AlignCAT: Visual-Linguistic Alignment of Category and Attribute for Weakly Supervised Visual Grounding

Yidan Wang*
Nanjing University of Aeronautics and Astronautics
Nanjing, China
wangyidan@nuaa.edu.cn

Chenyi Zhuang*
University of Trento
Trento, Italy
chenyi.zhuang@unitn.it

Wutao Liu
Nanjing University of Aeronautics and Astronautics
Nanjing, China
wutaoliu@nuaa.edu.cn

Pan Gao†
Nanjing University of Aeronautics and Astronautics
Nanjing, China
pan.gao@nuaa.edu.cn

Nicu Sebe
University of Trento
Trento, Italy
niculae.sebe@unitn.it

Abstract

Weakly supervised visual grounding (VG) aims to locate objects in images based on text descriptions. Despite significant progress, existing methods lack strong cross-modal reasoning to distinguish subtle semantic differences in text expressions due to category-based and attribute-based ambiguity. To address these challenges, we introduce AlignCAT, a novel query-based semantic matching framework for weakly supervised VG. To enhance visual-linguistic alignment, we propose a coarse-grained alignment module that utilizes category information and global context, effectively mitigating interference from category-inconsistent objects. Subsequently, a fine-grained alignment module leverages descriptive information and captures word-level text features to achieve attribute consistency. By exploiting linguistic cues to their fullest extent, our proposed AlignCAT progressively filters out misaligned visual queries and enhances contrastive learning efficiency. Extensive experiments on three VG benchmarks, namely RefCOCO, RefCOCO+, and RefCOCOg, verify the superiority of AlignCAT against existing weakly supervised methods on two VG tasks. Our code is available at: https://github.com/I2-Multimedia-Lab/AlignCAT.

CCS Concepts

• **Computing methodologies** → **Image segmentation; Scene understanding**.

Keywords

Weakly Supervised Visual Grounding, Multimodality

*Equal contribution.
†Corresponding author

MM '25, Dublin, Ireland., October 27–31, 2025

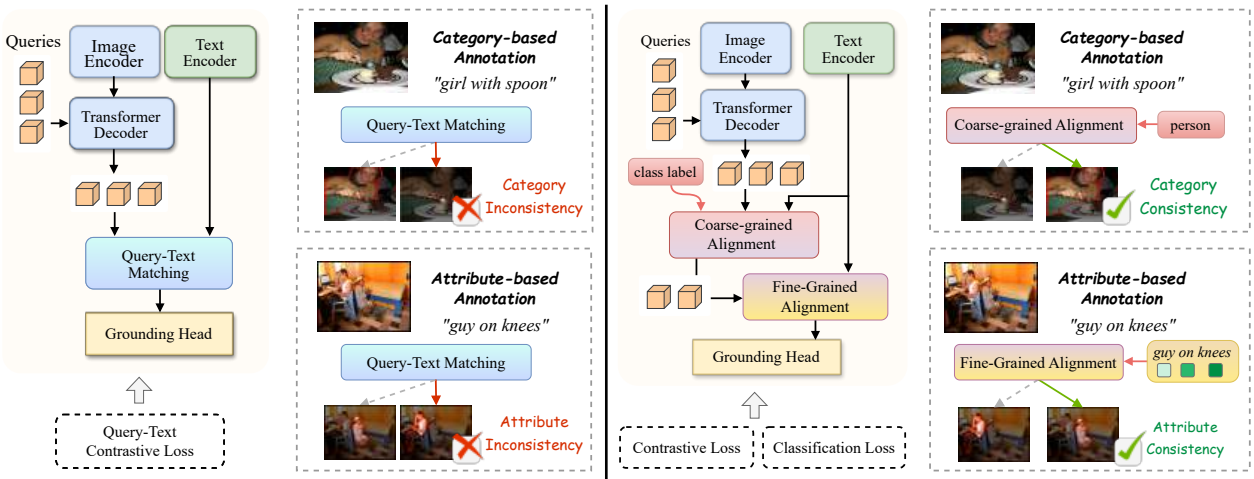
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755751

ACM Reference Format:
Yidan Wang, Chenyi Zhuang, Wutao Liu, Pan Gao, and Nicu Sebe. 2025. 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), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755751

1 Introduction

Visual Grounding (VG) aims to identify objects in an image corresponding to a given text description and has gained attention for its potential in open-ended detection for various computer vision applications [25, 31, 39]. While fully supervised methods [5, 22, 38, 42, 43] achieve high accuracy, they rely on instance-level annotations, which are both labor-intensive and time-consuming to obtain. To alleviate this burden, recent studies have explored weakly supervised learning in two grounding tasks, namely Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES). These works have reframed weakly supervised VG as a region-based [7, 33, 41], anchor-based [9, 23], or query-based [3] matching problem. However, understanding complex textual descriptions and associating referents in multi-object images remains challenging.

In Figure 1, we identify two types of text annotation in existing VG benchmarks: **(1) category-based annotations** that distinguish the referred object in its fundamental class from objects in other categories. For example, in the sentence *"girl with spoon"*, two objects *"girl"* and *"spoon"* belong to different classes. The model should identify the logical object *"girl"* rather than the contextual object *"spoon"*, and align this linguistic category information with visual features. **(2) attribute-based annotations** that describe specific characteristics of the referred object, such as colors and spatial relations. For example, the sentence *"guy on knees"* has no conflict in the person category, but its descriptive information *"on knees"* poses a challenge for the model in identifying the referred object as the image contains multiple persons. This requires an understanding of nuanced textual and visual semantics. However, the state-of-the-art query-based method [3] fails to produce reliable grounding results on both annotation types. While this method utilizes contrastive learning to amplify the alignment of target texts and positive queries, it is not conducive to discriminating nuanced



(a) Framework and visualization of QueryMatch

(b) Framework and visualization of our AlignCAT

Figure 1: Comparison of QueryMatch and the proposed AlignCAT. (a) QueryMatch fails to deal with category-based and attribute-based ambiguity in annotations. (b) AlignCAT progressively leverages linguistic cues from coarse (right-top) to fine (right-bottom) to filter visual queries, achieving category and attribute consistency.

semantics in object categories and attributes. For the category-based annotation, the contextual object *"spoon"* is used to enrich the information of the target object *"girl"*, yet it has created activation noise and hindered the accurate visual-linguistic alignment, showing category inconsistency. For the attribute-based annotation, it mismatches the visual features of the incorrect guy to the target action *"on knee"*, showing attribute inconsistency.

To address the aforementioned visual-linguistic inconsistencies, we propose **AlignCAT** (**Align C**ategory then **AT**tribute), a novel query-based VG framework. To ensure category consistency, we first design a **coarse-grained alignment** module that leverages category information and the global context from the input textual expression. This coarse alignment mitigates interference from irrelevant objects, effectively narrowing down the search space for ideal visual queries. To achieve attribute consistency, we further propose a **fine-grained alignment** module that employs adaptive phrase attention to capture word-level descriptive linguistic features. Such a finer alignment enhances cross-modal correspondences and resolves intra-class ambiguities when multiple visual objects belong to the same category. By aligning visual queries first at a coarse level and then at a finer level, AlignCAT highlights the key role of linguistic cues in understanding cross-modal representations. This category-then-attribute progressive alignment within a contrastive learning framework significantly enhances VG performance. In summary, the main contributions of this work are three-fold:

- We identify category inconsistency and attribute inconsistency in existing weakly supervised VG methods. To address these challenges, we propose a novel query-based category-then-attribute matching framework, modeling linguistic representations from general to detailed levels.
- To achieve visual-linguistic alignment, we design a coarse-grained module that leverages category information and global context to filter out category-inconsistent visual queries, and a fine-grained module that employs adaptive phrase attention to ensure attribute consistency.
- Evaluated on three benchmarks of REC and RES, our proposed method achieves state-of-the-art performance, demonstrating the potential of linguistic cues and the efficacy of the category-then-attribute matching strategy in enhancing visual-linguistic alignment.

2 Related Works

Referring Expression Segmentation (RES). This task aims to overcome the efficiency limitation of fully supervised learning schemes. Weakly supervised RES does not require intensive pixel-level annotation, which is less expensive and more efficient for training. Several works [10, 32] achieve region-text matching through multi-instance learning, but are far inferior to fully supervised methods. Instead of aggregating visual entities, TRIS [16] extracts rough object locations as pseudo-labels based on the input text to perform object localization. Lee et al. [11] relies on the linguistic relationship, which predicts significant maps for each word. However, the masks generated by these methods are highly noisy, resulting in less accurate segmentation.

Referring Expression Comprehension (REC). Compared to fully supervised REC, weakly supervised REC is more challenging due to the lack of bounding box annotations. To obtain additional supervision signals, existing REC methods [17, 36] incorporate external knowledge and align the region-based information with the corresponding phrases. Some works [2, 19] further utilize prior knowledge to filter out irrelevant region proposals. Recent advances also include leveraging language models to build negative samples [40], or pre-trained models to generate pseudo-labels [8, 21]. Yet, these two-stage methods lack generalization to real-world scenarios and large-scale tasks. To improve efficiency, anchor-based methods [9, 23] remove the region proposal stage towards a one-stage process. QueryMatch [3] further introduces a query-text matching scheme to improve the learning of object representations. Despite their advances in efficiency, we identify the challenges of category and attribute inconsistency in existing

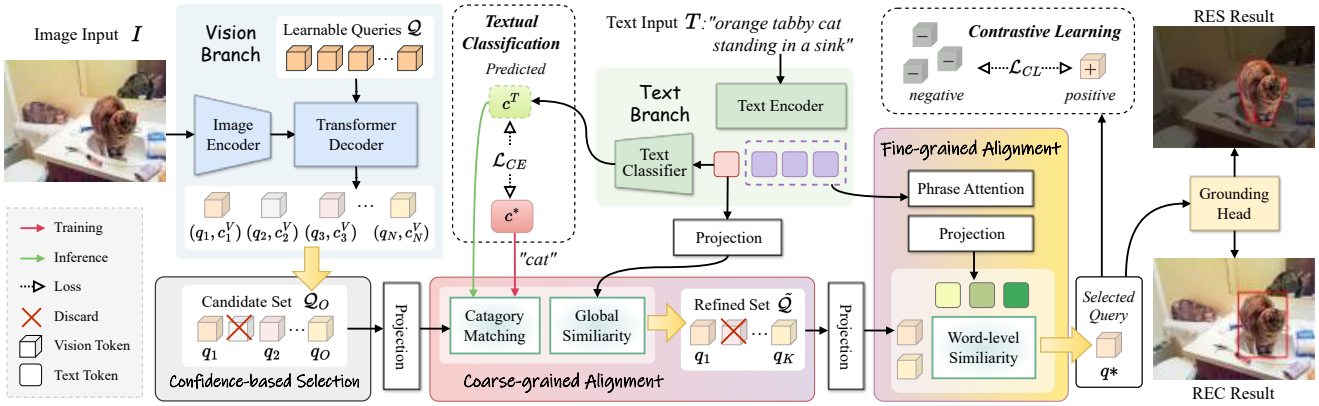


Figure 2: AlignCAT framework overview. AlignCAT filters visual queries by hierarchically leveraging linguistic cues. The coarse-grained alignment module utilizes category and global information to discard category-inconsistent candidates. The fine-grained alignment module employs adaptive phrase attention to select the attribute-consistent visual query.

one-stage methods. To address this problem, our method leverages linguistic cues from general to specific, integrating category information and global context for coarse-grained alignment, and then exploiting word-level descriptions for fine-grained alignment. The category-then-attribute matching framework significantly improves VG results for both RES and REC tasks, especially in multi-object scenarios with complex text expressions.

3 Preliminary

Following [3], we reformulate VG as a query-text matching problem by adopting a query-based detector Mask2Former [4]. It establishes one-to-one associations with objects in the image by N learnable vectors, namely *queries*, denoted as $\mathcal{Q} = \{q_1, ..., q_N\}$. QueryMatch filters out noisy and low-quality query features based on their confidence scores, resulting in a candidate set $\mathcal{Q}_O$, where O is a pre-defined hyperparameter. This method defines two metrics as *difficulty* and *uniqueness* to quantitatively estimate the quality of negative samples. Specifically, difficulty measures vision-language alignment, while uniqueness requires that high-quality negative queries significantly differ from other queries in the embedding space. Given a set of candidate queries, QueryMatch iteratively estimates the quality of i-th query as:

$$S_{d_i} = Norm(\text{sim}(f_{q_i}, f_t)), \tag{1}$$

$$S_{u_i} = Norm(-\max_{j=1}^{M} \cos(f_{q_i}, f_{q_j})), \tag{2}$$

where f_{q_i} is the feature of the current negative query, f_{q_j} is the feature of a previously selected negative query, f_t is the text feature. S_{d_i} is the difficulty of the i-th query, measured by the dot product similarity between visual and linguistic features, denoted $\text{sim}(f_{q_i}, f_t)$. S_{u_i} is the uniqueness of the i-th query, measured by the cosine similarity between two visual queries, denoted $\cos(f_{q_i}, f_{q_j})$. $Norm(\cdot)$ is the min-max normalization. The overall quality score of the negative query is defined as:

$$S_{q_i} = S_{d_i} \cdot S_{u_i}. \tag{3}$$

Ranked in descending order, an appropriate number of negative samples is selected to perform contrastive learning.

The effectiveness of QueryMatch relies on precise query-text matching. This method introduces an effective negative query selection scheme, while simply select the positive query by computing the similarity to the global text feature f_t. However, textual descriptions are expressive and require strong reasoning abilities to understand them. As presented in Figure 1, QueryMatch fails at identifying visual queries from the candidate set $\mathcal{Q}_O$ to achieve visual-linguistic consistency at the category and attribute levels. In this study, we leverage linguistic cues to their fullest extent, emphasizing both coarse and fine-grained information. Based on the category-then-attribute alignment strategy, our proposed AlignCAT can select high-quality positive visual queries and enhance query-text matching accuracy.

4 Methodology

4.1 Overview of AlignCAT

Given the input image I and the input expression T, we aim to locate the referred object through a bounding box (for REC) or a mask (for RES). To address the challenges of category and attribute inconsistencies, we introduce AlignCAT, a novel query-based weakly supervised VG framework. As illustrated in Figure 2, the main goal of AlignCAT is to select high-quality positive queries through a category-then-attribute matching mechanism for efficient contrastive learning. We follow [3] and adopt a query-based detector [4] to process the input image I. The encoded image features are then fed into the Transformer decoder to interact with N learnable queries, outputting the query features $\{f_{q_1}, ..., f_{q_N}\} \in \mathbb{R}^{d_v}$ for visual queries in $\mathcal{Q}$, where d_v is the visual dimension, and one-to-one classifications $\{c_1^V, ..., c_N^V\}$ predicted by the visual classifier.

Unlike existing query-based VG frameworks [3], we leverage linguistic cues in the input expression T to achieve visual-linguistic alignment. The text encoder transforms T into the global feature $f_t \in \mathbb{R}^{d_t}$ and the word-level features $\mathcal{F}_w \in \mathbb{R}^{l \times d_t}$, where l is the length of the input text and d_t is the dimension of the linguistic feature. Our method progressively filters out visual queries through three sequential selection modules: (1) a confidence-based filtering stage reduces the number of visual queries from N to O, forming a

candidate subset Q_O; (2) a coarse-grained alignment module evaluates category consistency and global query-text similarity to further refine the candidates into a refined query set $\widetilde{Q}$; (3) a fine-grained alignment captures attribute details by recalibrating the word-level features $\mathcal{F}_w$, ultimately selecting the most relevant query q^*. Finally, we can decode the selected visual query to obtain the bounding box or the mask of the referred object through a grounding head:

$$r^* = \text{Head}(q^*). \tag{4}$$

4.2 Coarse-grained Alignment

To address category inconsistency, we design a coarse-grained alignment module to first filter out irrelevant visual queries. We discern that the category information is readily available in the input text. For example, it is apparent from the input text *"orange tabby cat standing in a sink"* that the category of the referred object is *"cat"*. We are motivated to predict the specific category and inject this information to ensure that our selected queries belong to the target category. This category-based query-text matching, along with a global query-text matching, effectively mitigates interference from irrelevant objects in the candidate set Q_O. More specifically, at the category matching stage, we inject a Ground Truth (GT) category c^*, which is the class label annotation obtained from the dataset. For each query $q_i \in Q_O$, the Transformer decoder predicts its corresponding category $c_i^V \in \{1, 2, \ldots, C\}$ through a classification head, where C is a pre-defined number of total categories (e.g., $C = 80$ for MSCOCO [14]). The category score measures whether the predicted query category c_i^V is consistent with the GT category c^*. If they are the same, the category score $S_{\text{class},i}$ is set to 1; otherwise, it is set to 0. The above process can be formulated as:

$$S_{\text{class},i} = \begin{cases} 1, & \text{if } c_i^V = c^*, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The category-based query-text matching effectively filters out visual queries that belong to irrelevant categories. To fully exploit context in textual representation, we project the query feature f_{q_i} and global text feature f_t into a coarse-grained shared semantic space to learn global visual-linguistic alignment:

$$\bar{f}_{q_i} = f_{q_i} \cdot W_q + b_q, \tag{6}$$

$$\bar{f}_t = f_t \cdot W_t + b_t, \tag{7}$$

where W_q, W_t are projection matrices, b_q, b_t are biases to transform image and text features, respectively. After projection, we calculate the global query-text matching score:

$$S_{\text{global},i} = \text{sim}(\bar{f}_{q_i}, \bar{f}_t), \tag{8}$$

where $\text{sim}(\cdot)$ is the dot product similarity to measure the alignment between each visual query and the global linguistic feature.

Overall, we define the coarse-grained alignment score as the weighted sum of the category score and the global score:

$$S_{\text{coarse},i} = \alpha S_{\text{class},i} + S_{\text{global},i}, \tag{9}$$

where α is a hyperparameter to balance the value.

Designed to ensure category consistency, this coarse-grained alignment module filters out category-inconsistent visual queries. In other words, only the queries with $S_{class} = 1$ are selected to construct the refined set $\widetilde{Q}$. We also define a threshold K to curtail the query number based on the coarse-grained score S_{coarse}. More details are in the supplemental material.

4.3 Fine-grained Alignment

The above coarse-grained alignment module utilizes general linguistic cues to ensure category consistency and filter out category-inconsistent candidates. However, it is insufficient to discriminate nuanced semantics and achieve attribute consistency. We further introduce a fine-grained alignment that emphasizes descriptive information in word-level textual features, thereby capturing attribute-aware cross-modal correspondences.

Specifically, we adopt an adaptive phrase attention mechanism [35] to emphasize linguistic semantics within the word-level features $\mathcal{F}_w$. Instead of focusing on global context or category-level information, this module highlights fine-grained descriptive details by assigning higher attention weights to attribute words and lower weights to category words. For instance, the phrase *"standing in a sink"* provides more discriminative linguistic cues than other words when distinguishing between two cats, and is therefore given greater attention. More precisely, the word-level features $\mathcal{F}_w$ are first processed by a Bidirectional GRU (Bi-GRU) to recalibrate the importance of each word, which can be formulated as:

$$\widetilde{\mathcal{F}}_w = [\overrightarrow{\mathcal{F}}_w, \overleftarrow{\mathcal{F}}_w] = E(\mathcal{F}_w, \theta), \tag{10}$$

where E and θ represent the Bi-GRU module and its parameters, respectively. $\widetilde{\mathcal{F}}_w$ denotes the modulated word-level features that concatenate bidirectional outputs from the Bi-GRU network. To achieve a more adaptive aggregation, we dynamically balance the weights of the predicted features as:

$$\widetilde{\mathcal{F}}_w := \widetilde{\mathcal{F}}_w \cdot \text{softmax}(\text{FC}(\widetilde{\mathcal{F}}_w)), \tag{11}$$

where $\text{FC}(\cdot)$ is a fully connected layer to predict the weight assigned to each word.

To learn local visual-linguistic alignment, we aggregate these word-level features and project them into a fine-grained semantically shared space, where the query features are also projected. We formulate the above process as follows:

$$\widetilde{f}_w = f_w \cdot W_t' + b_t', \quad \text{where} \quad f_w = \sum_l \widetilde{\mathcal{F}}_w \tag{12}$$

$$\widetilde{f}_{q_i} = f_{q_i} \cdot W_q' + b_q', \tag{13}$$

where f_{q_i} is the i-th query in the refined query set $\widetilde{Q}$. The projection matrices are W_t' and W_q', and the bias terms are b_t' and b_q'. These parameters are used to transform the image and text features, respectively.

To select the visual query that best matches the text expression at the attribute level, we define the fine-grained alignment score as the dot product similarity between each visual query and the fine-grained adapted text feature, which can be expressed as:

$$S_{\text{fine},i} = \text{sim}(\widetilde{f}_{q_i}, \widetilde{f}_w), \tag{14}$$

Since the adapted word feature $\widetilde{f}_w$ encodes discriminative linguistic semantics, this fine-grained alignment module enables the model to differentiate candidate visual queries based on their local

Table 1: Comparisons with state-of-the-art methods on three RES benchmark datasets. Best in red and second in blue.

Method	Venue	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val-g
AMR [29]	*AAAI'22*	14.12	11.69	17.47	14.13	11.47	18.13	15.83
GroupViT [37]	*CVPR'22*	18.03	18.13	19.33	18.15	17.65	19.53	19.97
CLIP-ES [15]	*CVPR'23*	13.79	15.23	12.87	14.57	16.01	13.53	14.16
GbS [1]	*ICCV'21*	14.59	14.60	14.97	14.49	14.49	15.77	14.21
WWbL [30]	*NeurIPS'22*	18.26	17.37	19.90	19.85	18.70	21.64	21.84
TSEG [32]	*arXiv'20*	30.12	-	-	25.95	-	-	22.62
ALBEF [13]	*NeurIPS'21*	23.11	22.79	23.42	22.44	22.07	22.51	24.18
I-Chunk [12]	*ICCV'23*	31.06	32.30	30.11	31.28	32.11	30.13	32.88
TRIS [16]	*ICCV'23*	31.17	32.43	29.56	30.90	30.42	30.80	36.00
APL [23]	*ECCV'24*	55.92	54.84	55.64	34.92	34.87	35.61	40.13
QueryMatch [3]	*MM'24*	**59.10**	**59.08**	**58.82**	**39.87**	**41.44**	**37.22**	**43.06**
Ours	-	**61.83**	**62.75**	**60.02**	**42.05**	**46.39**	**37.53**	**49.06**

Table 2: Comparisons with state-of-the-art methods on three REC benchmark datasets.

Method	Venue	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val-g
VC [28]	*TPAMI'19*	-	32.68	27.22	-	34.68	28.10	29.65
ARN [18]	*ICCV'19*	32.17	35.25	30.28	32.78	34.35	32.13	33.09
KPRN [20]	*MM'19*	36.34	35.28	37.72	37.16	36.06	39.29	38.37
IGN [41]	*NeurIPS'20*	34.78	37.64	32.59	34.29	36.91	33.56	34.92
DTWREG [33]	*TPAMI'21*	38.35	39.51	37.01	38.91	39.91	37.09	42.54
Cycle-Free [34]	*TMM'21*	39.58	41.46	37.96	39.19	39.63	37.53	-
EARN [17]	*TPAMI'23*	38.08	38.25	38.59	37.54	37.58	37.92	45.33
TGKD [26]	*ICRA'23*	39.70	39.92	39.63	40.20	39.94	40.27	47.99
RefCLIP [9]	*CVPR'23*	60.36	58.58	57.13	40.39	40.45	38.86	47.87
APL [23]	*ECCV'24*	64.51	61.91	63.57	42.70	42.84	39.80	**50.22**
QueryMatch [3]	*MM'24*	**66.02**	**66.00**	**65.48**	**44.76**	**46.72**	**41.50**	48.47
Ours	-	**69.03**	**70.27**	**66.59**	**47.16**	**52.22**	**41.91**	**54.72**

representations, even when they belong to the same category. Finally, we select the query with the highest fine-grained alignment score $S_{\text{fine},i}$ as the optimal query:

$$q^* = \arg\max_i S_{\text{fine},i}. \tag{15}$$

4.4 Training and Inference

We adopt a query-text contrastive learning strategy [3] to achieve weakly supervised learning. A common choice for cross-modal contrastive learning objective is InfoNCE:

$$\mathcal{L}_{cl}(h_t, h_q^+, h_q^-) = -\log \frac{\mathcal{T}(h_t, h_q^+)}{\mathcal{T}(h_t, h_q^+) + \sum_{h_q^-} \mathcal{T}(h_t, h_q^-)}, \tag{16}$$

where $\mathcal{T} = \exp(\text{sim}(q, k^+)/\tau)$ is the dot product similarity. The text feature h_t should match the visual feature of its designated query h_q^+ over a set of negative samples h_q^- from other images.

In this study, we introduce two shared semantic spaces for visual-linguistic alignment. Therefore, the final contrastive learning objective of our AlignCAT is the sum of that from the two spaces:

$$\mathcal{L}_{CL} = \mathcal{L}_{cl}(\tilde{f}_t, \tilde{f}_q^+, \tilde{f}_q^-) + \mathcal{L}_{cl}(\widetilde{f_w}, \widetilde{f_q^+}, \widetilde{f_q^-}). \tag{17}$$

During training, we directly inject the GT category to calculate the category score. However, this information is not available at the inference stage. We are driven to train an auxiliary classifier and predict the category from the text side. Specifically, we add a text classifier to project the global linguistic feature f_t and produce the predicted category, denoted c^T. The standard cross-entropy loss is used to train this text classifier:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{18}$$

where y_i is the one-hot encoding of the GT category c^*, and $\hat{y}_i$ is the predicted probability for i-th observation belonging to one class. We note that this text category c^T differs from the query category c_i^V as they are predicted from the linguistic feature f_t and the visual feature f_{q_i}, respectively.

Overall, the weakly supervised learning objective for AlignCAT can be written as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CL} + \lambda_2 \mathcal{L}_{CE}, \tag{19}$$

where λ_1 and λ_2 are hyperparameters dynamically adjusted to control the strengths, detailed in the supplemental material.

5 Experiments

5.1 Datasets and Metric

We evaluate the proposed method on three benchmarks: RefCOCO [27], RefCOCO+ [27], and RefCOCOg [24]. All of them are based on MSCOCO [14], and each contains (image, expression) as: (19,994, 142,210), (19,992, 141,564), (26,711, 104,560). In these three datasets, each expression is associated with one class label, which is used as

Table 3: Ablation of the formula for query quality estimation.

Formula	val	testA	testB
S_{global}	65.89	65.94	65.47
$S_{\text{global}} + S_{\text{class}}$	$67.55_{\uparrow 1.66}$	$69.63_{\uparrow 3.69}$	$64.66_{\downarrow 0.81}$
$S_{\text{global}} + S_{\text{fine}}$	$67.21_{\uparrow 1.32}$	$67.93_{\uparrow 1.99}$	$66.21_{\uparrow 0.74}$
$S_{\text{fine}} + S_{\text{global}} + S_{\text{class}}$	$67.36_{\uparrow 1.47}$	$68.66_{\uparrow 2.72}$	$66.48_{\uparrow 1.01}$
$\mathbf{S_{global} + S_{class} + S_{fine}}$	$\mathbf{69.03}_{\uparrow 3.14}$	$\mathbf{70.27}_{\uparrow 4.33}$	$\mathbf{66.59}_{\uparrow 1.12}$

Table 4: Ablation study of the injected category information. "Train" and "Infer" refer to the training and inference stages. c^*: GT category. c^T: text classifier's predicted category.

c^* (*Train*)	c^T (*Train*)	c^T (*Infer*)	val	testA	testB
-	-	-	67.21	67.93	66.21
✓	-	-	$68.74_{\uparrow 1.53}$	$69.84_{\uparrow 1.91}$	$66.23_{\uparrow 0.02}$
-	-	✓	$67.61_{\uparrow 0.40}$	$68.18_{\uparrow 0.25}$	$66.40_{\uparrow 0.19}$
-	✓	✓	$64.64_{\downarrow 2.57}$	$64.90_{\downarrow 3.03}$	$63.53_{\downarrow 2.68}$
✓	-	✓	$\mathbf{69.03}_{\uparrow 1.82}$	$\mathbf{70.27}_{\uparrow 2.34}$	$\mathbf{66.59}_{\uparrow 0.38}$

the GT category in the coarse-grained alignment. Regarding the text expression, RefCOCO describes objects with absolute spatial information, while the other two datasets are more challenging. RefCOCO+ focuses more on relative spatial information and appearance (such as color and texture), and RefCOCOg provides longer expressions that are more complex and carry richer semantics. For the REC task, we follow [3, 9] that use IoU@0.5 as the metric. We count a prediction as correct if the IoU between the predicted and GT bounding boxes exceeds 0.5. For the RES task, we adopt mIoU [15, 30] as the metric that calculates the average IoU across all test samples. More details are in the supplementary material.

5.2 Implementation Details

Following [3], we employ the pretrained Mask2Former detector [4] and freeze its parameters when training our AlignCAT. The image resolution is set to 416×416. The text lengths for RefCOCO, RefCOCO+, and RefCOCOg are 15, 15, and 20, respectively. All experiments are conducted on two 24G Nvidia RTX 4090 GPUs. The batch size per GPU is 14. The query feature dimension is 256, and the dimensions for word-level features, text features, and the shared semantic space are all 512. During query selection, we set $O = 20$ for confidence-based filtering, $K = 10$ for the maximum selected queries after coarse-grained alignment. We set $\alpha = 100$ to emphasize the category information for calculating coarse-grained scores. We use the Adam optimizer [6] with a learning rate of $1e-4$ and set training epochs to 25.

5.3 Quantitative Analysis

In this section, we first validate AlignCAT by comparing it with comprehensive weakly supervised VG methods, and ablate key components of our approach.

Comparison to the state-of-the-arts. In Tables 1 and 2, we compare AlignCAT with a set of weakly supervised VG methods. The first observation is that AlignCAT significantly outperforms existing methods on all three benchmarks. Our method improves the average accuracy by +2.53% and +2.80% over QueryMatch on RefCOCO for RES and REC, respectively. The improvement on

Table 5: Ablation of the module to inject GT category.

Confidence-based Selection	Coarse-grained Alignment	val	testA	testB
-	-	67.61	68.18	66.40
✓	-	67.24	**71.50**	62.06
-	✓	**69.03**	70.27	**66.59**

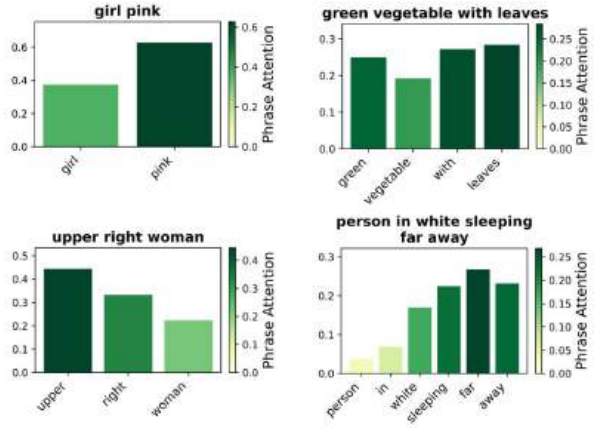


Figure 3: Visualization of adaptive phrase attention.

RefCOCOg is particularly notable, with AlignCAT increasing the accuracy of QueryMatch by more than 6% for both tasks. We also notice that AlignCAT excels on TestA of all datasets, where most categories of referred objects are *"person"*. With the help of category matching, AlignCAT effectively filters out visual queries not belonging to humans, before more fine-grained alignment. This validates the effectiveness of our innovative category-then-attribute mechanism in enhancing cross-modal alignment, with the capacity to tackle multi-object images and complex text expressions.

Ablation of AlignCAT. To validate the designs of AlignCAT, we have conducted various ablation studies on the RefCOCO dataset for weakly supervised REC. We first compare different settings of query selection. When ablating the design of global similarity, the corresponding contrastive learning objective is also removed. The same applies to the fine-grained alignment with word-level similarity calculation. These results are reported in Table 3. The baseline selects one positive visual query with the highest S_{global}. With category matching, the combination $S_{\text{class}} + S_{\text{global}}$ improves VG performance on two subsets, albeit with a slight decrease on testB. This suggests that category information benefits human-target localization, but struggles with non-human objects. Solely using the fine-grained alignment, $S_{\text{global}} + S_{\text{fine}}$ achieves 67.93% on RefCOCO testA, yet is worse than the former setting with 69.63%. This comparison highlights the importance of category-based filtering. We also experimented with the attribute-then-category order. The result of $S_{\text{fine}} + S_{\text{global}} + S_{\text{class}}$ shows a remarkable performance decline compared to $S_{\text{global}} + S_{\text{class}}$. We suspect that without category-based filtering, the text features of contextual objects create noise and affect the cross-modal alignment. Conversely, $S_{\text{class}} + S_{\text{global}} + S_{\text{fine}}$ with a category-then-attribute order achieves the best performance, demonstrating the effectiveness of the coarse-to-fine visual-linguistic matching scheme, as well as the complementary effect of three selection modules.

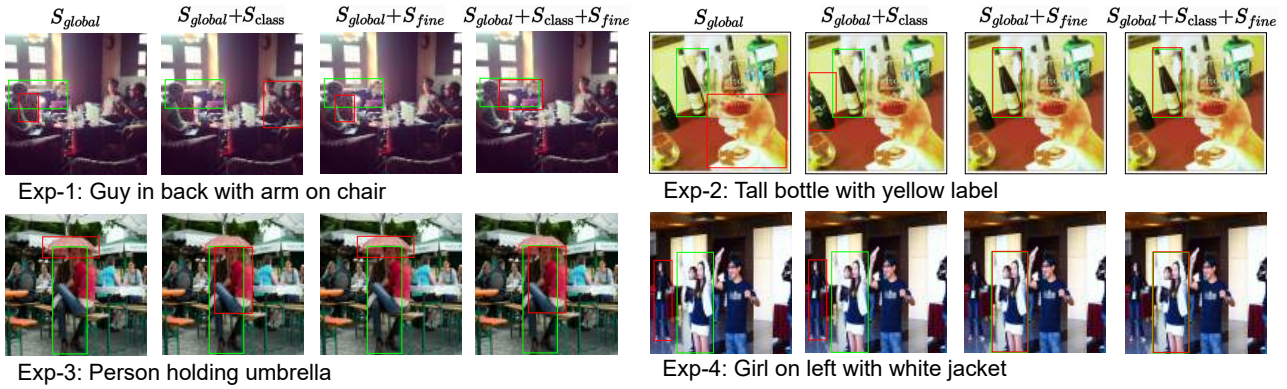


Figure 4: Visualization comparison of different selection designs of AlignCAT in weakly supervised REC. The red and green boxes are GT and predicted grounding results, respectively.

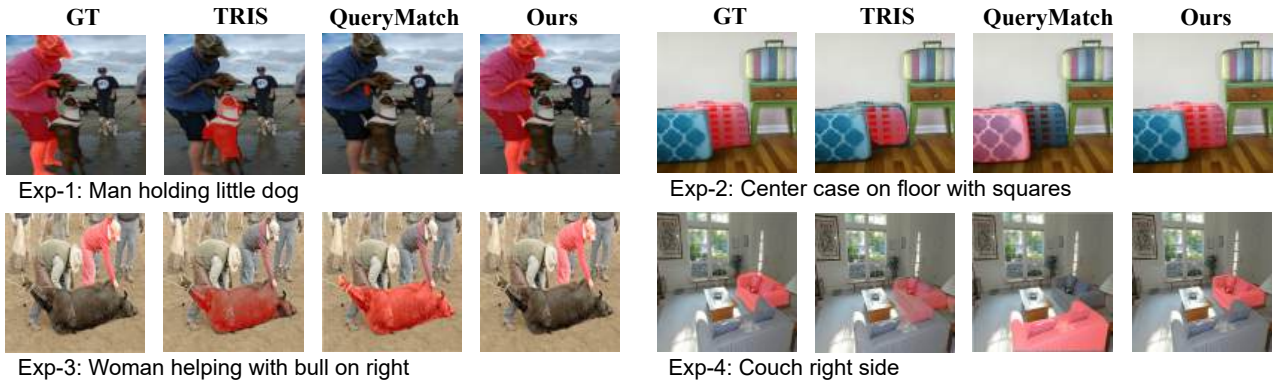


Figure 5: Visualization comparison to TRIS and QueryMatch for the weakly supervised RES task. The GT and predicted segmentation results are marked in red.

Next, we examine different strategies for using category information during training and inference. As shown in Table 4, the first row is the baseline without category information. The second row is the model trained with the GT category c^* while removing the category matching score during inference. This setting improves the model performance, which highlights the importance of category information in enhancing cross-modal correspondences. The third row is the result of training the text classifier but only injecting the predicted category c^T during inference, which presents a slight improvement. Notably, in the last second row, directly injecting the predicted category c^T during training indicates a significant performance decline. We suspect that the text classifier's predicted categories are largely inaccurate at the beginning of training, resulting in unreliable visual-linguistic alignment. The last row is our full model that uses the GT category during training and injects the predicted category during inference. This design enhances robustness and achieves the best performance across all subsets.

We further investigate alternative strategies for injecting GT category information. As shown in Table 5, we compare the confidence-based selection and the global feature alignment to incorporate the category score. In the former, Q_O is filtered by confidence and category matching, while $\widetilde{Q}$ relies on global similarity S_{global}. Although this improves testA performance, it leads to a significant performance degradation on testB. This issue arises due to suboptimal negative query selection, which are sampled from Q_O. Since a large proportion of referred objects in the training set belong to the *"person"* category, integrating category matching during confidence-based filtering results in the same category between most negative queries and the positive query. This reduces the diversity of negative samples and affects generalization. To address this, we inject the category information at the coarse-grained alignment module. This setting enhances negative sample quality and improves the model's robustness across comprehensive scenarios.

5.4 Qualitative Analysis

In Figure 3, we visualize the weights of the modulated word-level text features after adaptive phrase attention. These values illustrate how the model dynamically adjusts the importance of each word. For example, given the text *"girl pink"*, the model highlights the color attribute *"pink"* than the category word *"girl"*. Interestingly, the contextual object *"leaves"* is allocated with a higher value than the referred object *"vegetable"*. This observation explains the result in Table 3 for the inferior performance in the setting of the exchanged order. Overall, AlignCAT leverages descriptive information to mitigate intra-class ambiguity, thereby distinguishing objects belonging to the same category.

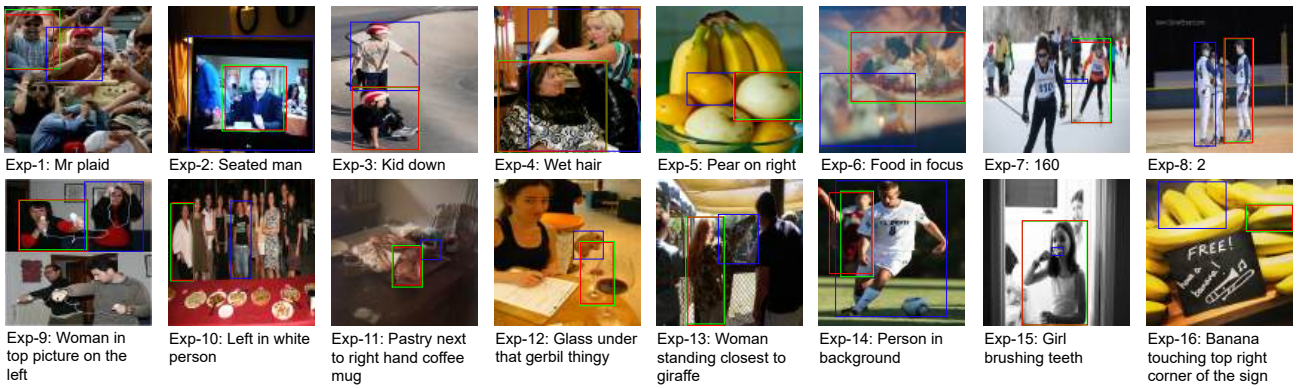


Figure 6: Visualization comparison in weakly supervised REC. Green: ground truth. Blue: QueryMatch. Red: Ours.

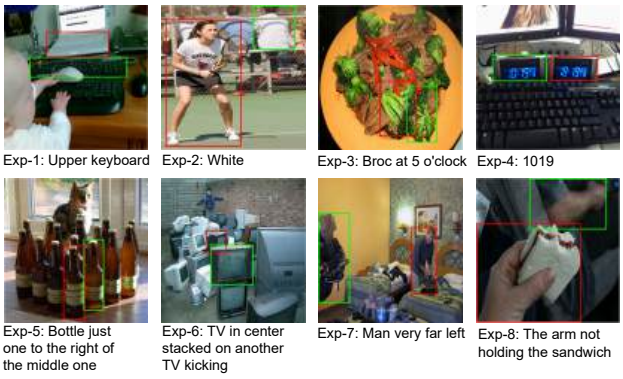


Figure 7: Failure cases of AlignCAT in weakly supervised REC. Green: ground truth. Red: Ours.

To gain in-depth insights into the category-then-attribute alignment mechanism, we ablate AlignCAT with four configurations and visualize the results in Figure 4. Utilizing solely the global feature similarity S_{global} struggles to achieve category and attribute consistencies, especially when images involve multiple objects. With the category matching score, the results present the category consistency. For instance, $S_{\text{global}} + S_{\text{class}}$ excludes the contextual object *"label"* in Exp-2, however, it fails to resolve intra-class ambiguities and selects another bottle. On the other hand, without S_{class}, the results of $S_{\text{global}} + S_{\text{fine}}$ still suffers from category inconsistency in Exp-3, which highlights the importance of the category matching. This discovery is consistent with the quantitative comparison in Table 3. In contrast, the full configuration $S_{\text{global}} + S_{\text{class}} + S_{\text{fine}}$ achieves category and attribute consistencies, whether the target is human or non-human, and the attribute is color (e.g., *"white"*) or spatial relation (e.g., *"in back"*). Notice that AlignCAT may fail to accurately locate the target object due to the occlusion problem (e.g., the arm behind the head in Exp-1).

In Figure 5, we compare AlignCAT with two state-of-the-art weakly supervised RES models, TRIS [16] and QueryMatch [3]. Given texts with complex relationships and intensive images with multiple objects, they incorrectly locate the contextual objects such as *"dog"* and *"bull"*. Conversely, the proposed method achieves higher segmentation accuracy. With stronger reasoning ability and better visual understanding, AlignCAT is more robust and reliable in extensive grounding scenarios.

Figure 6 compares the performance of AlignCAT and QueryMatch in the weakly supervised REC task. The analysis shows that AlignCAT has a clear advantage in maintaining category and attribute consistency. For example, in Exp-7, AlignCAT successfully aligns the abstract and vague query *"160"* with the *"person"* category, accurately localizing the target. In contrast, QueryMatch fails to understand this abstract query, leading to a localization failure. In complex multi-object scenarios, AlignCAT continues to effectively select the correct queries. For instance, in Exp-16, AlignCAT not only identifies the logical subject *"banana"*, but also captures its fine-grained spatial relationships, achieving precise localization. In summary, AlignCAT, with its robust coarse-to-fine semantic alignment, outperforms QueryMatch in complex scenarios.

To provide a more comprehensive analysis, we present typical failure cases of AlignCAT for the REC task in Figure 7. These failures occur due to dataset quality issues, including wrong ground truth annotations (Exp-1) and insufficient textual descriptions (Exp-2). Meanwhile, our method still lacks semantic understanding to comprehend out-of-distribution textual expressions (Exp-3), or to discern nuanced visual features for similar objects (Exp-4).

6 Conclusion

In this study, we identify that existing weakly supervised VG methods suffer from contextual ambiguities, showing category and attribute inconsistencies. To address these challenges, we propose a novel query-based VG framework, AlignCAT, with a category-then-attribute visual-linguistic alignment strategy to progressively filter out query candidates. To ensure category consistency, we design a coarse-grained alignment module that leverages category information and global context. For attribute consistency, we further propose a fine-grained alignment module to capture word-level linguistic features and emphasize attribute-based query-text alignment, effectively resolving intra-class ambiguities. Extensive experiments demonstrate that AlignCAT achieves state-of-the-art performance on three benchmarks for both REC and RES tasks. The proposed category-then-attribute alignment enhances category and attribute consistencies in comprehensive scenes. This work provides novel insights into leveraging linguistic cues for advancing weakly supervised visual grounding.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 62272227). It was also partly supported by the MUR PNRR project FAIR (PE00000013) funded by the NextGenerationEU, the EU Horizon projects ELIAS (No. 101120237) and ELLIOT (No. 101214398).

References

[1] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. 2021. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1801–1812.

[2] Kan Chen, Jiyang Gao, and Ram Nevatia. 2018. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4042–4050.

[3] Shengxin Chen, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, and Rongrong Ji. 2024. QueryMatch: A Query-based Contrastive Learning Framework for Weakly Supervised Visual Grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4177–4186.

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1290–1299.

[5] Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. 2024. SimVG: A Simple Framework for Visual Grounding with Decoupled Multi-modal Fusion. *arXiv preprint arXiv:2409.17531* (2024).

[6] P Kingma Diederik. 2014. Adam: A method for stochastic optimization. *(No Title)* (2014).

[7] Francisco Eiras, Kemal Oksuz, Adel Bibi, Philip HS Torr, and Puneet K Dokania. 2024. Segment, select, correct: A framework for weakly-supervised referring segmentation. In *European Conference on Computer Vision*. Springer, 326–342.

[8] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. 2022. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15513–15523.

[9] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. 2023. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2681–2690.

[10] Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. 2023. Shatter and gather: Learning referring image segmentation with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15547–15557.

[11] Jungbeom Lee, Sungjin Lee, Jinseok Nam, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. 2023. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 21870–21881.

[12] Jungbeom Lee, Sungjin Lee, Jinseok Nam, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. 2023. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 21870–21881.

[13] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[15] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. 2023. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15305–15314.

[16] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. 2023. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22124–22134.

[17] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. 2022. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3003–3018.

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

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

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

[21] Yang Liu, Jiahua Zhang, Qingchao Chen, and Yuxin Peng. 2023. Confidence-aware Pseudo-label Learning for Weakly Supervised Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2828–2838.

[22] 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*. 10034–10043.

[23] Yaxin Luo, Jiayi Ji, Xiaofu Chen, Yuxin Zhang, Tianhe Ren, and Gen Luo. 2025. APL: Anchor-Based Prompt Learning for One-Stage Weakly Supervised Referring Expression Comprehension. In *European Conference on Computer Vision*. Springer, 198–215.

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

[25] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.

[26] Jinpeng Mi, Song Tang, Zhiyuan Ma, Dan Liu, Qingdu Li, and Jianwei Zhang. 2023. Weakly supervised referring expression grounding via target-guided knowledge distillation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 8299–8305.

[27] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 792–807.

[28] Yulei Niu, Hanwang Zhang, Zhiwu Lu, and Shih-Fu Chang. 2019. Variational context: Exploiting visual and textual context for grounding referring expressions. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 347–359.

[29] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. 2022. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2117–2125.

[30] Tal Shaharabany, Yoad Tewel, and Lior Wolf. 2022. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *Advances in Neural Information Processing Systems* 35 (2022), 28222–28237.

[31] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 539–559.

[32] Robin Strudel, Ivan Laptev, and Cordelia Schmid. 2022. Weakly-supervised segmentation of referring expressions. *arXiv preprint arXiv:2205.04725* (2022).

[33] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. 2021. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2021), 4189–4195.

[34] Mingjie Sun, Jimin Xiao, Eng Gee Lim, and Yao Zhao. 2021. Cycle-free weakly referring expression grounding with self-paced learning. *IEEE Transactions on Multimedia* 25 (2021), 1611–1621.

[35] Wei Tang, Liang Li, Xuejing Liu, Lu Jin, Jinhui Tang, and Zechao Li. 2023. Context disentangling and prototype inheriting for robust visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[36] Josiah Wang and Lucia Specia. 2019. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4663–4672.

[37] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18134–18144.

[38] Ziyan Yang, Kushal Kafle, Franck Dernoncourt, and Vicente Ordonez. 2023. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19165–19174.

[39] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6281–6290.

[40] Ruisong Zhang, Chuang Wang, and Cheng-Lin Liu. 2023. Cycle-consistent weakly supervised visual grounding with individual and contextual representations. *IEEE Transactions on Image Processing* (2023).
[41] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems* 33 (2020), 18123–18134.
[42] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. 2021. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems* 34, 1 (2021), 134–143.
[43] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. 2022. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*. Springer, 598–615.