 2024)	66.7	73.0	61.8	44.4
GLEE-Lite + TaSe	**69.6**	**76.6**	**67.7**	**46.9**

Table 2: Zero-shot evaluation results on gRefCOCO (val, testA, testB) and the average performance across 13 datasets in ODinW (see Sec. C.1 for detailed results).

tion performance with textual queries.

Does hierarchical learning truly improve linguistic reasoning? We examine whether hierarchical learning of sentence embeddings improves linguistic compositionality. To evaluate the methodology, we compare our method with DesCo (Li et al., 2023b), a caption augmentation method. We randomly sample sentences from HiVG for Desco configuration. The selected sentence is concatenated with the original caption in either order (c [SEP] c^- or c^- [SEP] c), where c and c^- denote the original and negative captions, respectively. The augmented components are pooled separately and then averaged. The DesCo improves the GLEE model +2.4 AP on D^3 and +2.3 on OmniLabel.

Qualitative results. We present two qualitative examples in Fig. 4 to illustrate the effectiveness of our hierarchical entailment learning. The first case shows negative cases containing an attribute (i.e., blue), and the second case presents a positive case with attributes and descriptive relations (e.g., numeracy and text in the image). In the first case, GLEE incorrectly assigns a confident score of 0.92 to the language query. In the second case, GLEE predicts all bikes as positives, including those that do not correspond to the queried bike. On the other hand, TaSe captures contextual information related to category names, and hierarchical entailment helps reduce false positives.

	D^3	OmniLabel
Where-to-apply disentanglement		
w/o disentangling	27.8 (+1.0)	26.2 (+4.5)
Token-level disentangling	**30.7 (+3.1)**	**26.9 (+5.2)**
∟ Identity initialization	30.7 (+3.1)	26.9 (+5.2)
∟ Uniform initialization	28.8 (+1.2)	26.4 (+4.4)
After pooling	28.6 (+1.4)	22.6 (+0.9)
How-to-apply disentanglement		
Self-attention	29.0 (+1.4)	25.5 (+3.8)
Learnable query	29.6 (+2.0)	24.7 (+3.0)
Learnable key & value	**30.7 (+3.1)**	**26.9 (+5.2)**
Effectiveness of disentangling components (# of learnable vector)		
1 (w/o disentanglement)	29.4 (+1.8)	25.4 (+3.7)
2 (Object + Attribute)	29.5 (+1.9)	26.5 (+4.8)
3 (Object + Attribute + Relation)	**30.7 (+3.1)**	**26.9 (+5.2)**

Table 3: Ablation studies on disentangled representations (Sec. 3.2) with hierarchy entailment.

	D^3	OmniLabel
GLEE-Lite (Wu et al., 2024)	27.6	21.7
+ LoRA with HiVG data	27.5 (-0.1)	21.7 (+0.0)
Comparison of contrastive learning objectives with LoRA		
+ $\mathcal{L}_{CL}$	26.9 (-0.7)	23.9 (+2.2)
+ $\mathcal{L}_{RE}$	27.5 (-0.1)	25.7 (+4.0)
+ $\mathcal{L}_{H^\pm}$ in Eq. 6	28.6 (+1.0)	**26.2 (+4.5)**
∟ w/o $\mathcal{L}_{H^-}$	**28.8 (+1.2)**	24.8 (+3.1)
∟ w/o $\mathcal{L}_{H^+}$	27.7 (+0.1)	23.6 (+1.9)
+ Reverse objective: -$\mathcal{L}_{H^\pm}$	26.7 (-0.9)	22.1 (+0.4)
TriDe module ablation with LoRA on HiVG data		
+ TriDe module with $\mathcal{L}_{TriDe}$	**28.8 (+1.2)**	25.8 (+4.1)

Table 4: Ablation studies on hierarchy objectives (Sec. 3.3) and the TriDe module. Values in parentheses indicate performance changes over the GLEE baseline. $\mathcal{L}_{CL}$ and $\mathcal{L}_{RE}$ represent contrastive loss (Oord et al., 2018) and RE embedding objective, respectively.

4.3 Ablation Studies

Is negative caption augmentation alone sufficient? We ablate the hierarchy entailment loss to compare learning with positive and negative pairs. Fine-tuning with positives improves OmniLabel by +3.1 AP, while negatives yield +1.9 AP. We further conduct the objective and observe performance degradation when positives and negatives are aligned in reversed directions (Tab. 4). Hierarchical entailment improves sentence-level embeddings and strengthens semantic alignment with visual representations and performance. While prior captioning methods (Li et al., 2023b; Doveh et al., 2023) focus on negatives, our approach achieves further performance gains via positive hierarchy.

What advantages does our hierarchical loss offer over traditional contrastive loss? In Tab. 4, we conduct an ablation study to validate the effectiveness of our hierarchical loss. Within the base

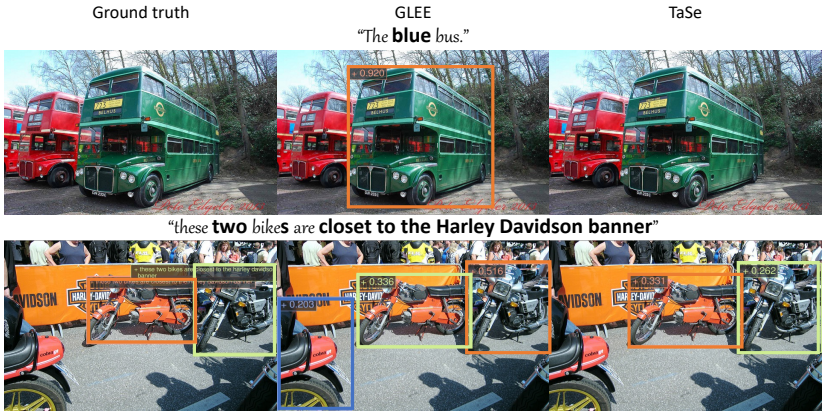


Figure 4: Qualitative analysis on OmniLabel. We compare the baseline and TaSe on language descriptions with attributes and relations.

Measure	Base	+TaSe
Grounding DINO-B (Liu et al., 2024b)		
D^3	20.7	**24.3** (+3.6)
OmniLabel	19.3	**28.6** (+7.3)
Latency	114	123 (+9)
GLIP-T (Li et al., 2022b)		
D^3	19.1	**21.5** (+2.4)
OmniLabel	19.3	**19.9** (+0.6)
Latency	122	138 (+16)
Qwen3-VL-2B (Yang et al., 2025)		
D^3	16.0	**41.6** (+25.6)
OmniLabel	3.8	**30.4** (+26.6)
Latency	582	583 (+1)

Table 5: Generalization and latency of TaSe across detectors

setting (GLEE with LoRA), we evaluate three configurations: (1) conventional contrastive loss ($\mathcal{L}_{\text{CL}}$); (2) the RE objective ($\mathcal{L}_{\text{RE}}$); and (3) the proposed method. For sentence-level hierarchy aggregation, our loss $\mathcal{L}_{\text{H}}$ outperforms contrastive baselines. Our reference-based hierarchy induction aligns them hierarchically and improves sentence-level meaning and performance.

Ablation on where and how to disentangle in text representation. Tab. 3 reports ablation studies analyzing the design choices of the TriDe module. The results show that the disentanglement location plays a key role in controlling the granularity of the representation. We compare three modes for constructing compositional text embeddings: (1) no disentanglement, (2) token-level disentanglement (before pooling), and (3) disentangled embeddings after pooling. Preserving text positional bias during token-level disentanglement improves performance and generalization. For the design of the TriDe module, we further examine three attention variants: (1) standard self-attention, (2) learnable queries, and (3) key–value configurations. The key-value setting improves performance by using keys to identify component types and values to encode the corresponding semantic contents.

Is it beneficial to disentangle the representation into three components? Conventional detectors (Li et al., 2023b; Yuksekgonul et al., 2022) disentangle objects and attributes, whereas we separate representations into three components—object, attribute, and relation—and evaluate their effectiveness. As shown at the bottom of Tab. 3, overall, the performance of three-component disentanglement is higher than two-component (i.e., object and attribute) disentanglement. Three-component disentanglement introduces an inductive bias for complex linguistic structures, making longer sentences more robust to negative samples. Finer-grained text embeddings better expose structural information in representations, motivating dataset composition criteria based on objects, attributes, and relations.

Generalization performance across diverse detector architectures. Experiments with GroundingDINO (Liu et al., 2024b), GLIP (Li et al., 2022b), and Qwen3-VL-2B (Yang et al., 2025) are conducted to validate the generalizability of TaSe. The results show consistent performance improvements (Tab. 5). The proposed approach improves performance on D^3 (FULL AP) and OmniLabel (AP_d) with only marginal latency overhead. Details of the learning setup are provided in Appendix Sec. B.3. TaSe improves compositional reasoning and robust VL alignment in the shared embedding space and generalizes across model architectures.

Additional analyses on HiVG. We report the performance of TaSe on HiVG datasets generated using different LLMs in Appendix A.3, and provide a human evaluation of HiVG in Appendix A.4.

5 Conclusion

This study proposed a disentanglement and hierarchy aggregation framework for constructing contextualized sentence representations within language-based object detection. TaSe improved linguistic compositionality, which serves as a key learning factor and leads to competitive results. The results indicated that hierarchy entailment allows learning the granularity of text embedding to distinguish descriptive sentences. This study highlights the need for further exploration of the underlying linguistic compositionality in future studies for downstream vision tasks.

Limitations

This section outlines the limitations of our proposed method. We propose a disentanglement and hierarchical aggregation framework to enhance linguistic compositionality in VLMs. The three disentangled components (i.e., object–attribute–relation) may not represent the optimal separation of language semantics. Despite this limitation, we adopt this decomposition to analyze the behavior of the embedding space. Our analysis shows that the proposed objective induces compositional structure and semantic disentanglement in the learned representations. In the detection task we consider, disentangling objects, attributes, and relations in the embedding space reduces false positives and improves performance.

Ethics Statements

As for VL-based detection models, we access all open-source LMs via the Hugging Face hub (Wolf et al., 2020). The number of parameters is presented in Tab. 7. All associated licenses permit user access for research purposes, and we have agreed to follow all terms of use. We validate the effectiveness of our proposed approach using the D3 (Xie et al., 2023), OmniLabel (Schulter et al., 2023), GRefCOCO (He et al., 2023), and ODinW (Li et al., 2022b) reference datasets. These datasets are sourced from publicly available resources on legitimate websites and do not contain any sensitive data.

We recaptioned our dataset, utilizing the publicly available Llama 3 (Dubey et al., 2024), GPT-4-nano (Achiam et al., 2023), and Qwen3 (Yang et al., 2025) models released on the Hugging Face hub (Wolf et al., 2020). Additional statistics and details of the dataset are presented in Appendix A.

During the preparation of this work, the authors used ChatGPT (Singh et al., 2025) in order to improve readability. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Morris Alper and Hadar Averbuch-Elor. 2024. Emergent visual-semantic hierarchies in image-text representations. In *European Conference on Computer Vision*, pages 220–238. Springer.

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. 2022. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4453–4462.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. 2023. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2021. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16321–16330.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, and 1 others. 2023. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36:76137–76150.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2024. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*, pages 1646–1655. PMLR.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. 2023. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*.

Jie Hong, Zeeshan Hayder, Junlin Han, Pengfei Fang, Mehrtash Harandi, and Lars Petersson. 2023. Hyperbolic audio-visual zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7873–7883.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. 2020. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6418–6428.

Fanjie Kong, Yanbei Chen, Jiarui Cai, and Davide Modolo. 2024. Hyperbolic learning with synthetic captions for open-world detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16762–16771.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and 1 others. 2022a. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. 2023a. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050.

Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. 2023b. Desco: Learning object recognition with rich language descriptions. *Advances in Neural Information Processing Systems*, 36:37511–37526.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, and 1 others. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Fangjian Lin, Jianlong Yuan, Sitong Wu, Fan Wang, and Zhibin Wang. 2023. Uninext: Exploring a unified architecture for vision recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3200–3208.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, and Christopher De Sa. 2020. Differentiating through the fréchet mean. In *International conference on machine learning*, pages 6393–6403. PMLR.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, and 1 others. 2022a. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, and 1 others. 2022b. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2025. Gpt-5 system card. Accessed: 2025-11-21.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. 2024. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*.

Kwanyong Park, Kuniaki Saito, and Donghyun Kim. 2024. Weak-to-strong compositional learning from generative models for language-based object detection. In *European Conference on Computer Vision*, pages 1–19. Springer.

Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. 2024. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8693–8702.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, and 1 others. 2024. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Samuel Schulter, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, and 1 others. 2023. Omnilabel: A challenging benchmark for language-based object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11953–11962.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. 2024. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13193–13203.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11998–12008.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2024. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9677–9696.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. 2024. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3795.

Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. 2023. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36:79095–79107.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. 2025a. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14358–14368.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. 2025b. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14358–14368.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. 2024. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14151.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. 2024. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*, pages 405–422. Springer.

Shiyu Zhao, Long Zhao, Yumin Suh, Dimitris N Metaxas, Manmohan Chandraker, Samuel Schulter, and 1 others. 2024. Generating enhanced negatives for training language-based object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13592–13602.

For a comprehensive understanding of the TaSe and HiVG dataset, we provide this Appendix. The following table of contents gives a concise overview and directs readers to specific sections of interest.

Contents

A. HiVG Dataset Construction 13
A.1 Data Generation Pipeline 13
A.2 Dataset Statistics 14
A.3 Data Generation with Different Large Language Models 14
A.4 Human Evaluation for Generated Captions 14

B. Experimental Setup 15
B.1 Implementation Details 15
B.2 Model Size and Budget 15
B.3 Experimental Setup for Additional Detectors 18

C. Additional Results 18
C.1 Detailed Quantitative Results 18
C.2 Additional Qualitative Results 19
C.3 Additional Ablation Studies 19
C.4 Pseudo Code for TaSe 19

D. Discussions 19
D.1 How are embeddings structured after disentanglement? 19
D.2 Comparative Analysis of Text Embeddings across State-of-the-Art Vision-Language Models 23
D.3 Exterior Angle Analysis 23
D.4 Analysis of Hierarchical Objectives 23
D.5 Failure Cases 24

A HiVG Dataset Construction

A.1 Data Generation Pipeline

Our hierarchical captioning pipeline is illustrated in Fig. 5. We generate a 10 K re-captioned dataset from the Visual Genome dataset. We first filter out all Visual Genome captions with fewer than six words to ensure sufficient semantic richness. To reduce hallucination, we enforced the hierarchical format "Tier 3=>Tier 2=>Tier 1" using the delimiter "=>". Captions that did not conform to this structure are filtered and regenerated. As shown in the Fig. 9, GPT-4 (Achiam et al., 2023) data demonstrate the highest performance, with Llama3 (Dubey et al., 2024) second and Qwen3 (Yang et al., 2025).

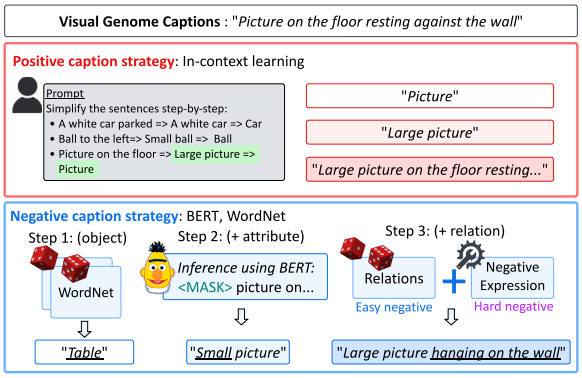


Figure 5: Overview of generating positive and negative captions. Positive captions are derived using in-context learning based on Llama3 (Dubey et al., 2024).

A.1.1 Creating Hierarchical Positive Captions.

For positive captions, we leverage in-context learning using LLaMA (Dubey et al., 2024) to transform the remaining Visual Genome captions into a three-tier hierarchical structure. To enhance attribute diversity in captions, we draw on common visual concepts (Huang et al., 2023; Lin et al., 2024) to define a set of visual attributes (spatial, color, number, and size) and randomly select one to modify the object's attributes. Based on our experiments, we set the randomization ratio to 50%. To fully align with the in-context learning demonstration format, samples that lack attributes or relations are also incorporated into the learning process, following the approach suggested by Min et al. (2022).

A.1.2 Creating Hierarchical Negative Captions.

Positive tier 1. One of the challenges in language-based object detection is effectively handling negative samples, which often report higher false negative rates compared to false positives. To address the issue, we focus on both re-captioning hard and easy negative samples. For hard negative samples in tier 1, we replace the positive object with an antonym (e.g., *man* is replaced by *woman*) or a random concrete noun. Easy negative samples are generated by selecting nouns from ImageNet-1K (Deng et al., 2009) classes and lexical databases such as WordNet and ConceptNet. Additionally, we insert a negative determiner to the object (e.g., *dog* is switched to *no dog*).

Positive tier 2. We reuse the same set of visual attributes from the positive captions but replace them with semantically distinct attributes (e.g., *tall*

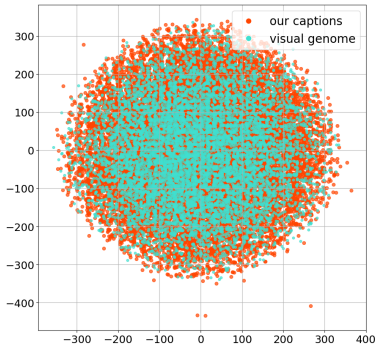


Figure 6: t-SNE plot comparing the clustering of our captions with Visual Genome captions. Visual Genome dataset and HiVG dataset are shown in the plot. Our captions exhibit a broader spread in the 2D space, indicating greater semantic diversity compared to the original Visual Genome captions.

building is replaced by *short building*) to generate hard negative samples. We use LLM-based mask-filling (Liu et al., 2019) to diversify attributes by substituting them with contextually plausible but semantically different terms or by prepending "not" to create hard negatives (e.g., tall building → not tall building).

Positive tier 3. We use a set of common spatial relations (e.g., above and beside) and object-specific relations from the Visual Genome dataset. These pre-defined object-specific relations ensure that the relation is contextually relevant to the object in question. To introduce hard negatives, we apply absence-based transformations by replacing affirmative relations with their negative counterparts (e.g., *with* is replaced by *without*). We also leverage LLaMA's (Dubey et al., 2024) sentence completion capability to generate further relation diversity.

Captions that do not adhere to the hierarchical structure (the arrow symbol "=>" is missing, or the second tier does not contain the previous tier) are filtered out and regenerated. By re-captioning using a multi-tiered set of positive and negative captions, our approach is intended to facilitate the learning of hierarchical representations, thereby improving linguistic compositionality.

A.2 Dataset Statistics

The combined dataset consists of 101,897 LLaMA-based annotations, with the majority containing 4.79 ± 26.8 words, as shown in Fig. 7. The Qwen-based dataset contains 98,261 annotations with 4.80 ± 15.0 words, while the GPT-based dataset includes 101,852 annotations with 3.09 ± 6.62 words. GPT generates the shortest captions on average. Caption length distributions across positive and negative samples remain similar within each tier, with negative captions in the LLaMA and Qwen datasets being slightly longer on average ($\approx$ 1 word). Tier 3 contains the most linguistically diverse and structurally rich captions, which may benefit semantic reasoning. We report examples from our hierarchical dataset, HiVG, in Fig. 8. Leveraging Visual Genome data for re-captioning, we create a more diverse dataset by incorporating a wider range of classes and the LLM and other datasets.

A.3 Performance Comparison based on Captions Generated by Different LLMs

We conduct a sensitivity study to examine whether the choice of large language models affects the generated dataset and the downstream performance. We constructed a caption dataset containing positive and negative examples using three LLMs: Llama3-8B (Dubey et al., 2024), Qwen3-8B (Yang et al., 2025), and GPT-4-nano (Achiam et al., 2023). Dataset construction yields 101,897 samples for the original pipeline, 98,261 samples with Qwen, and 101,852 samples with GPT.

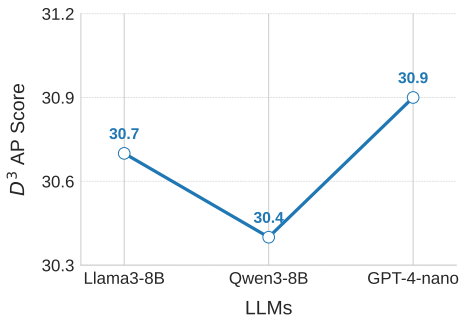


Figure 9: Evaluation of how captions generated by different LLMs affect object detection performance

All models use the same training configuration for fair comparison. GPT-4-nano (Achiam et al., 2023) achieves the best performance on D^3 with a score of 30.9. Performance remains consistent across datasets generated by different LLMs, which indicates low sensitivity to LLM choice. Fig.7 shows that the GPT-generated dataset exhibits a near-Gaussian word-length distribution. The balanced caption lengths likely contribute to the superior performance.

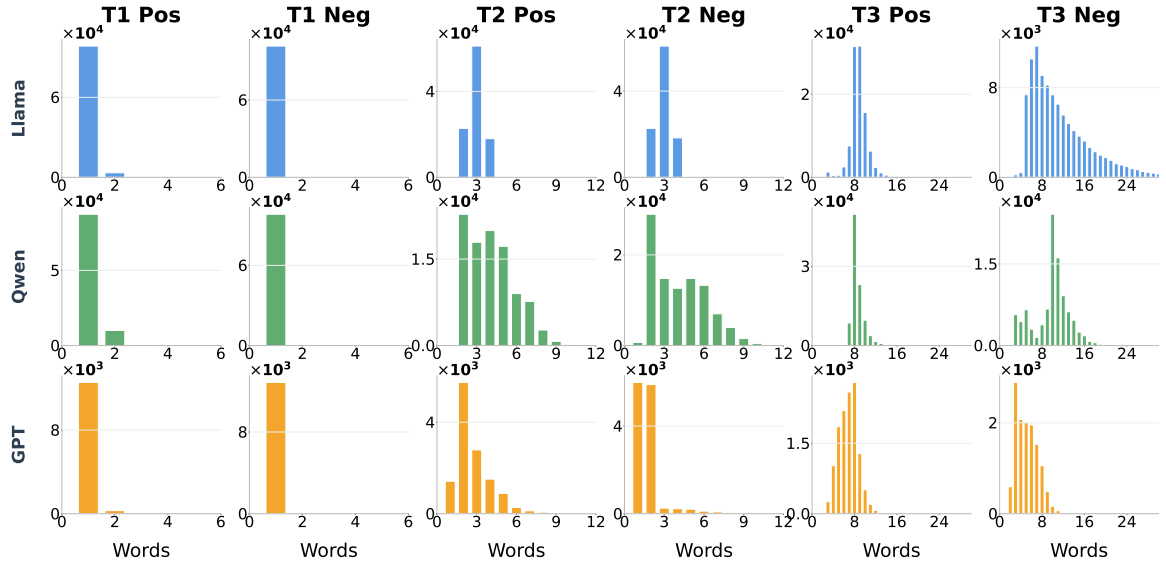


Figure 7: Statistics of the HiVG dataset generated with different LLMs. Rows correspond to datasets generated by LLaMA-3-8B, Qwen3-8B, and GPT-4-nano. Columns show the word-count distributions of hierarchical caption tiers, including T1 (category), T2 (attribute), and T3 (relation), each separated into positive (Pos) and negative (Neg) captions. The distributions remain consistent across LLMs.

A.4 Human Evaluation for Generated Captions

To validate the quality of HiVG data, we conducted a human evaluation. Specifically, we aimed to assess whether the three-tier hierarchies in the LLM-generated captions accurately reflect our intended design. To this end, we asked 50 participants to verify whether the captions in each tier describe the same underlying objects. The caption sets were randomly assigned to participants, and a total of 500 samples were evaluated. A screenshot of the evaluation interface is provided in Fig. 10.

The quality of the proposed HiVG dataset is remarkably high, supporting the effectiveness of our data-generation pipeline. 98.8% of samples followed the intended hierarchy; only 6 of 500 depicted different target objects. The six failure cases are listed below:

Tier 1	Tier 2	Tier 3
pen	a white pen	papers and a white pen on top of tab
cpu	a black cpu unit	a black cpu unit on the floor under a computer monitor
man	a black man	a bartender wearing a mustache and a tie
boat	the yellow boat	the yellow boat masts in the water
woman	woman and little girl	woman and little girl reading on a beige couch
numbers	red numbers	red numbers 2008 on the side of the building

B Experimental Setup

Baselines This paper compares the language-based object detection models on OFA (Wang et al., 2022), OWL (Minderer et al., 2022b), G-DINO (Liu et al., 2024b), GLIP (Li et al., 2022b), UNINEXT (Lin et al., 2023), Desco (Li et al., 2023b), GLIP-GEN (Zhao et al., 2024), and GLEE (Wu et al., 2024).

B.1 Implementation Details

We sample 16 images per batch and further select 6 corresponding sentences per image for hierarchy learning. We employ AdamW (Loshchilov and Hutter, 2017) to optimize the trainable model, using a learning rate of 1×10^{-4} for the TriDe module and 5×10^{-6} for LoRA. For the comparison with baselines, our detector was trained for 60 K iterations, the same as in the ablation studies. Following Alper and Averbuch-Elor (2024), the RE loss was set with a positive-to-negative ratio of 10:4. In case of $\mathcal{L}_{\mathrm{H}}$, we conduct experiments with positive-to-negative ratio of 2:1. The value of γ is set to 0.1.

B.2 Model Size and Budget

For fine-tuning the pre-trained GLEE, we only train LoRA layers, TriDe module, and VL fusion layers. We train a total of 5,447,680 parameters, which is an efficient approach that reduces memory usage by 2.93% of the model parameters. On average,

Tier 1	Tier 2	Tier 3	Tier 1	Tier 2	Tier 3
car	a red car	red car are parked on the side of the road	pen	A red pen	A red writing utensil ontop of a notebook
discard	a sports car	a red car parked on the side of a road: BMW 5 Sedan (G30)	bedstraw	A blank pen	A red pen next to a yellow legal pad with the words 'I don't know what to write about it

Tier 1	Tier 2	Tier 3	Tier 1	Tier 2	Tier 3
table	wooden table	The wooden table and chairs are made of wood	building	Yellow building	Yellow building opposite were the men are standing
string orchestra	summary table	wooden tableware set wooden tableware set dining room chairs	equestrienne	Not yellow building	Yellow building witout the letter l on it

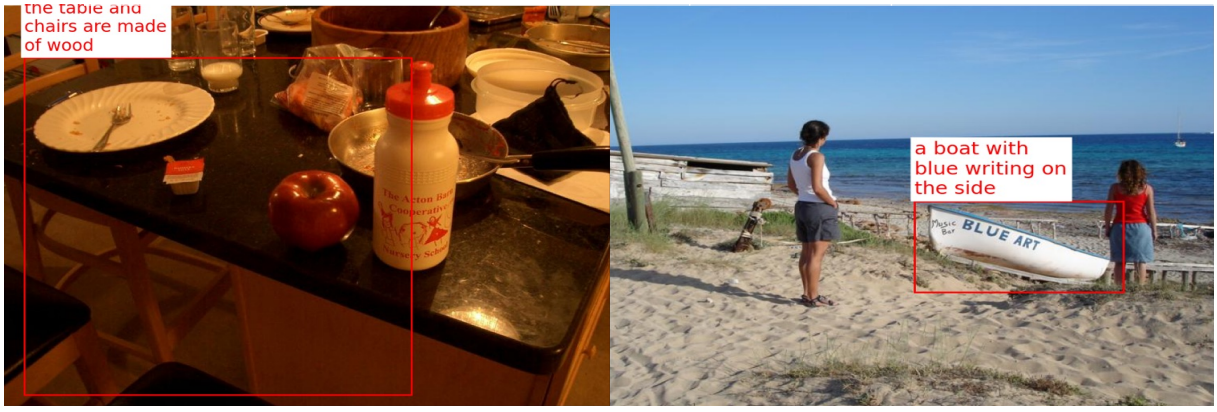


Figure 8: Re-captioning data examples

Human Evaluation for Caption Generation

Each question presents a set of images structured in a hierarchical manner:

- **Tier 1: Object only**
  Example: *Segway*
- **Tier 2: Object + Attribute**
  Example: *Black Segway*
  Represents characteristics of the object, such as **size, color, shape, or spatial information**.
- **Tier 3: Object + Attribute + Relation**
  Example: *Black Segway with a man*
  Represents a **relationship** between the object and another entity.

Your task is to determine whether the three tiers describe the **same object** or **different objects**.
Please examine the images carefully and indicate whether they refer to the **same underlying object across tiers**.

Does Tier 1, Tier 2, and Tier 3 refer to the same object? *
Tier 1: window
Tier 2: A green window
Tier 3: A green window in the front of the building

- ◯ Same object
- ◯ Different objects

```javascript
function generateMultipleForms() {
  const ss = SpreadsheetApp.openById("");
  const sheet = ss.getSheetByName("");
  const baseFormId = "";

  const values = sheet.getRange(2,5,sheet.getLastRow()-1,1).getValues();
  const texts = values.map(row => row[0]).filter(t => t !== "");

  for (let i = 1; i <= 50; i++) {
    const newFormFile = DriveApp.getFileById(baseFormId).makeCopy(`Survey ${i}`);
    const newForm = FormApp.openById(newFormFile.getId());

    const shuffled = texts.sort(() => Math.random() - 0.5).slice(0, 10);

    newForm.getItems().forEach(item => newForm.deleteItem(item));

    shuffled.forEach(q => {
      const item = newForm.addMultipleChoiceItem();
      item.setTitle(q)
          .setChoices([
            item.createChoice("Same object"),
            item.createChoice("Different objects")
          ])
          .setRequired(false);
    });

    Logger.log(`📌 Created: ${newForm.getTitle()} → ${newForm.getEditUrl()}`);
  }
}
```

Figure 10: Google form-based questionnaire for human evaluation

Params.	Value
Batch size	4
Optimizer	AdamW
Optimizer momentum	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Rank of LoRA	16
scaling factor of LoRA	16
learning rate of LoRA	5e-6
learning rate of TriDe	1e-4
Input resolution	800×800
loss of class ($\mathcal{L}_{class}$)	4.0
loss of bbox ($\mathcal{L}_{bbox}$)	5.0
loss of gIoU ($\mathcal{L}_{giou}$)	2.0
loss of TaSe ($\mathcal{L}_{\text{TaSe}}$)	5.0
λ	0.1

Table 6: Hyperparameters setting

Layers	# of Params. (M)
Image backbone	23.5
Text encoder	126.3
Detector	31.5
LoRA	0.4
TriDe	1.9
VL	3.2
Trainable params.	5.4 (2.93%)

Table 7: Parameter comparison across networks

10 iterations are executed, requiring 62.2 ms per inference. The disentangling and aggregation step adds an additional 1.57 ms (2.3%) overhead. Experiments were conducted using 4 NVIDIA A6000 GPUs for model training.

B.3 Experimental Setup for Additional Detectors

We apply TaSe to Grounding DINO-B (Liu et al., 2024b), GLIP-T (Li et al., 2022b), and Qwen3-VL-2B (Yang et al., 2025) to evaluate its model-agnostic effectiveness. We apply LoRA (Hu et al., 2021) and the disentanglement network, and tune only these modules. The learning rate is selected from $\{10^{-4}, 5\times10^{-5}, 10^{-5}\}$, with 5×10^{-5} yielding the best validation performance across all models. The rank and alpha are fixed to 16. All models are trained only on the proposed HiVG data using the original training loss combined with $\mathcal{L}_{\text{TaSe}}$. We compute $\mathcal{L}_{\text{H}^{\pm}}$ using pooled embeddings to learn hierarchical structures. For Qwen3-VL-2B, we use the corresponding tokenized span instead.

GLIP Training Target Construction. Each class in GLIP-T needs a positive map constructed from the start and end indices. For the negative captions in HiVG, we randomly select one tier and concatenate the preceding and following strings to construct the positive map for training.

Qwen3-VL Target Construction. We generate HiVG proposals and use their bounding boxes as prediction targets. We train the model with the cross-entropy loss and $\mathcal{L}_{\text{TaSe}}$.

C Additional Results

C.1 Detailed Quantitative Results

Detailed results on D^3 and OmniLabel. We report detailed results corresponding to the scores summarized in Tab. 1 and Tab. 2. The average performance on the D3 and Omnilabel datasets, along with the results across different text lengths, is presented in Tab. 9. Tab. 10 and Tab. 11 provide detailed results on the GRefCOCO (He et al., 2023) and ODinW (Li et al., 2022b) datasets, respectively. As shown in Tab. 5, the performance consistently improves not only in terms of the average score but also across increasing text lengths. To further analyze this trend, we conduct a more fine-grained evaluation by dividing the results into smaller text-length groups. As shown in the Tables, performance improvements are observed across multiple OOD datasets by adjusting only the text embeddings.

Performance comparison by text length. We further evaluate performance as a function of text length. As illustrated in Fig. 11, the proposed method achieves strong performance across all text-length ranges. Decomposing into only three components (object, attribute, and relation) effectively models long textual descriptions for detection tasks and improves performance across all caption lengths.

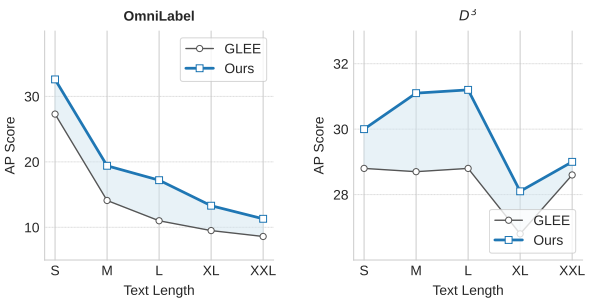


Figure 11: Performance analysis of Omnilabel and D^3 based on sentence length. S: [0, 3], M: [4, 7], L: [7, 10], XL: [10, 13], and XXL: [13, ∞].

Efficiency Comparison with MLLM Baselines. Recent MLLMs have shown strong zero-shot generalization ability (Yang et al., 2025; Chen et al., 2024). We compare inference latency with Qwen3-VL (Yang et al., 2025) and InternVL2 (Chen et al., 2024) on 100 samples to assess the efficiency of our model against MLLM baselines. Compared with these MLLM baselines, our model requires only 0.18B parameters and achieves substantially lower latency, demonstrating higher efficiency for detection-oriented inference. We emphasize the need to develop efficient detectors for real-time detection tasks.

Model	Params	FULL↑	PRES↑	ABS↑	Latency↓
Qwen3-VL-2B-Instruct (Yang et al., 2025)	2B	16.0	17.0	11.2	582
Qwen3-VL-4B-Instruct (Yang et al., 2025)	4B	24.3	25.8	19.7	1,791
InternVL2-8B-Instruct (Chen et al., 2024)	8B	9.8	11.0	6.2	932
GLEE + TaSe (Ours)	**0.18B**	**30.7**	**29.9**	**33.2**	**131**

Table 8: Comparison with MLLM-based baselines on D^3. Latency is measured in milliseconds.

C.2 Additional Qualitative Results

We present qualitative results to analyze the behavior of TaSe. For comparison, we evaluate text queries containing both attribute and relational cues. As shown in Fig. 12 and Fig. 13, we visualize the results of two scenarios containing an absent example. Given that sentences become longer, many VLMs focus on specific words, such as "running" to detect objects. For example, our model improves performance by capturing richer semantic information, such as the attribute "pink," and by understanding contextual meaning, such as recognizing "girl" as the subject of "running." In contrast, our model demonstrates greater robustness in detecting complex relations and predicts bounding boxes more accurately by better understanding object states and relative information.

C.3 Additional Ablation Studies

We conducted an ablation study on the model parameters. We compared and analyzed the effects of adjusting the LoRA (Hu et al., 2021) rank (see Fig. 14), the number of tokens (see Fig. 15), and the margin parameter (see Fig. 16) for the disentangled loss. Performance remains stable across different parameter settings. Positive and negative sentences in higher tiers partially share the embedding space due to hierarchical semantic structures. Strict disentanglement constrains representation learning. Token disentanglement and aggregation benefit from preserving the original token length.

Algorithm 1: Pseudo code for disentangling and hierarchical aggregating paradigm

Input: Image-text pair $(\mathrm{v}_i, t_i)_{i=1}^{B}$; Text encoder with LoRA $\mathcal{T}_\theta$; Learnable vectors $\mathbf{V}_\mathrm{O}, \mathbf{V}_\mathrm{A}, \mathbf{V}_\mathrm{R}$; Projection embedding δ; Vision backbone $\mathcal{V}_\psi$; MaskDINO f_ψ

Output: Bounding box and class Y; Total Loss $\mathcal{L}$;

for *each training iteration* **do**

$\mathbf{X}_\mathrm{v} = \mathcal{V}_\psi(\mathrm{v})$

$\mathbf{X} = \mathcal{T}_\theta(t) \cdot \delta$

`// TriDe: token-level disentangling`

$\mathbf{X} = \mathbf{FFN}(\mathbf{X})$

$\mathbf{O} = \mathrm{CrossAttn}(\mathbf{X}, \mathbf{V_O})$

$\mathbf{A} = \mathrm{CrossAttn}(\mathbf{X}, \mathbf{V_A})$

$\mathbf{R} = \mathrm{CrossAttn}(\mathbf{X}, \mathbf{V_R})$

$\mathbf{E} = \mathrm{Pool}(\mathrm{FFN}(\mathbf{O} + \mathbf{A} + \mathbf{R}))$

$\mathbf{Y}, \mathcal{L} = f_\psi(\mathbf{X}_\mathrm{v}, \mathbf{E})$

Update θ by minimize $\mathcal{L}$

Maintaining the standard sequence length of 77 tokens preserves positional information and stabilizes token representations during manipulation. Increasing the LoRA rank yields moderate performance improvements. Rank 16 provides the best balance between performance and parameter efficiency.

C.4 Pseudo Code for TaSe

We provide the details of the disentanglement modes employed for compositional learning as shown in Alg. 1. The first mode adopts traditional mean-pooling and uses the resulting representation for contrastive learning. The second and third modes involve disentanglement via a TriDe module, followed by contrastive learning based on the aggregated compositional embedding. Specifically, the second mode applies the TriDe module at the token level to leverage information across all tokens, whereas the third mode applies the module after pooling, focusing on sentence-level semantics.

D Discussions

D.1 How Are Embeddings Structured After Disentanglement?

While a few negative pairs still exhibit large angles, Fig. 17 confirms that positive and negative

Figure 12: Qualitative analysis on Omnilabel data (Schulter et al., 2023). We visualize and compare the results between GLEE and TaSe. We visualized the prediction results for both positive and negative captions of the same image.

Figure 13: Qualitative analysis on Omnilabel data (Schulter et al., 2023). We visualize and compare the results between GLEE and TaSe.

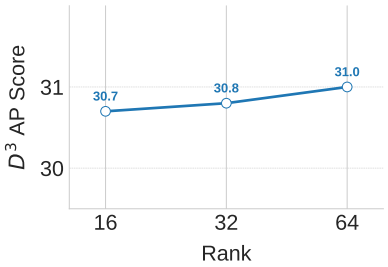


Figure 14: Analysis on the effect of different LoRA ranks

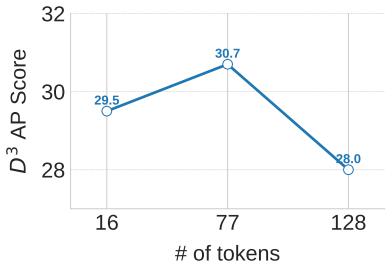


Figure 15: Analysis on the number of tokens

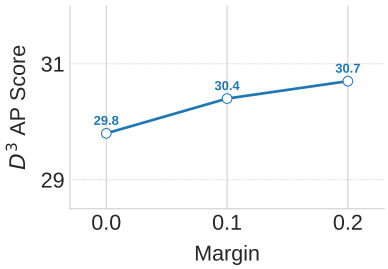


Figure 16: Analysis on the margin m in the $\mathcal{L}_{\text{TriDe}}$

Model	Param.	D^3 (default)			D^3 (length)				OmniLabel (default)			OmniLabel (length)		
		Full	Pres	Abs	S	M	L	XL	AP	AP_c	AP_d	S	M	L
Specialist models														
OFA-L (Wang et al., 2022)	0.5B	4.2	4.1	4.6	4.9	5.4	3.0	2.1	2.7	2.7	2.6	3.6	2.7	2.3
MDETR (Kamath et al., 2021)	0.1B	-	-	-	-	-	-	-	-	4.7	9.1	6.4	4.6	4.0
OWL (Minderer et al., 2022b)	0.2B	9.6	10.7	6.4	20.7	9.4	6.0	5.3	8.0	15.6	5.4	5.7	5.4	6.2
UNINEXT (Lin et al., 2023)	-	21.6	23.7	15.4	23.6	22.6	20.5	18.4	22.2	27.2	18.8	-	-	-
G-DINO-B (Liu et al., 2024b)	0.2B	20.7	20.1	22.5	22.6	22.5	18.9	16.5	19.3	23.6	16.4	29.4	14.8	8.2
GEN (Zhao et al., 2024)	-	21.4	20.6	23.7	28.1	24.5	17.4	11.5	22.2	27.2	18.8	-	-	-
GLIP-T (Li et al., 2022b)	0.2B	19.1	18.3	21.5	22.4	22.0	16.6	10.6	19.3	23.6	16.4	29.4	14.8	8.2
GLIP-T + DesCo (Li et al., 2023b)	0.2B	24.2	22.9	27.8	24.3	21.9	16.4	11.5	23.8	27.4	21.0	33.7	19.0	13.7
GLEE-Lite* (Wu et al., 2024)	0.2B	27.6	26.8	30.1	30.0	27.6	26.9	17.2	21.7	36.6	15.4	28.4	13.8	10.3
Generalist models														
Qwen3-VL-2B (Yang et al., 2025)	2B	14.9	14.5	16.1	11.0	18.1	13.8	11.5	-	-	3.9	3.3	4.6	3.9
InternVL2-8B (Chen et al., 2024)	8B	9.8	11.0	6.2	-	-	-	-	-	-	4.5	6.2	4.2	3.9
Griffon (Zhan et al., 2024)	13B	12.3	12.4	12.2	-	-	-	-	-	-	7.2	12.5	6.3	4.5
Groma (Ma et al., 2024)	7B	16.0	15.9	16.3	-	-	-	-	-	-	15.6	21.4	15.6	11.3
ROD-MLLM (Yin et al., 2025b)	7B	29.7	30.0	28.7	-	-	-	-	-	-	25.3	31.8	24.5	21.0
GPT-5.4-nano (OpenAI, 2025)	-	1.0	3.5	3.5	1.5	1.2	0.1	0.1	-	-	3.9	3.3	4.6	3.9
Ours														
GLEE-Lite + TaSe	0.2B	30.7	29.9	33.2	31.8	31.2	30.3	19.8	26.9	36.8	21.2	33.1	19.3	14.8
G-DINO-B + TaSe	0.2B	24.3	22.9	28.5	27.7	26.2	22.4	18.8	28.6	40.5	22.1	31.1	44.7	18.7
GLIP-T + TaSe	0.2B	21.5	21.2	22.7	26.4	23.3	18.8	16.0	21.5	24.3	19.9	34.6	17.3	10.5
Qwen3-VL-2B + TaSe	2B	41.6	43.0	37.3	42.7	44.1	38.3	38.7	-	-	30.4	40.3	29.1	25.1

Table 9: Evaluation on D^3 (Xie et al., 2023) and OmniLabel (Schulter et al., 2023). D^3 provides three types of descriptions: absence (ABS), presence (PRES), and full (FULL). Models are grouped into specialist models, generalist models, and ours. Param. indicates the publicly reported model size when available. For MLLM-based baselines, only AP_d is reported on OmniLabel. For OmniLabel, the final AP is computed as the geometric mean of category-level (AP_c) and description-level (AP_d) scores.

Methods	Visual Encoder	Textual Encoder	val		testA		testB	
			Pr@0.5	N-acc.	Pr@0.5	N-acc.	Pr@0.5	N-acc.
MCN (Luo et al., 2020)	DarkNet-53	GRU	28.0	30.6	32.3	32.0	26.8	30.3
VLT (Ding et al., 2021)	DarkNet-53	GRU	36.6	35.2	40.2	34.1	30.2	32.5
MDETR	ResNet-101	RoBERTa	42.7	36.3	50.0	34.5	36.5	31.0
UNINEXT† (Lin et al., 2023)	ResNet-50	BERT	58.2	50.6	46.4	49.3	42.9	48.2
GLEE (Wu et al., 2024)	RN50	CLIP	66.7	62.4	73.0	64.1	61.8	51.9
GLEE + TaSe	RN50	CLIP	**69.6**	**69.9**	**76.6**	**69.6**	**67.7**	**58.9**

Table 10: Zero-shot evaluation results on gRefCOCO (He et al., 2023). Following (He et al., 2023), Pr@0.5 (F1=1, IoU≥0.5) denotes a true positive when the predicted bounding box achieves an IoU greater than 0.5 with the ground truth. N-acc. represent positive rate (TN / Positive).

	Pascal VOC	Aerial Drone	Aquarium	Rabbits	Ego Hands	Mushrooms	Packages	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal	Average
GLEE	61.2	5.0	23.9	71.9	46.2	**57.8**	25.6	**56.8**	33.1	60.6	57.1	**25.3**	**52.5**	44.4
TaSe	**61.3**	**14.4**	**24.8**	**81.4**	**54.5**	20.8	**63.1**	50.6	**41.0**	**60.8**	**61.5**	18.9	46.7	**49.3**

Table 11: Zero-shot evaluation results across 13 datasets in ODinW (Li et al., 2022a).

pairs are effectively aligned in the embedding space. To examine whether this alignment follows the intended structure after disentanglement and hierarchical aggregation, we visualize the t-SNE projection of the trained TaSe embeddings. As shown in Fig. 18a, GLEE is dispersed around category names, whereas TaSe realigns embeddings around objects and preserves robust angular distance for negatives corresponding to attributes or relations. This is evident in the "segway" object, where the captions "black segway with a woman" and "black segway with a man" lie at different angles from the reference point "black segway."

D.2 Comparative Analysis of Text Embeddings across State-of-the-Art Vision-Language Models

VLMs are pre-trained for text–image alignment, and Fig. 1b shows that they often demonstrate bag-of-words behavior. Recent VLMs incorporate approaches such as MAE-based architectures (Fang et al., 2024) and LLM-connected visual encoders (Jiang et al., 2024) to improve representation quality. We investigate whether the text embeddings of these models improve object detection. For this analysis, we visualize text embeddings from SigLIP (Zhai et al., 2023) and E5-V (Jiang et al., 2024) using the same example shown in Fig. 18d.

Fig. 18d shows that SigLIP demonstrates embeddings that closely align across distinct objects (e.g., black segway with a man and black segway with a woman), similar to CLIP. For E5-V, the LLM-based embeddings were more object-centered (e.g., woman), yet white segway and black segway remained closer to each other than black segway with a man. This observation indicates that even advanced VLMs can retain ambiguities in object-level representation, as they rely on textual embedding similarity to correlate visual content.

D.3 Exterior Angle Analysis

We validate the effectiveness of our proposed hierarchical learning approach by visualizing the impact of the angular loss. For the experimental setup, we randomly initialize 50 two-dimensional embeddings and train them using the original hierarchy loss and our extended loss function.

Based on the results, the baseline method pushes both positive and negative samples away in the embedding space. The proposed method pulls positive samples closer and pushes negative samples farther apart. The proposed method moves text embeddings with stable optimization and limited training.

We further analyze the exterior angles of text embeddings after training in Fig. 17 and Fig. 19. Fig. 17 compares the embedding distributions of GLEE and TaSe. For GLEE, positive and negative exterior angles show substantial overlap, suggesting less distinct boundaries between positive and negative pairs. TaSe shows a more distinct separation of positive and negative pairs in angle space. To examine negative pairs with small angles, we compare angle distributions across semantic tiers in Fig. 19. Figures (a), (b), and (c) present Tier 1, Tier 2, and Tier 3 results. Each figure reports exterior angles between positive and negative samples. Tier 1 shows a large angular separation between positive and negative samples. Tier 2 and Tier 3 show smaller separations for negative samples. Shared embeddings and pre-trained angular structures explain this pattern. The proposed method increases angular separation compared to the initial configuration. Positive and negative embeddings show meaningful disentanglement after training.

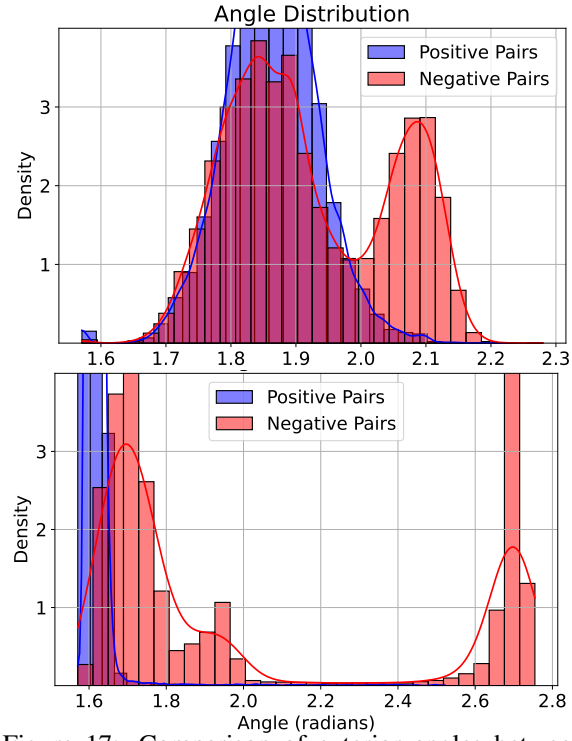


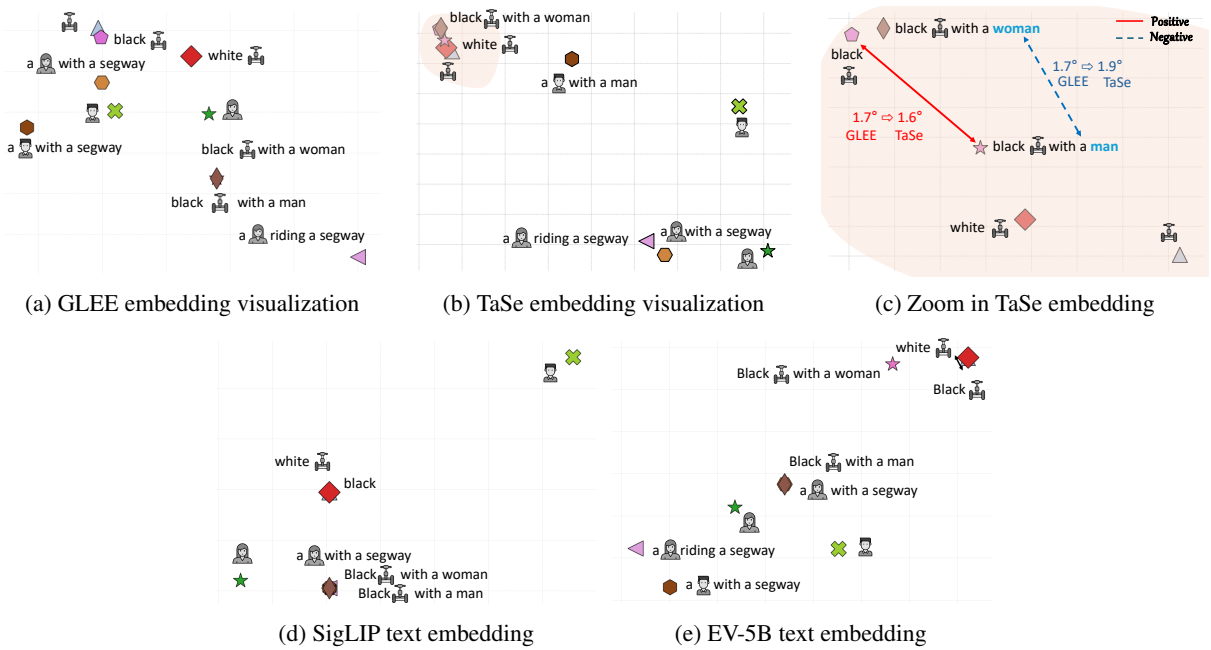
Figure 17: Comparison of exterior angles between GLEE and TaSe

D.4 Analysis of Hierarchical Objectives

To assess the efficacy of the proposed H^+ and H^- objectives in hierarchical representation learning, we performed experiments on synthetic datasets. Fifty pairs were randomly labeled as positive or negative with respect to a root, and the resulting embeddings were optimized using either the baseline or the proposed loss function with a fixed random seed. As shown in Fig. 20, random initialization fails to produce structured representations. The baseline objective causes pairs to drift away from both the root and each other. In contrast, the proposed objectives maintain proximity to the root while ensuring clear separation by clustering positive pairs and distancing negative ones.

D.5 Failure Cases

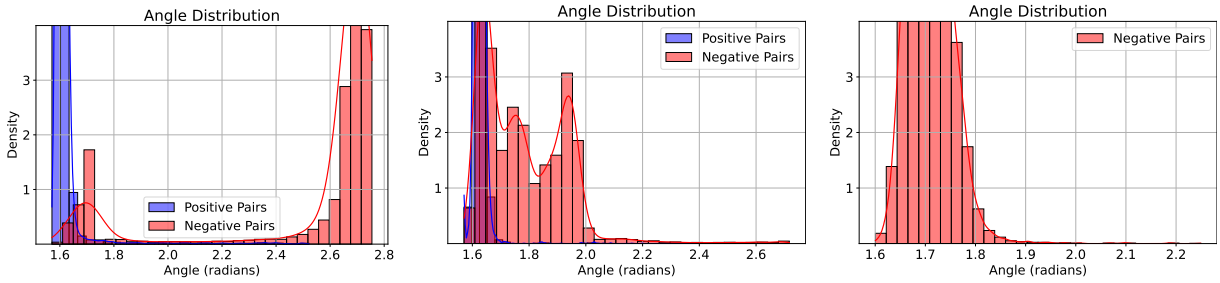
We present qualitative results that indicate directions for improving the model. As shown in Fig. 21, accurately detecting objects remains challenging when additional attributes or relations are included. Natural language contains complex structures, such as negation or compound structures, beyond the O-A-R components. These diverse forms make it difficult to understand finer-grained semantic meaning, and this remains an important research problem. TaSe reveals two major failure patterns: i) The model struggles with ambiguous expressions (e.g., looking toward the camera); and ii) Datasets such as RefCOCO contain only a single instance per object category, often resulting in predictions that merge multiple instances into one bounding box.



(a) GLEE embedding visualization (b) TaSe embedding visualization (c) Zoom in TaSe embedding

(d) SigLIP text embedding (e) EV-5B text embedding

Figure 18: Comparison of text embeddings across models and tiers. (a) GLEE embedding visualization. (b) TaSe embedding visualization. (c) Zoomed view of TaSe embedding. (d) SigLIP text embedding. (e) EV-5B text embedding.



(a) Tier 1 (b) Tier 2 (c) Tier 3

Figure 19: Exterior angle comparison across tiers. Each subplot shows the angular separation between positive and negative text embeddings.

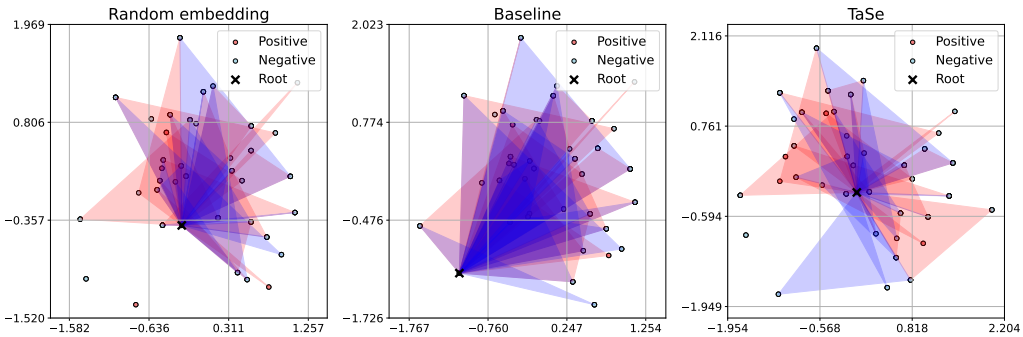


Figure 20: Visualization of angular embeddings (without dynamic reference). Triangles illustrate learned pairs with respect to the root: positive pairs and negative pairs are connected to depict directional behavior. Positive pairs are expected to align in similar directions from the root, while negative pairs should diverge. While the baseline tends to increase radial distance more than meaningful angular adjustment, our objective function encourages more structured representations guided by directional alignment.

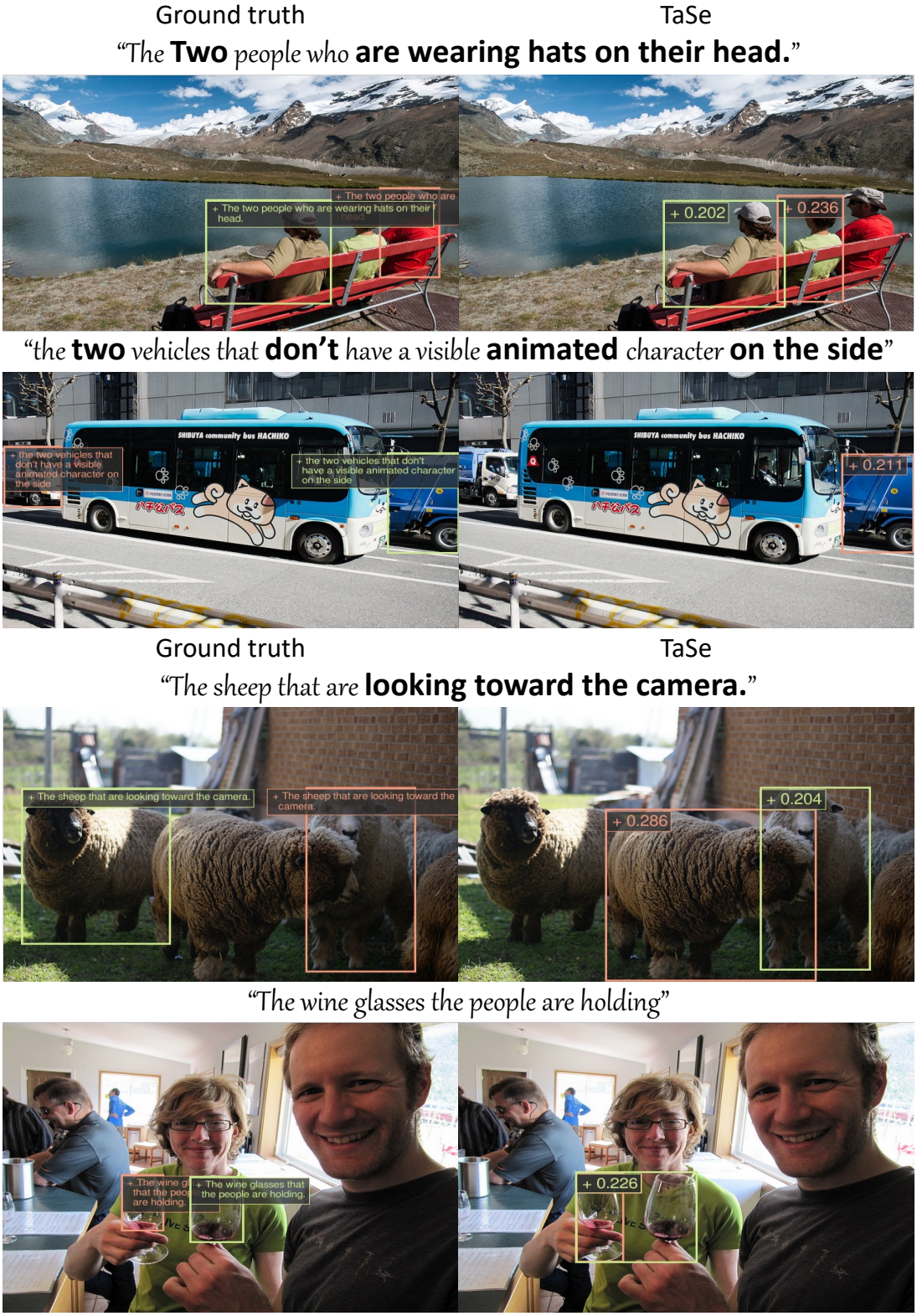


Figure 21: Qualitative results for failure cases.