Object-Centric Alignment and Anchor Distillation for Weakly Supervised Referring Expression Comprehension

Yi Tian[1], **Cheng Yang**[1], **Qingbao Huang**[1*]
[1]School of Electrical Engineering, Guangxi University
{2412391054, 2212391065}@st.gxu.edu.cn, qbhuang@gxu.edu.cn

Abstract

Weakly supervised Referring Expression Comprehension (WREC) aims to localize referred objects using only image-text pairs without box-level annotations. Existing one-stage methods predominantly rely on anchor-level alignment, which suffers from two fundamental limitations: (1) anchors represent local visual patches rather than holistic objects, and (2) they lack the capability to model inter-object relations. To address these issues, we propose **OCAAD**, an Object-Centric Alignment and Anchor Distillation framework. Our key insight is that different self-attention heads in DINOv2 naturally attend to distinct semantic regions, effectively capturing object-level information. Building on this, OCAAD introduces two complementary modules: (1) **Anchor-Object Distillation Module (AODM)**, which bridges the semantic gap between anchors and objects by transferring object-level knowledge from attention heads to anchors via overlap-aware contrastive learning; and (2) **Intra-Modal Relation Consistency (IRC)**, which explicitly models inter-object relations by enforcing the relational structure among linguistic entities to match that of their visual counterparts. Extensive experiments on RefCOCO, RefCOCO+, and RefCOCOg demonstrate that OCAAD achieves new state-of-the-art performance, validating the effectiveness of object-centric alignment for WREC. Code is available at https://github.com/VILAN-Lab/OCAAD.

1 Introduction

Localizing objects through natural language without box annotations remains a fundamental yet challenging problem. Weakly supervised Referring Expression Comprehension (WREC) tackles this challenge by learning to ground referred objects using only image-text pairs as supervision [Deng *et al.*, 2021]. Recent one-stage methods [Jin *et al.*, 2023; Chen *et al.*, 2025] formulate this task as anchor-text contrastive learning, where anchor features from a pre-trained detector are directly matched with text features. Despite their efficiency, these methods share two fundamental limitations that hinder their performance on complex expressions.

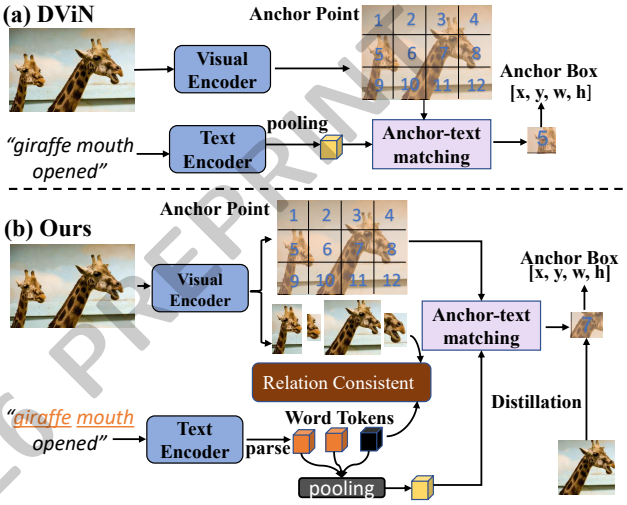

Figure 1: Comparison between DViN and our OCAAD. (a) DViN matches anchor features with the text CLS token, lacking object-level understanding. (b) OCAAD distills object-level semantics from DINOv2 attention heads to anchors, while enforcing relation consistency between linguistic entities and their visual counterparts.

First, existing methods treat anchors as isolated visual patches rather than complete semantic entities. As illustrated in Figure 1(a), anchor features capture only local appearances, lacking holistic object-level information. This makes it difficult to distinguish objects with similar local patterns but different overall semantics, such as two persons wearing similar clothes.

Second, without explicit relation modeling, current methods fail to reason about structural dependencies among multiple objects. For expressions like "the player holding a bat", understanding the relation between "player" and "bat" is crucial for accurate grounding. However, anchor-level alignment provides no mechanism for such relational reasoning.

To overcome these limitations, we propose **OCAAD**, an Object-Centric Alignment and Anchor Distillation framework. Our approach is motivated by a key observation: different self-attention heads in DINOv2 [Oquab *et al.*, 2023] naturally attend to distinct semantic regions within an im-

*Contact Author.

age [Barsellotti *et al.*, 2025]. These attention maps exhibit remarkable object-level consistency without any supervision, making them ideal for bridging the gap between local anchors and holistic objects.

Building on this observation, OCAAD introduces two complementary modules that address the aforementioned limitations respectively:

- **Anchor-Object Distillation Module (AODM)** bridges the semantic gap between anchors and objects. Given a text-guided selected anchor, AODM identifies attention heads with maximum spatial overlap and performs contrastive distillation to transfer object-level semantics into the anchor representation. This enables anchors to capture holistic object information beyond local patches.
- **Intra-Modal Relation Consistency (IRC)** explicitly models inter-object relations. Each linguistic entity is paired with its most similar attention head, and the pairwise relation matrices in both modalities are constrained to be consistent. This enables the model to learn structural dependencies crucial for complex expressions.

We evaluate OCAAD on three standard benchmarks: RefCOCO, RefCOCO+, and RefCOCOg. Extensive experiments demonstrate that our method outperforms previous state-of-the-art, with improvements of up to 3.29% on RefCOCO testA. Notably, OCAAD shows particularly strong gains on RefCOCO+, which relies heavily on relational descriptions, confirming the effectiveness of our relation modeling.

Our contributions are summarized as follows:

- We identify two fundamental bottlenecks in WREC: the lack of object-level semantics in anchors and the absence of relational reasoning, and propose object-centric alignment as a unified solution.
- We design two complementary modules, AODM and IRC, to inject object-level semantics and relational awareness into anchor representations respectively, enabling precise and structured grounding.
- OCAAD achieves state-of-the-art on three REC benchmarks, demonstrating the effectiveness of object-centric alignment for weakly supervised visual grounding.

2 Related Work

2.1 Referring Expression Comprehension

Referring Expression Comprehension localizes image regions described by natural language [He *et al.*, 2023]. Two-stage approaches [Hu *et al.*, 2016; Liu *et al.*, 2019a; Yu *et al.*, 2018; Rohrbach *et al.*, 2016; Zhao *et al.*, 2024] generate region proposals (e.g., via Faster-RCNN [Ren *et al.*, 2015]) and rank them by text similarity, while one-stage methods [Liao *et al.*, 2020; Luo *et al.*, 2020; Zhu *et al.*, 2022; Deng *et al.*, 2021] perform end-to-end prediction for higher efficiency. Recent works leverage content-structure modeling [Zheng *et al.*, 2025], auto-focus mechanisms on locked pre-trained vision models [Zheng *et al.*, 2026], Transformer architectures [Vaswani *et al.*, 2017; Kim *et al.*, 2021], and large language models [Sui *et al.*, 2023] for stronger cross-modal reasoning. However, both paradigms require dense box annotations, motivating weakly supervised alternatives.

2.2 Weakly Supervised REC

WREC learns object grounding using only image-text pairs. Early two-stage methods [Liu *et al.*, 2019b; Liu *et al.*, 2019c; Wang *et al.*, 2021] employ sentence reconstruction [Liu *et al.*, 2021] or contrastive learning [Gupta *et al.*, 2020; Zhang *et al.*, 2020] as weak supervision, but suffer from high computational costs due to dense region proposals.

One-stage alternatives have recently gained attention for their efficiency. RefCLIP [Jin *et al.*, 2023] introduces anchor-text contrastive learning; APL [Luo *et al.*, 2024] improves anchor representations via prompt learning; DViN [Chen *et al.*, 2025] fuses multiple pre-trained encoders through dynamic visual routing. More recently, AlignCAT [Wang *et al.*, 2025] and MC-Bench [Xu *et al.*, 2025] extend to settings with extra category labels or MLLM-based multi-image grounding, both beyond our scope. Existing methods treat anchors as isolated patches, lacking object-level semantics and relational reasoning. Our OCAAD addresses this by introducing object-centric alignment and explicit relation modeling.

3 Preliminary

We briefly review DViN [Chen *et al.*, 2025], a recent one-stage WREC method that serves as our baseline.

Problem Formulation. Given an image $I \in \mathbb{R}^{H\times W\times 3}$ and a referring expression $T = \{w_1, \ldots, w_L\}$, the goal is to localize the referred object by predicting a bounding box $b^* \in \mathbb{R}^4$ using only image-text pairs as supervision.

Visual Feature Extraction. DViN extracts anchor features $\{F_v^i\}_{i=1}^3$ from DarkNet [Redmon and Farhadi, 2018] at three scales (13×13, 26×26, 52×52 for 416×416 input). Each spatial location generates anchor features $f_a \in \mathbb{R}^d$ that can be decoded into bounding boxes via the detection head $\mathcal{D}(\cdot)$.

Language Feature Extraction. An LSTM encoder processes the expression into word-level features $\{f_t^l\}_{l=1}^L$ and a global sentence feature $\bar{f}_t \in \mathbb{R}^d$ from the last hidden state.

Dynamic Visual Routing. To enhance fine-grained visual capability, DViN incorporates multiple visual encoders (DINOv2, EfficientSAM) through a vision-conditioned router. The routing weights are computed as:

$$r = \mathrm{softmax}\Big(\sum_i F_{v0}^i W_r + b_r\Big), \tag{1}$$

where F_{v0} denotes DarkNet features. Only the top-1 expert is selected along with DarkNet to maintain efficiency:

$$S_e = \{0\} \cup \{\arg\max_j r_j\}. \tag{2}$$

The routed anchor feature is a weighted combination:

$$\hat{f}_a = \sum_{j\in S_e} r_j f_a^j. \tag{3}$$

Anchor-Text Matching. Best-matching anchor satisfies:

$$a^* = \arg\max_{f_a\in\mathcal{F}_v} \phi(\hat{f}_a, \bar{f}_t), \tag{4}$$

where $\phi(\cdot,\cdot)$ denotes cosine similarity. The final bounding box is obtained as $b^* = \mathcal{D}(a^*)$.

Training Objective. DViN adopts anchor-text contrastive learning with additional alignment losses:

$$\mathcal{L} = \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_{dt} \mathcal{L}_{dt} + \lambda_{dv} \mathcal{L}_{dv}, \tag{5}$$

where $\mathcal{L}_c$ is the contrastive loss, $\mathcal{L}_r$ is the diversity-aware routing loss, and $\mathcal{L}_{dt}$, $\mathcal{L}_{dv}$ are text-guided and vision-guided alignment losses respectively.

Limitations. Despite the progress, DViN suffers from two limitations. **First**, anchor features represent local patches rather than complete semantic entities, lacking holistic object-level information. This makes it difficult to distinguish objects with similar local appearances but different overall semantics. **Second**, without explicit relation modeling, DViN cannot reason about structural dependencies among objects. For expressions like "the giraffe with mouth closed", understanding the relation between "giraffe" and "mouth" is crucial for accurate grounding, yet anchor-level alignment provides no mechanism for relational reasoning.

4 Method

Given an image $I \in \mathbb{R}^{H \times W \times 3}$ and a referring expression $T = \{w_1, \ldots, w_L\}$, our goal is to localize the referred object using only image-text pairs as supervision. As shown in Figure 2, OCAAD builds upon a dual-branch architecture and introduces two key modules: Anchor-Object Distillation Module (AODM) and Intra-Modal Relation Consistency (IRC). In the following, we first describe the feature extraction process, then detail our object-centric visual representation, and finally present the two proposed modules.

4.1 Feature Extraction

Following DViN [Chen *et al.*, 2025], we adopt the same feature extraction pipeline for visual and language modalities to ensure fair comparison. The key difference lies in our use of DINOv2 attention maps for object-centric representation.

Visual Features. We extract multi-scale visual features $\{F_v^i\}_{i=1}^3$ from DarkNet [Redmon and Farhadi, 2018] for anchor-based detection. Following DViN [Chen *et al.*, 2025], we also incorporate features from EfficientSAM [Xiong *et al.*, 2024] and DINOv2 [Oquab *et al.*, 2023] through a vision-conditioned router. The routed features are fused with DarkNet features, yielding enhanced multi-scale representations $\{\hat{F}^i\}_{i=1}^3$. Additionally, we extract self-attention maps $\{A_k\}_{k=1}^{N_h}$ from the last transformer layer of DINOv2, where $A_k \in \mathbb{R}^{N_p}$ captures the attention weights from the CLS token to spatial patches for head k, and N_p denotes the number of spatial patches.

Language Features. An LSTM encoder encodes the expression into word-level features $\{f_t^l\}_{l=1}^L$ and a global sentence feature $\bar{f}_t \in \mathbb{R}^d$. To enable object-level reasoning, we employ an entity extractor to identify object noun indices $\mathcal{O} = \{o_1, \ldots, o_M\}$ from the expression, yielding entity-level features $\{f_t^{o_m}\}_{m=1}^M$.

4.2 Object-Centric Visual Representation

A key observation motivating our approach is that different self-attention heads in DINOv2 naturally capture diverse semantic regions within an image [Barsellotti *et al.*, 2025]. Research on bridging self-supervised vision backbones with language has demonstrated that these attention maps computed between the CLS token and spatial tokens align with relevant objects, exhibiting remarkable semantic and local consistency. We exploit this property to construct object-centric visual representations that bridge the gap between local anchors and holistic objects.

Head-Weighted Visual Embeddings. For each attention head k, we compute a weighted embedding by aggregating the routed features according to the attention distribution:

$$v^{A_k} = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathrm{softmax}(A_k)_i \cdot \hat{F}^i, \tag{6}$$

where $v^{A_k} \in \mathbb{R}^d$ encapsulates the semantic content captured by head k. This yields N_h distinct visual embeddings $\{v^{A_k}\}_{k=1}^{N_h}$, each attending to different image regions.

Entity-Head Hard Selection. To associate each linguistic entity with its corresponding visual region, we compute the similarity between entity features and all head embeddings, then perform hard selection:

$$k_m^* = \arg\max_{k \in [1, N_h]} \mathrm{sim}(f_t^{o_m}, v^{A_k}), \tag{7}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity. The selected embedding $v^{A_{k_m^*}}$ serves as the object-centric visual representation for entity o_m.

4.3 Intra-Modal Relation Consistency

Complex referring expressions often describe target objects through their relations with other entities, such as "the player wearing a helmet and holding a bat". To capture such structural dependencies, we introduce IRC to enforce consistency between relational structures in visual and linguistic modalities, as illustrated in the left part of Figure 2.

Cross-Modal Entity Matching. Each linguistic entity o_m is first paired with its most similar visual head embedding:

$$\hat{v}_m = v^{A_{k_m^*}}, \tag{8}$$

Pairwise Relation Modeling. For samples with $M \geq 2$ entities, we construct pairwise relation matrices that capture relationships among entities within each modality:

$$R_{\mathrm{lang}}[i, j] = \frac{\exp(f_t^{o_i} \cdot f_t^{o_j} / \tau_{\mathrm{rel}})}{\sum_k \exp(f_t^{o_i} \cdot f_t^{o_k} / \tau_{\mathrm{rel}})}, \tag{9}$$

$$R_{\mathrm{vis}}[i, j] = \frac{\exp(\hat{v}_i \cdot \hat{v}_j / \tau_{\mathrm{rel}})}{\sum_k \exp(\hat{v}_i \cdot \hat{v}_k / \tau_{\mathrm{rel}})}, \tag{10}$$

where τ_{rel} is the relation temperature parameter and softmax normalization is applied row-wise.

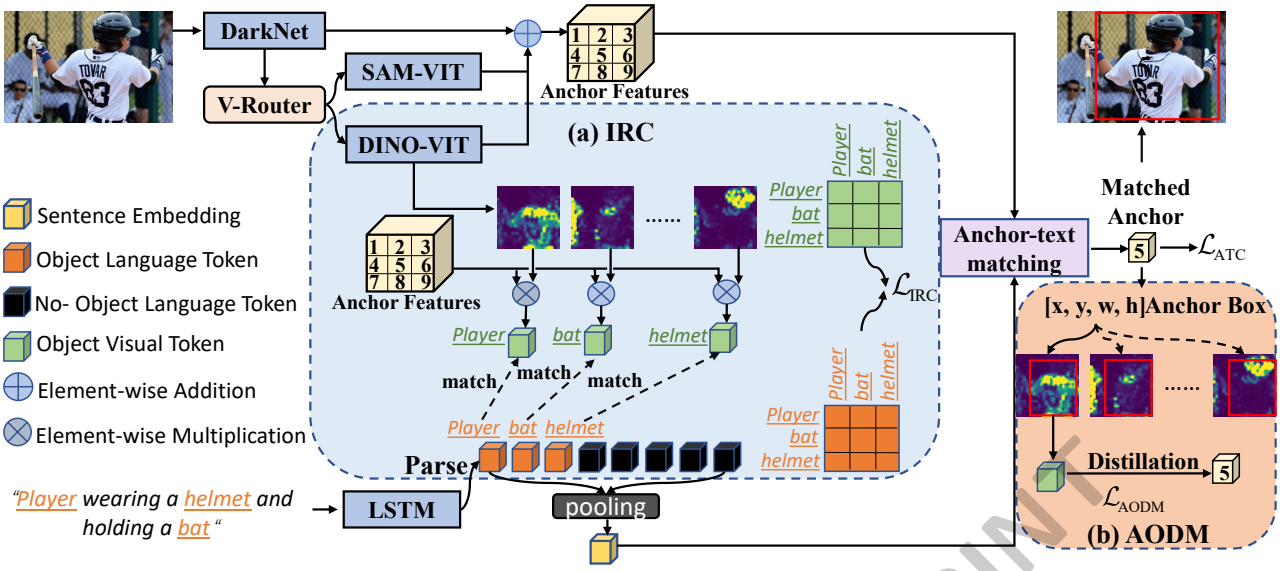

Figure 2: Overall framework of OCAAD. Given an image and referring expression, we extract visual features from DarkNet, EfficientSAM and DINOv2, and language features from LSTM with entity extraction. (a) IRC module: DINOv2 attention heads are matched with linguistic entities extracted from the expression. Pairwise relation matrices are computed in both language and visual spaces, then aligned via MSE loss to enforce relation consistency. (b) AODM module: The matched anchor's bounding box is used to compute spatial overlap with attention maps, selecting the top heads for contrastive distillation.

Relation Alignment Loss. The IRC loss enforces that the linguistic relation structure guides visual relation learning:

$$\mathcal{L}_{\text{IRC}} = \frac{1}{|\mathcal{V}|} \sum_{b \in \mathcal{V}} \|R^{(b)}_{\text{lang}} - R^{(b)}_{\text{vis}}\|^2_F, \tag{11}$$

where $\mathcal{V}$ denotes the set of valid samples containing at least two entities, and $\|\cdot\|_F$ denotes the Frobenius norm. The key insight is that if two entities are semantically close in language space, their visual counterparts should also exhibit similar proximity.

4.4 Anchor-Object Distillation Module

Although anchor features enable efficient one-stage detection, they represent local visual patches rather than complete semantic entities. This creates a fundamental gap between anchor representations and holistic object understanding. To address this issue, AODM transfers object-level knowledge from attention heads to anchors through spatially overlap-aware contrastive learning.

The detailed process of AODM consists of four steps, as illustrated in the right part of Figure 2.

Step 1: Text-Guided Anchor Selection. Given anchor features $\{f^j_a\}_{j=1}^{N_a}$ and the global text feature $\bar{f}_t$, we first identify the anchor most relevant to the text query:

$$j^* = \arg\max_{j \in [1, N_a]} \text{sim}(f^j_a, \bar{f}_t). \tag{12}$$

Step 2: Spatial Overlap Computation. For the selected anchor j^*, we retrieve its associated anchor box b_{j^*} with the highest objectness score. Then, we compute the spatial overlap between this anchor box and each attention map to identify which heads best capture the target region. The attention maps are reshaped to spatial grids $\hat{A}_k \in \mathbb{R}^{H_a \times W_a}$, and the overlap score is computed as:

$$\text{overlap}_k = \frac{\sum_{(i,j) \in \mathcal{B}} \hat{A}_k[i,j]}{\sum_{(i,j)} \hat{A}_k[i,j]}, \tag{13}$$

where $\mathcal{B}$ denotes the set of spatial positions within the anchor box region.

Step 3: Top-2 Head Selection. Based on the overlap scores, we select the top-2 attention heads:

$$\{k^{\text{top}}_1, k^{\text{top}}_2\} = \text{Top2}\big(\{\text{overlap}_k\}_{k=1}^{N_h}\big). \tag{14}$$

The head k^{top}_1 with maximum overlap serves as the positive target for distillation, while k^{top}_2 serves as a hard negative since it typically captures nearby or partially overlapping regions that are easily confused with the target.

Step 4: Overlap-Aware Contrastive Distillation. Finally, we train the anchor feature to align with the best-matching head while contrasting against negative samples. Let $s^+ = \text{sim}(f^{j^*}_a, v^{A_{k^{\text{top}}_1}})$ denote the positive similarity. The distillation loss is formulated as:

$$\mathcal{L}_{\text{AODM}} = -\log \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \mathcal{N}}, \tag{15}$$

where the negative term $\mathcal{N} = \sum_{n \neq b} \exp(s^{(1)}_n/\tau) + \sum_{n=1}^{B} \exp(s^{(2)}_n/\tau)$. Here, $s^{(1)}_n$ and $s^{(2)}_n$ represent similarities to the top-1 and top-2 heads of other samples in the batch, B is the batch size, and τ is the temperature parameter. Importantly, the head embeddings v^{A_k} are detached during training, making DINOv2 heads serve as a frozen teacher that distills

object-level knowledge into learnable anchor representations. This formulation encourages anchors to learn holistic object semantics while maintaining cross-sample discrimination.

4.5 Training and Inference

Training Objective. The overall training loss combines the anchor-text contrastive loss with our proposed modules:

$$\mathcal{L} = \mathcal{L}_{\text{ATC}} + \lambda_1 \mathcal{L}_{\text{AODM}} + \lambda_2 \mathcal{L}_{\text{IRC}}, \tag{16}$$

where $\mathcal{L}_{\text{ATC}}$ is the standard anchor-text contrastive loss [Jin *et al.*, 2023], and λ_1, λ_2 are balancing weights.

Inference. During inference, we select the anchor with the highest text similarity and regress the final bounding box through the detection head:

$$\mathbf{O}_b = \mathcal{D}\Big(\arg\max_{f_a \in F_a} \langle f_a, \bar{f}_t \rangle \Big). \tag{17}$$

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate OCAAD on three standard benchmarks: RefCOCO [Yu *et al.*, 2016], RefCOCO+ [Yu *et al.*, 2016], and RefCOCOg [Mao *et al.*, 2016]. RefCOCO and RefCOCO+ contain over 140K expressions for 50K objects, with RefCOCO+ excluding absolute spatial references. RefCOCOg features longer expressions (average 8.4 words) covering attributes and spatial relations.

Implementation Details. We build upon WeakMCN [Cheng *et al.*, 2025], which shares its architecture with DViN [Chen *et al.*, 2025]. For fair comparison, we adopt the single-task WREC setting without auxiliary segmentation. The detection backbone is YOLOv3-DarkNet [Redmon and Farhadi, 2018], with EfficientSAM [Xiong *et al.*, 2024] and DINOv2 ViT-B/14 [Oquab *et al.*, 2023] as visual experts in the router. We extract attention maps from the 12 heads ($N_h = 12$, $d = 768$) of DINOv2's last transformer layer. An LSTM encodes language features. All visual encoders are frozen; only the LSTM is trainable. Images are resized to 416×416, with maximum text length 15 for RefCOCO/RefCOCO+ and 20 for RefCOCOg. Entity nouns are extracted via Stanza [Qi *et al.*, 2020]. We train for 25 epochs using Adam (learning rate 1×10^{-4}, batch size 64) with cosine decay. Loss weights are $\lambda_1 = 0.1$ and $\lambda_2 = 10$; temperatures are $\tau = 0.07$ and $\tau_{\text{rel}} = 0.1$. All experiments use a single NVIDIA RTX 4090 GPU.

Evaluation Metric. We use Precision@0.5 (IoU ≥ 0.5).

5.2 Comparison with State-of-the-Art

As shown in Table 1, OCAAD achieves state-of-the-art on most splits, with improvements of 1.59% on RefCOCO val, 3.29% on testA, and 1.06% on testB over the baseline. On RefCOCO+, gains reach 3.12% on val and 2.66% on testA.

Notably, gains are more pronounced on testB splits, which feature non-person objects requiring fine-grained attribute understanding, validating that OCAAD captures object-level semantics better than anchor-level methods. On RefCOCO+, which relies on attributes and relative relations rather than absolute spatial references, OCAAD's strong gains confirm IRC's effectiveness in modeling inter-object relations.

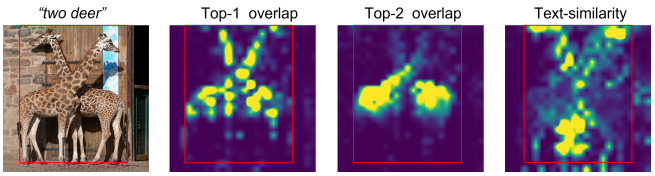

Figure 3: Visualization of different head selection strategies. The top-1 overlapping head captures the complete target object, while the top-2 head attends to partial or incomplete regions, serving as an effective hard negative. The text-similarity head may attend to semantically related but spatially misaligned regions.

5.3 Ablation Study

Component Analysis

Table 2 presents the ablation results of AODM and IRC.

Effect of AODM. AODM achieves notable gains on person-centric splits, improving RefCOCO testA by 2.10% and RefCOCO+ val by 1.92%. This validates that distilling object-level semantics from DINOv2 attention heads enriches anchors for better holistic object understanding.

Effect of IRC. IRC shows strong improvements on RefCOCO+, with gains of 2.70% on val and 1.86% on testB. Since RefCOCO+ excludes absolute spatial references, expressions must rely on relative relations, demonstrating IRC's effectiveness in modeling inter-object relations.

Complementary Nature of AODM and IRC. The full model outperforms individual components with synergistic effects. On RefCOCO testA, AODM alone provides +2.10% and IRC alone +1.85%, while their combination yields +3.29%. This complementarity stems from their distinct roles: AODM enhances object identity recognition while IRC captures structural dependencies among entities.

AODM Design Analysis

To validate AODM design choices, we ablate the head selection strategy in Table 3 and Figure 3.

Importance of Hard Negatives. Using only the top-1 overlapping head (without the second-best as hard negative) decreases performance by 0.46% on val and 1.17% on testA. As shown in Figure 3, the top-2 overlapping head typically attends to nearby or similar regions, serving as an effective hard negative that helps the model learn more discriminative anchor representations.

Spatial Overlap vs. Text Similarity. Replacing spatial overlap with text-similarity for head selection leads to performance drops of 0.94% on val and 1.45% on testA. Figure 3 illustrates that text-similarity may select heads that are semantically related but spatially misaligned with the target object, while spatial overlap provides more accurate correspondence between anchors and attention heads.

Robustness to Noisy Anchors. Although early-stage anchors can be noisy, AODM remains robust: *frozen* DINOv2 heads provide stable supervision, and the overlap score (Eq. 13) varies smoothly with box shifts. Empirically (Table 4), anchor-head overlap and accuracy co-evolve in a self-reinforcing loop. Better anchors yield better head selection, which refines anchors in return.

Method	Venue	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val-g
GT Proposals:								
VC [Zhang *et al.*, 2018]	CVPR'18	-	33.29	30.13	-	34.60	31.58	30.26
ARN [Liu *et al.*, 2019b]	ICCV'19	38.05	36.43	36.47	34.53	36.40	36.12	39.62
KPRN [Liu *et al.*, 2019c]	MM'19	36.34	35.28	37.72	37.16	36.06	39.29	38.37
DTWREG [Sun *et al.*, 2021]	TPAMI'21	39.21	41.14	37.72	39.18	40.01	38.08	43.24
Det Proposals:								
ARN [Liu *et al.*, 2019b]	ICCV'19	32.17	35.25	30.28	32.78	34.35	32.13	33.09
DTWREG [Sun *et al.*, 2021]	TPAMI'21	38.35	39.51	37.01	38.91	39.91	37.09	42.54
RefCLIP [Jin *et al.*, 2023]	CVPR'23	60.36	58.58	57.13	40.39	40.45	38.86	47.87
APL [Luo *et al.*, 2024]	ECCV'24	64.51	61.91	63.57	42.70	42.84	39.80	50.22
DViN [Chen *et al.*, 2025]	CVPR'25	67.67	70.90	59.39	52.39	58.51	43.87	55.04
WeakMCN [Cheng *et al.*, 2025]	CVPR'25	69.20	69.88	62.63	51.90	57.33	43.10	54.62
OCAAD (Ours)	-	**70.79**	**73.17**	**63.69**	**55.02**	**59.99**	**45.02**	**56.85**

Table 1: Comparison with state-of-the-art methods on RefCOCO, RefCOCO+, and RefCOCOg. Metric is Precision@0.5 (%). Best results are shown in **bold**.

AODM	IRC	RefCOCO			RefCOCO+		
		val	testA	testB	val	testA	testB
		69.20	69.88	62.63	51.90	57.33	43.10
✓		69.65	71.98	63.02	53.82	59.43	44.14
	✓	69.96	71.73	63.00	54.60	59.87	44.96
✓	✓	**70.79**	**73.17**	**63.69**	**55.02**	**59.99**	**45.02**

Table 2: Ablation study on the effectiveness of AODM and IRC.

Head Selection Strategy	val	testA	testB
Text-similarity	52.88	57.98	43.40
Top-1 overlapping head only	53.36	58.26	43.59
Top-1 & Top-2 heads (Ours)	**53.82**	**59.43**	**44.14**

Table 3: Ablation on AODM head selection strategies (RefCOCO+). All variants use Baseline+AODM configuration.

Epoch	1	5	10	15	20	25
Avg Overlap	0.31	0.39	0.46	0.51	0.55	0.58
Val Acc (%)	34.61	47.69	51.78	53.11	54.53	55.02

Table 4: Co-evolution of anchor-head overlap and validation accuracy during training on RefCOCO+ val.

Token Selection	val	testA	testB
w/o Head features	53.66	58.59	43.24
All word tokens	54.23	59.01	44.09
All nouns	54.47	59.68	44.22
Parsed entities (Ours)	**54.60**	**59.87**	**44.96**

Table 5: Ablation on IRC token selection strategies (RefCOCO+). All variants use Baseline+IRC configuration.

IRC Design Analysis

Table 5 compares token selection strategies in IRC.

Necessity of Disentangled Head Features. Replacing disentangled head features with raw anchor features ("w/o Head features") causes significant drops across all splits (0.94% on val, 1.28% on testA, 1.72% on testB). This confirms that DINOv2 attention heads provide superior object-centric representations for relation modeling.

Entity Extraction Improves Precision. Using all word tokens introduces noise from non-entity words (e.g., prepositions, articles), resulting in 0.37% lower performance on val. Filtering to all nouns improves results but still includes irrelevant nouns. Our parsed entity extraction achieves the best performance by precisely identifying object-relevant tokens, enabling more accurate entity-visual correspondence.

5.4 IRC Effectiveness on Different Entity Counts

Grouping RefCOCO+ val samples by entity count (Table 6), IRC yields progressively larger gains as the count increases: +1.62% for one entity, +2.00% for two, and +3.19% for three or more, confirming that IRC's effectiveness scales with relational complexity.

5.5 Cross-Dataset Transfer

To assess generalization, we transfer OCAAD across datasets without fine-tuning. As shown in Table 7, OCAAD consistently outperforms WeakMCN, with the largest gains under diverse linguistic styles (e.g., RCg→RC: +5.2% on testA), confirming that object-centric and relational features transfer better than anchor-level ones. Cross-dataset evaluation thus warrants wider adoption in WREC, since in-domain accuracy can mask poor generalization.

5.6 Visualization of DINOv2 Attention Heads

Figure 5 visualizes the attention maps from the last transformer layer of DINOv2 on a sample image. Different heads consistently focus on distinct object-level concepts (e.g., A_4: full body, A_6: head, A_{11}: ball, A_{12}: racket) without supervision, empirically supporting the object-centric assumption underlying both AODM and IRC.

Figure 4: Qualitative comparison between DViN and our OCAAD. Green boxes: ground truth; red boxes: incorrect predictions.

Entity Count	Baseline	+IRC	Δ
1	57.37	58.99	+1.62
2	49.20	51.20	+2.00
≥3	42.03	45.22	+3.19
Total	**51.90**	**54.60**	**+2.70**

Table 6: IRC effectiveness across entity counts (RefCOCO+ val).

Train→Test	WeakMCN	OCAAD	Δ
RC→RC+ (v/tA/tB)	52.57/57.72/43.65	**52.63/58.66/43.91**	+0.1/+0.9/+0.3
RC+→RC (v/tA/tB)	49.38/55.59/40.41	**52.32/58.30/41.96**	+2.9/+2.7/+1.6
RC→RCg (val-g)	55.66	**56.90**	+1.2
RCg→RC (v/tA/tB)	53.64/56.73/49.32	**57.20/61.92/50.15**	+3.6/+5.2/+0.8
RC+→RCg (val-g)	52.21	**54.83**	+2.6
RCg→RC+ (v/tA/tB)	47.84/52.62/41.52	**50.56/55.22/43.32**	+2.7/+2.6/+1.8

Table 7: Cross-dataset transfer results. RC=RefCOCO, RC+=RefCOCO+, RCg=RefCOCOg.

5.7 Qualitative Results

Figure 4 compares DViN and OCAAD on six examples. **Action understanding (Exp-1):** for "woman playing tennis", the baseline encompasses multiple people, while OCAAD identifies the active player. **Spatial relation reasoning (Exp-2, 3):** for "white pants behind batter" (Exp-2), the baseline focuses on the batter; OCAAD localizes the person behind. IRC enables correct grounding for "striped couch nearest tan chair" (Exp-3). **Multi-entity relational grounding (Exp-4):** for "man in blue shirt by guy in red shirt", the baseline fails to associate entities; IRC correctly grounds the target by modeling inter-object relations. **Attribute-based distinction (Exp-5):** for "white shirt oww my toe", AODM captures holistic clothing semantics that the baseline misses. **Object-person association (Exp-6):** for "old man with cooler", OCAAD identifies the person-object association. These cases confirm that AODM enhances object-level semantics while IRC enables structured relational reasoning.

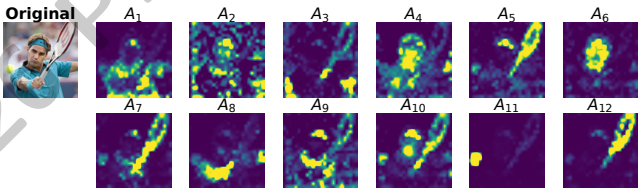

Figure 5: Visualization of DINOv2 attention heads.

6 Conclusion

In this paper, we proposed OCAAD, an Object-Centric Alignment and Anchor Distillation framework for Weakly supervised Referring Expression Comprehension (WREC). Our key contribution is identifying that anchor-level alignment fundamentally limits fine-grained reasoning, and proposing object-centric alignment as a solution. Specifically, AODM transfers object-level semantics from DINOv2 attention heads to anchors through overlap-aware contrastive distillation, while IRC enforces relation consistency between linguistic and visual entity pairs. Experiments demonstrate that OCAAD achieves state-of-the-art performance on three benchmarks, validating the effectiveness of object-centric alignment for weakly supervised visual grounding.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62276072), the Guangxi Natural Science Foundation Key Project (2025GXNSFDA069017), and the Bagui Scholar Program of Guangxi (2025).

References

[Barsellotti *et al.*, 2025] Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22035, 2025.

[Chen *et al.*, 2025] Xiaofu Chen, Yaxin Luo, Gen Luo, Jiayi Ji, Henghui Ding, and Yiyi Zhou. Dvin: Dynamic visual routing network for weakly supervised referring expression comprehension. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14347–14357, 2025.

[Cheng *et al.*, 2025] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9175–9185, 2025.

[Deng *et al.*, 2021] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1769–1779, 2021.

[Gupta *et al.*, 2020] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020.

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

[Hu *et al.*, 2016] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European conference on computer vision*, pages 108–124. Springer, 2016.

[Jin *et al.*, 2023] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2681–2690, 2023.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[Liao *et al.*, 2020] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10880–10889, 2020.

[Liu *et al.*, 2019a] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4673–4682, 2019.

[Liu *et al.*, 2019b] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620, 2019.

[Liu *et al.*, 2019c] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 539–547, 2019.

[Liu *et al.*, 2021] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5612–5621, 2021.

[Luo *et al.*, 2020] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1274–1282, 2020.

[Luo *et al.*, 2024] 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 *European Conference on Computer Vision*, pages 198–215. Springer, 2024.

[Mao *et al.*, 2016] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Qi *et al.*, 2020] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *ACL: System Demonstrations*, 2020.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[Rohrbach *et al.*, 2016] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[Sui *et al.*, 2023] Xiuchao Sui, Shaohua Li, Hong Yang, Hongyuan Zhu, and Yan Wu. Language models can do zero-shot visual referring expression comprehension. In *ICLR Tiny Papers*, 2023.

[Sun *et al.*, 2021] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4189–4195, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2021] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14090–14100, 2021.

[Wang *et al.*, 2025] Yidan Wang, Chenyi Zhuang, Wutao Liu, Pan Gao, and Nicu Sebe. Aligncat: Visual-linguistic alignment of category and attribute for weakly supervised visual grounding. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, 2025.

[Xiong *et al.*, 2024] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iber, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *CVPR*, 2024.

[Xu *et al.*, 2025] Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[Yu *et al.*, 2016] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[Yu *et al.*, 2018] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

[Zhang *et al.*, 2018] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018.

[Zhang *et al.*, 2020] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134, 2020.

[Zhao *et al.*, 2024] P. Zhao, S. Zheng, W. Zhao, D. Xu, P. Li, Y. Cai, and Q. Huang. Rethinking two-stage referring expression comprehension: A novel grounding and segmentation method modulated by point. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7487–7495, 2024.

[Zheng *et al.*, 2025] S. Zheng, P. Zhao, Z. Zheng, P. He, H. Cheng, Y. Cai, and Q. Huang. Look around before locating: Considering content and structure information for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1656–1664, 2025.

[Zheng *et al.*, 2026] S. Zheng, P. Zhao, Q. Huang, Y. Cai, H. Cheng, and Q. Wu. Implement referring expression comprehension by extending auto-focus lens to locked vision model. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(2):1–24, 2026.

[Zhu *et al.*, 2022] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.