

GrabVG: Graph-Attentive Binding for Visual Grounding in UAV Imagery

Chaowei Wang^{1,*} Yan Di^{2,*} Jingjun Sun¹ Baozhe Liu³
Jiaxu Tian¹ Yuheng Li¹ Guangqian Guo¹ Shan Gao^{1,†}

¹Northwestern Polytechnical University ²Harbin Institute of Technology

³The Hong Kong Polytechnic University

chaowei.wang@mail.nwpu.edu.cn diyan@hit.edu.cn gaoshan@nwpu.edu.cn

*Equal contribution. †Corresponding author.

Abstract

Visual grounding in Unmanned Aerial Vehicle (UAV) imagery aims to localize a target object in complex bird's-eye-view scenes according to a natural language description. However, the abundance of small, densely distributed, and visually similar objects creates high visual redundancy, while repetitive local configurations give rise to strong topological ambiguity. Existing approaches mainly focus on visual-language feature alignment or dense contextual interaction, yet they struggle to distinguish subtle inter-instance differences and effectively exploit spatial topological structures, leading to inaccurate grounding in highly crowded scenarios. To address these challenges, we propose **GrabVG**, a novel visual grounding framework inspired by human visual search. GrabVG explicitly decomposes grounding into two sequential stages: *preattentive hypothesis search* and *graph-attentive feature binding*. Specifically, we first generate a compact set of reliable object hypotheses through distillation-guided proposal induction and text-aware hypothesis filtering, substantially reducing background distractions and semantic mismatches. These hypotheses are then organized into a sparse graph, where language-guided intra-instance visual cues and inter-instance topological relationships are jointly bound and propagated via graph attention, enabling efficient spatial reasoning and accurate target localization. Extensive experiments on AerialVG and AerialSense show that GrabVG achieves a favorable accuracy-speed trade-off, reaching 67.31% and 80.34% Acc@0.5 and outperforming the corresponding baselines by 10.55 and 8.76 percentage points, respectively.

1. Introduction

Visual grounding in Unmanned Aerial Vehicle (UAV) imagery aims to localize a target object in a bird's-eye-view

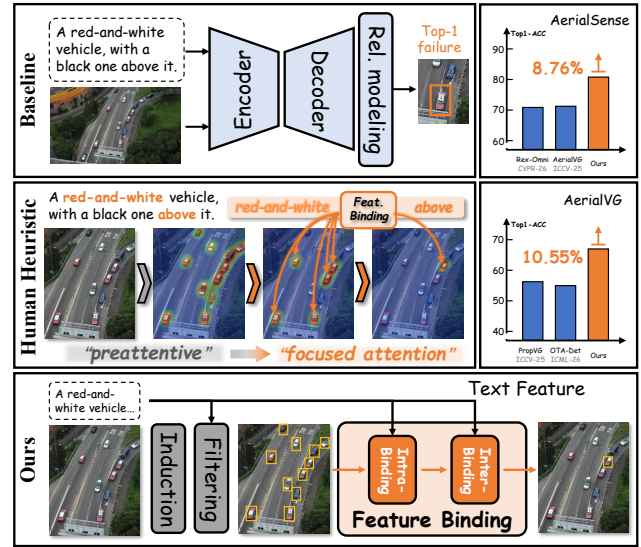


Figure 1. Comparison of visual grounding paradigms. **Baseline** [19]: Existing methods rely on global decoding or dense relational modeling, often leaving visually similar hypotheses insufficiently distinguished. **Human Visual Search** [25, 30]: Preattentive guidance first narrows the search, followed by focused binding of discriminative evidence. **Ours**: GrabVG follows this progressive process through hypothesis induction, hypothesis filtering, and graph-attentive feature binding.

scene according to a natural-language expression [4, 9, 15, 19]. Compared with conventional ground-level imagery, UAV scenes typically contain numerous small and densely distributed objects, many of which exhibit highly similar appearances [19, 31, 42, 46]. The limited visual distinctions among these instances create *high visual redundancy*, making appearance cues alone insufficient for distinguishing the referred object from nearby distractors [19]. Moreover, the elevated viewpoint weakens perspective and depth cues while bringing a larger number of objects into view si-

multaneously. As a result, objects are often compressed into dense and repetitive two-dimensional arrangements, causing different target hypotheses to share similar local neighborhoods and relative spatial configurations. Such structurally similar neighborhoods make topological cues alone insufficient to uniquely identify the referent, giving rise to *strong topological ambiguity*.

Existing visual grounding approaches address these challenges primarily through stronger vision–language alignment and contextual relation modeling. Vision–language alignment methods enhance expression-conditioned visual representations [7, 29], while relation-based approaches incorporate spatial configurations among instances [19]. However, cross-modal alignment is often performed across numerous intermediate queries, which may insufficiently encode subtle candidate-level differences. Meanwhile, dense relation modeling captures extensive pairwise interactions, potentially allowing distant or weakly relevant relations to interfere with informative local context. Consequently, existing approaches may still struggle to obtain discriminative candidate representations and reliable contextual evidence in crowded UAV scenes. These limitations arise from both the modeled evidence and the organization of the grounding process.

Human visual search offers a useful principle for organizing this process. Guided Search suggests that preattentive feature signals bias visual attention toward a restricted set of likely target locations [30], while Feature Integration Theory posits that focused attention is required to bind distributed visual features into coherent object representations [25]. Inspired by this progressive process, we formulate UAV visual grounding as *preattentive hypothesis search* followed by *graph-attentive feature binding*: the former constrains the search space to plausible target hypotheses, whereas the latter associates each surviving hypothesis with discriminative evidence.

As illustrated in Fig. 1, we instantiate this formulation as **GrabVG**, which separates candidate-space construction from evidence-based referent selection. During *Preattentive Hypothesis Search*, Distillation-Guided Proposal Induction transfers expression-aligned object proposals from an external teacher pipeline, while Text-Aware Hypothesis Filtering filters out likely background queries and retains foreground candidates aligned with the referring expression. During *Graph-Attentive Feature Binding*, the remaining hypotheses are organized as a sparse geometric graph. Intra-Instance Appearance Binding supplements each graph node with language-guided visual details sampled from the corresponding region and its immediate surroundings, whereas Inter-Instance Topological Binding exchanges edge-conditioned spatial information among neighboring nodes. This separation allows GrabVG to preserve subtle instance-level appearance cues while limiting

relational reasoning to geometrically meaningful neighborhoods, thereby reducing interference from dense all-pairs interactions.

Extensive experiments on AerialVG and AerialSense demonstrate the effectiveness and efficiency of GrabVG. It achieves 67.31% Acc@0.5 on AerialVG and 80.34% Acc@0.5 on AerialSense, outperforming the corresponding baselines by 10.55 and 8.76 percentage points, respectively, while maintaining competitive inference speed. Our main contributions are summarized as follows:

- We reformulate crowded UAV visual grounding as a search-before-binding problem, separating the construction of plausible object candidates from the comparison of appearance and relational evidence. This formulation provides an explicit alternative to directly decoding the referent from dense visual tokens or unrestricted object interactions.
- We develop two complementary mechanisms for this formulation. Distillation-guided proposal induction and text-aware filtering establish a well-constrained, expression-aligned candidate space, while language-guided adaptive sampling and edge-aware graph attention enhance node appearance and model local geometric relations, respectively.
- Comprehensive experiments on AerialVG and AerialSense demonstrate the effectiveness and efficiency of GrabVG. Ablation studies further examine the contributions of proposal supervision, candidate filtering, adaptive appearance sampling, graph topology, and different pseudo-annotation sources.

2. Related Work

Visual Grounding. Visual grounding localizes an image region according to a natural-language expression. Early proposal-based methods rank detected regions using appearance, location, and contextual cues [38, 39], while one-stage and transformer-based approaches directly regress the referred region from cross-modal representations [7, 13, 18, 36, 45]. Subsequent studies improve visual–language correspondence through language-guided feature refinement [23, 37], dynamic visual sampling [22], language-irrelevant token removal [24], and decoupled multimodal fusion [5]. Grounded pre-training further unifies language-conditioned localization with open-vocabulary detection [17, 20], while PropVG revisits proposal-driven grounding [6] and recent work explores progressive refinement for small-object REC [8]. However, most of these methods either operate on dense multimodal features or rank generic region proposals, without explicitly separating candidate screening from instance-level relational reasoning.

UAV and Remote-Sensing Visual Grounding. Remote-sensing visual grounding is challenged by scale variation, cluttered backgrounds, high-resolution imagery, and small targets. Existing studies improve cross-modal alignment [14, 29, 40], relational reasoning [19, 44], and semantic-geometric modeling [3, 41]. ProVG further adopts a progressive *survey-locate-verify* strategy that sequentially injects global, relational, and attribute cues into dense visual features [16]. In contrast, GrabVG first screens explicit object candidates and then reasons over their local visual attributes and geometric neighborhoods.

Reasoning-Centric Remote-Sensing Grounding. Recent methods such as RSGround-R1 and Geo-R1 improve grounding through supervised reasoning traces and reinforcement fine-tuning [11, 43]. These approaches offer explicit textual rationales and improved low-shot reasoning, but require additional reasoning supervision or rollout-based optimization and incur autoregressive inference overhead. GrabVG instead performs fixed-depth differentiable reasoning over sparse candidate graphs, offering directly inspectable node-edge interactions without generating textual reasoning trajectories.

Relation Modeling for Visual Grounding. Relational context is commonly used to distinguish visually similar candidates. Early methods compare target proposals with surrounding objects or decompose expressions into subject, location, and relation components [38, 39]. Later graph-based methods employ language-guided node and edge attention [27], cross-modal relation graphs [33], dynamic multi-step reasoning [34], or aligned language and visual scene graphs [35]. However, these relations are typically modeled over generic proposals or broad semantic structures. AerialVG explicitly captures positional relations among aerial objects [19], but dense interaction may introduce context from distant or weakly relevant instances. GrabVG instead applies language-conditioned graph attention to filtered candidates connected by local geometric proximity, thereby limiting message passing from distant or weakly relevant instances.

3. Method

3.1. Overview

Given a UAV image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and a referring expression $\mathbf{T} = \{w_l\}_{l=1}^L$, visual grounding aims to localize the unique instance described by $\mathbf{T}$ with a normalized bounding box $\mathbf{b}^* \in [0, 1]^4$. As illustrated in Fig. 2, GrabVG resolves the referent through explicit candidate construction and graph-based comparison rather than direct decoding from an unconstrained set of visual tokens.

The vision and language encoders first produce multi-scale image features and token-level text features:

$$\begin{aligned}\mathbf{F}_{img} &= \{\mathbf{F}_{img}^s\}_{s=1}^S = \mathcal{E}_v(\mathbf{I}), \\ \mathbf{F}_{txt} &= \mathcal{E}_l(\mathbf{T}).\end{aligned}\quad (1)$$

A DETR-style proposal generator $\mathcal{F}_{ind}$ [1, 47] first maps the multi-scale visual features $\mathbf{F}_{img}$ to N initial hypotheses, $\mathcal{H}^0 = \mathcal{F}_{ind}(\mathbf{F}_{img}) = \{(\mathbf{q}_i^0, \mathbf{b}_i^0)\}_{i=1}^N$, where each hypothesis consists of a decoder representation and its normalized bounding box. GrabVG then progressively filters and binds these hypotheses through four consecutive states:

$$\mathcal{H}^0 \xrightarrow{\mathcal{F}_{filt}} \mathcal{H}^f \xrightarrow{\mathcal{F}_{app}} \mathcal{H}^a \xrightarrow{\mathcal{F}_{topo}} \mathcal{H}^g. \quad (2)$$

Here, Text-Aware Hypothesis Filtering $\mathcal{F}_{filt}$ scores the initial hypotheses using $\mathbf{F}_{txt}$ and retains K foreground candidates, forming $\mathcal{H}^f$. Intra-Instance Appearance Binding $\mathcal{F}_{app}$ updates their representations with sampled local image features to obtain $\mathcal{H}^a$. Inter-Instance Topological Binding $\mathcal{F}_{topo}$ subsequently exchanges language-conditioned messages over neighboring candidates, yielding $\mathcal{H}^g$. The bounding boxes remain unchanged throughout the two binding modules.

Let $\mathcal{H}^g = \{(\mathbf{q}_j^g, \mathbf{b}_j^g)\}_{j=1}^K$. We adopt the Contrastive-based Refer Scoring (CRS) module from PropVG [6] for final referent selection. CRS evaluates the compatibility between each graph-aware hypothesis and the referring expression to produce a referring score, while leaving the associated box unchanged:

$$\begin{aligned}s_j^{ref} &= \mathcal{F}_{crs}(\mathbf{q}_j^g, \mathbf{F}_{txt}), \\ \mathbf{b}^* &= \mathbf{b}_{j^*}, \quad j^* = \arg \max_j s_j^{ref}.\end{aligned}\quad (3)$$

3.2. Preattentive Hypothesis Search

Grounding annotations identify only the referred instance and provide little supervision for representing other plausible foreground objects. Consequently, proposal queries trained solely with the final referring objective may receive insufficient supervision to form well-constrained object hypotheses in crowded scenes. Preattentive Hypothesis Search addresses this issue in two steps: proposal induction transfers expression-aligned supervision from offline pseudo annotations, and text-aware filtering removes background queries before the more costly binding operations.

Distillation-Guided Proposal Induction. To supplement this sparse supervision, we follow the principle of teacher-student knowledge transfer [10] and employ an external teacher pipeline $\mathcal{F}_{tea}$ [19, 20] to generate an expression-aligned pseudo-annotation set offline:

$$\begin{aligned}\mathcal{Y}_{tea} &= (\{\bar{\mathbf{b}}_r\}_{r=1}^R, r^*) \\ &= \mathcal{F}_{tea}(\mathbf{I}, \mathbf{T}),\end{aligned}\quad (4)$$

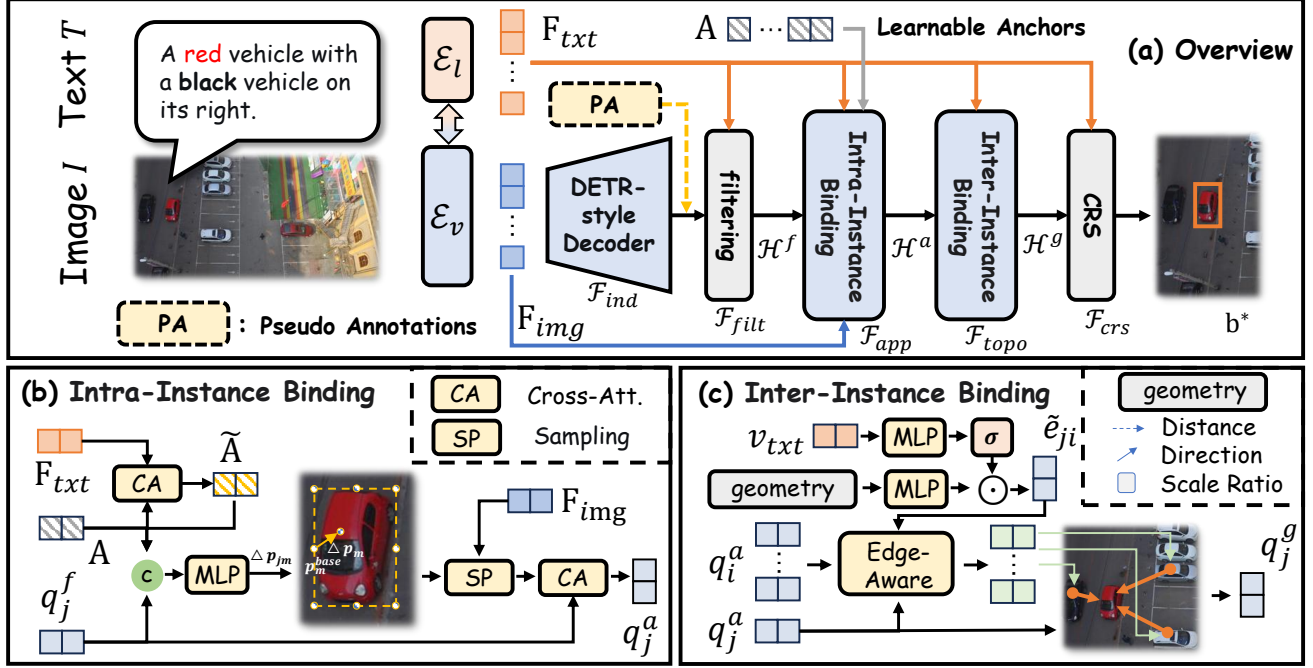


Figure 2. **Overall architecture of GrabVG.** (a) Given a UAV image and a referring expression, Distillation-Guided Proposal Induction first produces an initial hypothesis set $\mathcal{H}^0$. Text-Aware Hypothesis Filtering then removes background queries and retains an expression-aligned candidate subset. The resulting hypotheses $\mathcal{H}^f$ are processed by Intra-Instance Appearance Binding and Inter-Instance Topological Binding, followed by CRS for final target selection. (b) Intra-Instance Appearance Binding samples language-guided visual features from each hypothesis region and its immediate surroundings to obtain appearance-aware representations $\mathbf{Q}^a$. (c) Inter-Instance Topological Binding exchanges relational information over a sparse geometric graph using language-conditioned edge features, producing graph-aware representations $\mathbf{Q}^g$.

where $\bar{\mathbf{b}}_r$ denotes a pseudo foreground box and r^* identifies the pseudo annotation corresponding to the referred instance. These annotations are generated only for the training split and are not used during validation or inference. To guarantee positive referent supervision during training, the ground-truth referent box is appended when no retained teacher prediction satisfies the predefined matching criterion.

Using $\mathbf{F}_{img}$ as the visual memory, the proposal generator $\mathcal{F}_{ind}$ employs N learnable queries and stacked query self-attention and multi-scale deformable cross-attention layers to produce the initial hypothesis set $\mathcal{H}^0$ defined in the Overview. A foreground classification head and a box regression head are attached to each decoder query. Their predictions are assigned to $\{\bar{\mathbf{b}}_r\}_{r=1}^R$ through one-to-one bipartite matching and optimized using the standard DETR classification, ℓ_1 regression, and generalized IoU losses. This supervision encourages the decoder queries to cover multiple expression-aligned foreground hypotheses with accurate boxes, rather than concentrating only on the annotated referent.

Text-Aware Hypothesis Filtering. Although distillation improves proposal coverage, $\mathcal{H}^0$ may still contain background regions and hypotheses that are weakly aligned with the referring expression. We therefore employ a lightweight text-aware filter to perform coarse candidate screening before the more expensive binding operations. Its goal is to retain foreground hypotheses aligned with the referring expression, while fine-grained attribute and relational disambiguation is delegated to the subsequent binding and selection modules.

Let $\mathbf{Q}^0 = [\mathbf{q}_1^0, \dots, \mathbf{q}_N^0]$. A single cross-attention block [26] injects linguistic context into the proposal features for relevance estimation:

$$\begin{aligned} \mathbf{U} &= \text{LN}(\mathbf{Q}^0 + \text{MHA}(\mathbf{Q}^0, \mathbf{F}_{txt}, \mathbf{F}_{txt})), \\ \bar{\mathbf{Q}} &= \text{LN}(\mathbf{U} + \text{FFN}(\mathbf{U})), \quad s^f = \sigma(\mathbf{W}_f \bar{\mathbf{Q}}). \end{aligned} \quad (5)$$

Here, $\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes a multi-head attention block. The top- K scoring hypotheses are retained:

$$\begin{aligned} \Omega &= \text{TopK}(s^f, K) = \{\omega_1, \dots, \omega_K\}, \\ \mathcal{H}^f &= \{(\mathbf{q}_{\omega_j}^0, \mathbf{b}_{\omega_j}^0)\}_{j=1}^K \triangleq \{(\mathbf{q}_j^f, \mathbf{b}_j)\}_{j=1}^K. \end{aligned} \quad (6)$$

The auxiliary features $\bar{\mathbf{Q}}$ are used only to estimate fil-

tering scores, while the selected original decoder features $\{\mathbf{q}_{\omega_j}^0\}_{j=1}^K$ are forwarded to the binding stage. This separates language-guided hypothesis selection from subsequent feature binding, with the filtering scores supervised by $\mathcal{L}_{filt}$.

3.3. Graph-Attentive Feature Binding

Graph-Attentive Feature Binding updates the graph through node-level visual sampling and edge-conditioned message passing. Intra-Instance Appearance Binding enriches each node using features sampled around its associated box, whereas Inter-Instance Topological Binding exchanges spatial information among geometrically neighboring candidates.

Intra-Instance Appearance Binding. Within the hypothesis graph, each node initially contains a decoder feature $\mathbf{q}_j^f$ and its associated box $\mathbf{b}_j = (\mathbf{c}_j, \boldsymbol{\rho}_j)$, where $\mathbf{c}_j$ and $\boldsymbol{\rho}_j$ denote the normalized box center and size, respectively. To supplement the decoder feature with candidate-specific visual evidence, we associate each hypothesis with M box-relative anchors $\{(\mathbf{a}_m, \mathbf{p}_m^{base})\}_{m=1}^M$, where $\mathbf{a}_m$ is a learnable query embedding and $\mathbf{p}_m^{base}$ denotes its predefined base location relative to the hypothesis box. These query embeddings are first conditioned on the referring expression, after which candidate-specific offsets are jointly predicted from the hypothesis representation and the resulting language-aware anchor embeddings:

$$\begin{aligned} \mathbf{A}^t &= \text{MHA}(\mathbf{A}, \mathbf{F}_{txt}, \mathbf{F}_{txt}), \\ \Delta \mathbf{p}_{jm} &= \text{MLP}_{off}([\mathbf{q}_j^f; \mathbf{a}_m; \mathbf{a}_m^t]), \\ \mathbf{p}_{jm} &= \text{clip}_{[0,1]^2}(\mathbf{c}_j + (\mathbf{p}_m^{base} + \Delta \mathbf{p}_{jm}) \odot \boldsymbol{\rho}_j). \end{aligned} \quad (7)$$

Here, $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_M]$ and $\mathbf{A}^t = [\mathbf{a}_1^t, \dots, \mathbf{a}_M^t]$ denote the original and language-aware anchor embeddings, respectively. The predicted locations therefore adapt to both the candidate representation and the linguistic context while remaining relative to the candidate box. We then sample multi-scale visual features at these locations and aggregate them to update the corresponding hypothesis:

$$\begin{aligned} \mathbf{F}_j^{loc} &= \text{Sample}(\mathbf{F}_{img}, \{\mathbf{p}_{jm}\}_{m=1}^M), \\ \mathbf{q}_j^a &= \text{MHA}(\mathbf{q}_j^f, [\mathbf{q}_j^f; \mathbf{F}_j^{loc}], [\mathbf{q}_j^f; \mathbf{F}_j^{loc}]). \end{aligned} \quad (8)$$

Here, Sample denotes differentiable bilinear sampling from all visual feature levels followed by cross-scale feature fusion. The resulting appearance-aware hypothesis set is $\mathcal{H}^a = \{(\mathbf{q}_j^a, \mathbf{b}_j)\}_{j=1}^K$.

Inter-Instance Topological Binding. After appearance refinement, the hypothesis boxes remain unchanged while the node attributes are replaced by $\{\mathbf{q}_j^a\}_{j=1}^K$. To capture

local spatial configurations without introducing distracting all-pairs interactions, we construct a symmetric k -nearest-neighbor graph from the detached normalized hypothesis centers and include a self-loop for every node:

$$\mathcal{N}(j) = \text{SymKNN}_k(\mathbf{c}_j) \cup \{j\}. \quad (9)$$

The self-loop allows each node to directly balance its own appearance evidence against messages from neighboring hypotheses during attention aggregation. The graph connectivity is shared across all graph-attention layers. For each directed edge $j \leftarrow i$, we construct a geometric descriptor $\mathbf{r}_{ji}$ from the relative center displacement, distance, direction, and log-scale difference between the two boxes. A pooled sentence representation $\mathbf{v}_{txt}$ then modulates the projected geometric descriptor:

$$\tilde{\mathbf{e}}_{ji} = \text{MLP}_{geo}(\mathbf{r}_{ji}) \odot \sigma(\text{MLP}_{txt}(\mathbf{v}_{txt})). \quad (10)$$

Starting from $\mathbf{z}_j^{(0)} = \mathbf{q}_j^a$, we perform edge-aware graph attention over the local neighborhood. For clarity, the attention-head index is omitted:

$$\begin{aligned} a_{ji}^\ell &= \frac{1}{\sqrt{d}} \left(\mathbf{W}_q^\ell \mathbf{z}_j^{(\ell-1)} \right)^\top \left(\mathbf{W}_k^\ell \mathbf{z}_i^{(\ell-1)} + \mathbf{W}_{ek}^\ell \tilde{\mathbf{e}}_{ji} \right) \\ &\quad + \left(\mathbf{W}_{eb}^\ell \right)^\top \tilde{\mathbf{e}}_{ji}, \\ \alpha_{ji}^\ell &= \text{Softmax}_{i \in \mathcal{N}(j)}(a_{ji}^\ell), \\ \mathbf{m}_j^\ell &= \sum_{i \in \mathcal{N}(j)} \alpha_{ji}^\ell \left(\mathbf{W}_v^\ell \mathbf{z}_i^{(\ell-1)} + \mathbf{W}_{ev}^\ell \tilde{\mathbf{e}}_{ji} \right), \\ \mathbf{z}_j^{(\ell)} &= \text{GraphBlock}^\ell(\mathbf{z}_j^{(\ell-1)}, \mathbf{m}_j^\ell). \end{aligned} \quad (11)$$

The edge-conditioned key and bias terms determine the attention allocated to neighboring hypotheses, while the edge-conditioned value injects relational information into the aggregated message $\mathbf{m}_j^\ell$. The graph block then combines this message with the previous node state through residual and feed-forward updates. After L_g graph-attention layers, a learned sigmoid gate adaptively blends the graph-propagated state $\mathbf{z}_j^{(L_g)}$ with the appearance-aware representation $\mathbf{q}_j^a$ to produce $\mathbf{q}_j^g$. The resulting graph-aware hypothesis set is $\mathcal{H}^g = \{(\mathbf{q}_j^g, \mathbf{b}_j)\}_{j=1}^K$.

3.4. Training Objective

GrabVG is jointly optimized with proposal-induction, hypothesis-filtering, grounding, and auxiliary segmentation objectives:

$$\begin{aligned} \mathcal{L} &= \lambda_{det} \mathcal{L}_{det} + \lambda_{filt} \mathcal{L}_{filt} \\ &\quad + \lambda_{ref} \mathcal{L}_{ref} + \lambda_{mask} \mathcal{L}_{mask}. \end{aligned} \quad (12)$$

The proposal-induction loss $\mathcal{L}_{det}$ follows DETR [1] and combines foreground classification, ℓ_1 box regression, and

generalized IoU losses [21]. The grounding loss $\mathcal{L}_{ref}$ and auxiliary segmentation loss $\mathcal{L}_{mask}$ follow PropVG [6].

For hypothesis filtering, let $\mathcal{M}$ denote the bipartite matching between induced hypotheses and pseudo annotations. A hypothesis is treated as a positive filtering target if it is matched to any pseudo box. Let i^r denote the hypothesis matched to the referred pseudo annotation and Ω the retained index set. We combine binary relevance supervision with a retention penalty that discourages pruning the referred hypothesis:

$$\begin{aligned} y_i^f &= \mathbb{I}[\exists r, (i, r) \in \mathcal{M}], \\ \mathcal{L}_{filt} &= \lambda_{bce} \text{BCE}(\mathbf{s}^f, \mathbf{y}^f) \\ &\quad + \lambda_{keep} \mathbb{I}[i^r \notin \Omega] \left[-\log(s_{i^r}^f + \epsilon) \right]. \end{aligned} \quad (13)$$

The filtering objective performs coarse foreground screening, while final referent discrimination is supervised by $\mathcal{L}_{ref}$ over the retained hypotheses.

4. Experiments

4.1. Setup

Datasets. We evaluate GrabVG on **AerialVG** [19] and the visual-grounding subset of **AerialSense** [9]. AerialVG contains 5,000 high-resolution images, 50,000 referring expressions, and 103,000 bounding boxes, and we follow its official train-test split. AerialSense covers diverse resolutions, illumination conditions, and UAV scenes; as no official split is available, we use the first 75% of samples for training and the remaining 25% for testing.

Metrics. We report **Acc@0.5 (Top1)**, **Top-5 Acc@0.5 (Top5)**, and **mIoU**. Top1 and Top5 measure whether the top-1 prediction or any of the five highest-scoring boxes has IoU above 0.5, respectively, while mIoU is the mean IoU of the top-1 predictions.

Baselines. Our comparisons cover general visual grounding methods [6], UAV-specific visual grounding methods [19, 29], and zero-shot grounding methods [12].

Implementation Details. Following PropVG [6], we adopt the BEiT-3 Base [28] backbone and a three-layer deformable proposal decoder. Images are resized with the aspect ratio preserved and padded to 586×586 , and the maximum text length is 64. We initialize $N = 256$ hypotheses for AerialVG and $N = 128$ for AerialSense, of which the top-scoring 50% are retained. Appearance binding uses 48 adaptive anchors, while topological binding employs four graph-attention layers over a symmetrized 6-NN graph. The models are trained using Adam for 36 epochs

Method	Top1 (%)	Top5 (%)	mIoU (%)
<i>Zero-Shot Methods</i>			
ReX-Omni [12]	28.37	30.09	-
<i>Supervised Methods</i>			
TransVG [7]	11.53	13.68	-
D-MDETR [22]	19.87	29.87	-
G-DINO [20]	29.36	78.87	-
AerialVG [19]	50.03	87.00	-
PropVG [6]	49.40	-	49.38
PropVG* [6]	56.76	74.05	46.76
OTA-Det [29]	54.90	-	-
GrabVG (Ours)	67.31	89.43	53.34

Table 1. **Comparison with state-of-the-art methods on the AerialVG test set.** “*” indicates results obtained with hyperparameter optimization. Best results are in **bold**. Top1 and Top5 denote Acc@0.5 and Top-5 Acc@0.5, respectively.

on AerialVG and 45 epochs on AerialSense, with learning rates of 1.5×10^{-4} for the newly introduced modules and 1.5×10^{-5} for the backbone. Unless otherwise stated, we use AerialVG-derived pseudo annotations for AerialVG and Grounding-DINO-derived pseudo annotations for AerialSense. The filtering layer uses eight attention heads. Training uses a batch size of 3 per GPU, gradient clipping at 0.15, and two epochs of linear warm-up. The learning rate is subsequently reduced by a factor of 0.1; for AerialVG, the decay is applied at epochs 24 and 32.

4.2. Comparison with State of the Art

Results on AerialVG. As shown in Table 1, we compare the proposed GrabVG with recent methods, including both zero-shot multimodal approaches and fully supervised aerial-specific models. GrabVG achieves state-of-the-art performance on the AerialVG test set, with 67.31% Top1, 89.43% Top5, and 53.34% mIoU. Compared with the strongest one-stage baseline, OTA-Det (54.90%), and the dense-relation model AerialVG (50.03%), GrabVG improves Top1 by 12.41 and 17.28 percentage points, respectively. Furthermore, GrabVG outperforms the hyperparameter-optimized two-stage baseline PropVG* by 10.55 percentage points in Top1. These results suggest that expression-aligned proposal supervision, local visual sampling, and neighborhood-restricted graph reasoning provide complementary benefits in crowded UAV scenes.

Results on AerialSense. To evaluate the generalization capability of GrabVG across diverse UAV scenarios, we report results on AerialSense in Table 2. GrabVG achieves the best performance, reaching 80.34% Top1 and 72.00% mIoU. It surpasses the zero-shot ReX-Omni and the supervised AerialVG baseline by 8.87 and 8.76 percentage points

Method	Type	Top1 (%)	mIoU (%)
<i>Large Multimodal Models (Zero-Shot)</i>			
ReX-Omni [12]	Zero-shot	71.47	-
<i>Domain-Specific Models</i>			
AerialVG [19]	Supervised	71.58	-
PropVG* [6]	Supervised	61.30	67.76
GrabVG (Ours)	Supervised	80.34	72.00

Table 2. **Performance on the AerialSense benchmark.** We compare GrabVG with a recent zero-shot multimodal model and aerial-specific supervised baselines.

in Top1, respectively. These results demonstrate that the proposed candidate construction and graph-attentive binding framework generalizes effectively beyond AerialVG.

4.3. Comprehensive Ablation Study

Component-wise Contribution. Table 3 incrementally evaluates each component from the optimized PropVG baseline, which is trained on the original ground-truth annotations and achieves 56.76% Top1. Introducing **Distillation-Guided Proposal Induction** with expression-aligned pseudo annotations improves Top1 to 62.84%, indicating that denser foreground supervision produces a more reliable and better-covered hypothesis space. Adding **Intra-Instance Appearance Binding** further improves Top1 by 2.75 points to 65.59%, demonstrating the benefit of recovering candidate-specific local cues for distinguishing visually similar objects. In comparison, **Inter-Instance Topological Binding** alone yields a 2.26-point gain to 65.10%, showing that local geometric relations among neighboring hypotheses provide effective contextual evidence. Combining both binding modules reaches 67.01%, confirming that fine-grained appearance cues and inter-instance relational information are complementary. Finally, **Text-Aware Hypothesis Filtering** achieves the best Top1 of 67.31% by suppressing weakly relevant hypotheses before feature binding.

4.4. Further Experimental Analysis

Filtering Strategy. Table 4 examines where and how often hypotheses should be screened. All variants use the same AerialVG-derived pseudo annotations for proposal induction. Replacing the learned filter with non-maximum suppression (NMS) produces a similar Top1 result (67.25% versus 67.31%) and a higher Top5 result, but NMS does not exploit the referring expression. Applying the learned filter both before and after appearance binding reduces Top1 to 66.75%. Repeated pruning can therefore discard ambiguous but correct hypotheses before topological evidence becomes available. A single early text-aware filter provides the best Top1 while keeping the subsequent binding stages focused on a compact candidate set.

Baseline	GrabVG Components				Test		Val	
	$\mathcal{V}_{tea}$	$\mathcal{F}_{app}$	$\mathcal{F}_{topo}$	$\mathcal{F}_{filt}$	Top1	Top5	Top1	Top5
✓					56.76	74.05	55.96	72.28
✓	✓				62.84	85.54	61.82	85.04
✓	✓	✓			65.59	87.89	65.42	87.13
✓	✓		✓		65.10	88.02	64.15	87.86
✓	✓	✓	✓		67.01	89.52	67.13	88.96
✓	✓	✓	✓	✓	67.31	89.43	67.41	88.78

Table 3. **Component-wise Ablation on AerialVG.** The baseline is a hyperparameter-optimized PropVG model trained with the original ground-truth annotations. $\mathcal{V}_{tea}$ denotes fixed expression-aligned pseudo annotations generated offline by a trained AerialVG teacher; $\mathcal{F}_{app}$, $\mathcal{F}_{topo}$, and $\mathcal{F}_{filt}$ denote appearance binding, topological binding, and text-aware filtering, respectively. $\mathcal{F}_{topo}$ uses a symmetric k -NN graph.

Pipeline Strategy	Top1 (%)	Top5 (%)
$\mathcal{F}_{ind} + \mathcal{F}_{topo}$ (no appearance binding)	65.10	88.02
$\mathcal{F}_{ind} + \mathcal{F}_{app} + \mathcal{F}_{topo} + \text{NMS}$	67.25	90.03
$\mathcal{F}_{ind} + \mathcal{F}_{filt} + \mathcal{F}_{app} + \mathcal{F}_{filt} + \mathcal{F}_{topo}$	66.75	89.46
$\mathcal{F}_{ind} + \mathcal{F}_{filt} + \mathcal{F}_{app} + \mathcal{F}_{topo}$	67.31	89.43

Table 4. **Pipeline design choices on AerialVG.** Only the post-induction candidate-processing strategy is changed.

Anchor Topology	Scales	Top1 (%)	Top5 (%)
Single Ring	[1.0]	64.59	87.15
Double Ring	[0.5, 1.0]	64.57	87.51
Double Ring	[1.0, 1.2]	64.36	87.56
Triple Ring (Ours)	[0.5, 1.0, 1.2]	65.59	87.89

Table 5. **Ablation of Intra-Instance Appearance Binding on AerialVG.** All variants use proposal induction followed by appearance binding; filtering and topological binding are disabled.

Appearance-Binding Design. Table 5 studies the spatial layout of the adaptive sampling anchors. These experiments isolate appearance binding by disabling filtering and topological binding. Single- and double-ring layouts achieve Top1 values between 64.36% and 64.59%. The triple-ring layout with scales [0.5, 1.0, 1.2] improves Top1 to 65.59% and Top5 to 87.89%. Covering the hypothesis interior, boundary, and immediate surroundings therefore provides more discriminative evidence than concentrating all sampling points at one or two scales.

Topology and Propagation Depth. Table 6 compares dense interaction, Delaunay triangulation, and the default KNN graph while keeping the remaining pipeline fixed. Dense all-pairs interaction reaches only 63.98% Top1, consistent with the hypothesis that distant or weakly relevant candidates introduce distracting context. Delaunay triangulation improves Top1 to 66.21%, whereas the 6-NN

Relational Topology	Layers	Top1 (%)	Top5 (%)
<i>Topology Strategy</i>			
Dense (AerialVG-style)	4	63.98	87.07
KNN ($k = 6$)	4	67.31	89.43
Delaunay Triangulation	4	66.21	88.65
<i>Graph Depth with KNN ($k = 6$)</i>			
KNN	2	64.34	88.32
KNN	4	67.31	89.43
KNN	6	67.96	88.51
KNN	8	63.79	87.58

Table 6. **Ablation of Inter-Instance Topological Binding on AerialVG.** The upper block changes the graph structure; the lower block changes only the depth of the default KNN graph.

Prior Source	Source Baseline	With GrabVG	Gain (pp)
Training GT Boxes	56.76	58.94	+2.18
LLM + Grounding-DINO [20]	56.83	60.06	+3.23
LLM + SAM3 [2]	56.97	60.36	+3.39
AerialVG-Derived [19]	62.84	67.31	+4.47
Oracle (All GT Boxes)*	69.62	74.23	+4.61

Table 7. **Robustness to hypothesis priors on the AerialVG test set.** All values are Top1 (Acc@0.5). LLM denotes Qwen3 [32]. *The oracle setting supplies all ground-truth boxes and omits $\mathcal{F}_{filt}$.

graph reaches 67.31%. Sparse local connectivity is therefore more effective than unrestricted interaction in crowded UAV scenes. The depth comparison reveals a complementary trade-off. Increasing the graph from two to four layers improves Top1 by 2.97 percentage points and produces the highest Top5 value. Six layers obtain the highest Top1 (67.96%) but reduce Top5, while eight layers degrade both metrics. We use four layers because they provide the most balanced localization and candidate-recall performance without excessive feature propagation.

Robustness to Hypothesis Priors. GrabVG can be trained with proposal supervision from different sources. Table 7 compares original training boxes, two open-vocabulary pseudo-annotation pipelines, AerialVG-derived pseudo annotations, and an oracle setting containing all ground-truth boxes. Relative to the corresponding source baseline, the complete refinement stack improves Top1 by 2.18–4.47 percentage points for all non-oracle sources. With perfect oracle hypotheses, appearance and topological binding still yield a 4.61-point gain even though filtering is omitted. These consistent improvements indicate that the binding modules are not tied to one teacher or proposal distribution; they remain useful under both imperfect and ideal candidate coverage.

Inference Efficiency. Table 8 reports inference speed under the same evaluation setup. AerialVG runs at 13.65 FPS, whereas GrabVG reaches 28.51 FPS and approaches the 32.00 FPS of OTA-Det. Without hypothesis filtering, $\mathcal{F}_{app}$ and $\mathcal{F}_{topo}$ run at 25.09 FPS. Introducing $\mathcal{F}_{filt}$ reduces the number of candidates processed by the two binding modules and increases the speed to 28.51 FPS. These results show that early candidate screening recovers part of the computational overhead introduced by graph-attentive binding.

Method	FPS (img/s)
<i>Existing Baselines</i>	
AerialVG [19]	13.65
OTA-Det [29]	32.00
<i>GrabVG Variants</i>	
$\mathcal{F}_{ind}$	30.32
+ $\mathcal{F}_{app}$ (Intra-binding)	30.03
+ $\mathcal{F}_{app}$ + $\mathcal{F}_{topo}$ (Inter-binding)	25.09
+ $\mathcal{F}_{filt}$ + $\mathcal{F}_{app}$ + $\mathcal{F}_{topo}$ (Full GrabVG)	28.51

Table 8. **Inference-speed comparison.** FPS results for GrabVG, its component variants, and existing baselines under the same evaluation setup. All FPS results are measured on a single NVIDIA RTX 4090 GPU.

4.5. Qualitative Results

Progressive Disambiguation. Figure 3 illustrates how the two binding modules progressively resolve appearance and relational ambiguity. Without either module, the model selects a blue vehicle whose appearance and relative position are both inconsistent with the expression. After introducing **Intra-Instance Appearance Binding**, the prediction shifts to a white van, indicating that language-guided local visual sampling captures the referred appearance attribute; however, the selected instance still violates the specified “bottom-right” relation. With **Inter-Instance Topological Binding** further incorporated, the model propagates spatial evidence from neighboring hypotheses and identifies the white van whose surrounding configuration satisfies the expression.

Binding Mechanism Visualization. Figure 4 illustrates how GrabVG binds topological and appearance evidence. In the upper part, thin yellow arrows denote sparse graph edges, the thick red arrow marks the highest-attended edge, and the polar plots summarize the directions of highly weighted edges. For “a white SUV on the left” and “a silver sedan at the top right,” the dominant directions point west and northeast, respectively, showing that **Inter-Instance Topological Binding** emphasizes neighboring hypotheses consistent with the described spatial relations. In the lower part, the top-8 sampling points from **Intra-Instance Appearance Binding** concentrate on the referred objects and

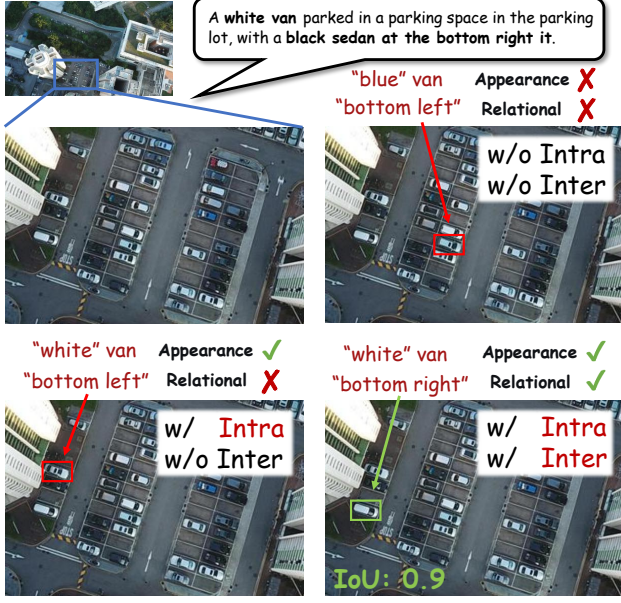


Figure 3. **Qualitative ablation of progressive feature binding.** Appearance binding corrects the target attribute, while topological binding resolves the remaining spatial ambiguity.

discriminative regions, such as the dark body of the black SUV and the white roof and orange body of the bus. Together, these visualizations show that the two modules capture expression-relevant relational directions and fine-grained local appearance evidence, respectively.

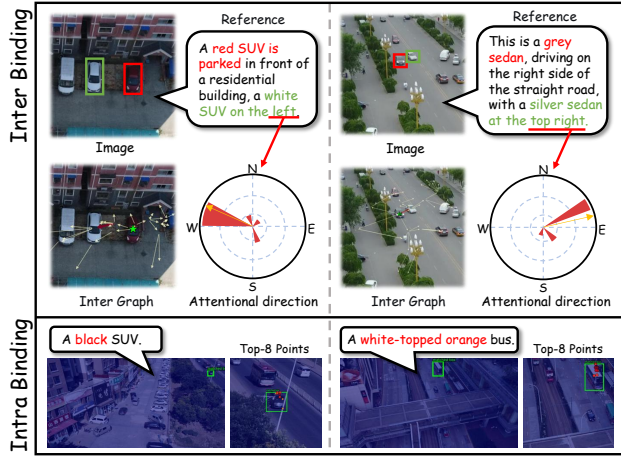


Figure 4. **Visualization of Graph-Attentive Feature Binding.** Top: sparse graph edges and directional attention distributions, where thick red arrows indicate the highest-attended edges. Bottom: top-8 adaptive sampling points highlighting appearance-relevant regions.

5. Conclusion

We presented GrabVG for visual grounding in crowded UAV imagery. The central idea is to separate candidate-space construction from evidence-based referent selection: expression-aligned pseudo-annotation supervision and text-aware filtering first establish a manageable candidate pool, after which adaptive visual sampling and sparse graph attention compare local appearance and geometric neighborhood information. This design limits the influence of background queries and distant candidate interactions while retaining cues needed to distinguish nearby, appearance-similar objects. GrabVG achieves 67.31% Top1 on AerialVG and 80.34% on AerialSense, exceeding the corresponding baselines by 10.55 and 8.76 percentage points. The ablation results further show that proposal supervision, appearance sampling, local graph propagation, and early filtering provide complementary improvements across the evaluated settings. Future work will explore scene-adaptive retention and neighborhood settings.

References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [2] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.
- [3] Jinming Chai, Lingling Li, Licheng Jiao, Xiaoqiang Lu, Long Sun, Xu Liu, Wenping Ma, and Weibin Li. Recs4r: Bridging semantics and geometry for referring remote sensing interpretation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 42213–42224, 2026.
- [4] Chenglizhao Chen, Shaofeng Liang, Runwei Guan, Xiaolou Sun, Haocheng Zhao, Haiyun Jiang, Tao Huang, Henghui Ding, and Qing-Long Han. Aerialmind: Towards referring multi-object tracking in uav scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2805–2813, 2026.
- [5] Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *Advances in neural information processing systems*, 37:121670–121698, 2024.
- [6] Ming Dai, Wenxuan Cheng, Jiedong Zhuang, Jiang-jiang Liu, Hongshen Zhao, Zhenhua Feng, and Wankou Yang. Propvg: End-to-end proposal-driven visual grounding with multi-granularity discrimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7058–7068, 2025.
- [7] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual ground-

- ing with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1769–1779, 2021.
- [8] Kanoko Goto, Takumi Hirose, Mahiro Ukai, Shuhei Kurita, and Nakamasa Inoue. Referring expression comprehension for small objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21231–21242, 2025.
 - [9] Mingning Guo, Mengwei Wu, Shaoxian Li, Haifeng Li, and Chao Tao. Towards accurate uav image perception: Guiding vision-language models with stronger task prompts. *arXiv preprint arXiv:2512.07302*, 2025.
 - [10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
 - [11] Shiqi Huang, Shuting He, and Bihan Wen. Rsground-r1: Rethinking remote sensing visual grounding through spatial reasoning. *arXiv preprint arXiv:2601.21634*, 2026.
 - [12] Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25472–25483, 2026.
 - [13] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.
 - [14] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27831–27840, 2024.
 - [15] Hengyou Li, Xinyan Liu, and Guorong Li. A benchmark for uav-view natural language-guided tracking. *Electronics*, 13(9):1706, 2024.
 - [16] Ke Li, Ting Wang, Di Wang, Yongshan Zhu, Yiming Zhang, Tao Lei, and Quan Wang. Provg: Progressive visual grounding via language decoupling for remote sensing imagery. *arXiv preprint arXiv:2604.01893*, 2026.
 - [17] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.
 - [18] 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.
 - [19] Junli Liu, Qizhi Chen, Zhigang Wang, Yiwen Tang, Yiting Zhang, Chi Yan, Dong Wang, Xuelong Li, and Bin Zhao. Aerialvg: A challenging benchmark for aerial visual grounding by exploring positional relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5177–5187, 2025.
 - [20] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Euro-pean conference on computer vision*, pages 38–55. Springer, 2024.
 - [21] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.
 - [22] Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1181–1198, 2024.
 - [23] Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866, 2023.
 - [24] Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. Scan-former: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13449–13458, 2024.
 - [25] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.
 - [26] 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.
 - [27] Peng Wang, Qi Wu, Jiawei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1960–1968, 2019.
 - [28] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.
 - [29] Guoting Wei, Xia Yuan, Yang Zhou, Haizhao Jing, Yu Liu, Xianbiao Qi, Chunxia Zhao, Haokui Zhang, and Rong Xiao. Open-text aerial detection: A unified framework for aerial visual grounding and detection. *arXiv preprint arXiv:2602.07827*, 2026.
 - [30] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.
 - [31] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection

- in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.
- [32] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [33] Sibe Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4145–4154, 2019.
- [34] Sibe Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4644–4653, 2019.
- [35] Sibe Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9952–9961, 2020.
- [36] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4683–4693, 2019.
- [37] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15502–15512, 2022.
- [38] 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.
- [39] 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.
- [40] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE transactions on geoscience and remote sensing*, 61:1–13, 2023.
- [41] Peirong Zhang, Yidan Zhang, Luxiao Xu, Jinliang Lin, Zonghao Guo, Fengxiang Wang, Xue Yang, Kaiwen Wei, and Lei Wang. Geovis: Geospatially rewarded visual search for remote sensing visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14335–14345, 2026.
- [42] Xindi Zhang, Ebroul Izquierdo, and Krishna Chandramouli. Dense and small object detection in uav vision based on cascade network. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.
- [43] Zilun Zhang, Zian Guan, Tiancheng Zhao, Haozhan Shen, Yuxiang Cai, Zhonggen Su, Yongheng Shang, Zhaojun Liu, Jianwei Yin, and Xiang Li. Geo-r1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:113–129, 2026.
- [44] Yue Zhou, Jue Chen, Zilun Zhang, Penghui Huang, Ran Ding, Zhentao Zou, Pengfei Gao, Yuchen Wei, Ke Li, Xue Yang, et al. Dvgbench: Implicit-to-explicit visual grounding benchmark in uav imagery with large vision–language models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232:831–847, 2026.
- [45] 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.
- [46] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.
- [47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.