GRABVG: GRAPH-ATTENTIVE BINDING FOR VISUAL GROUNDING IN UAV IMAGERY

Anonymous authors
Paper under double-blind review

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. The code will be made publicly available.

1 INTRODUCTION

Visual grounding in Unmanned Aerial Vehicle (UAV) imagery aims to localize a target object in a bird's-eye-view scene according to a natural-language expression Liu et al. (2025); Guo et al. (2025); Chen et al. (2026); Li et al. (2024). Compared with conventional ground-level imagery, UAV scenes typically contain numerous small and densely distributed objects, many of which exhibit highly similar appearances Xia et al. (2018); Zhu et al. (2018); Zhang et al. (2019); Liu et al. (2025). The limited visual distinctions among these instances create *high visual redundancy*, making appearance cues alone insufficient for distinguishing the referred object from nearby distractors Liu et al. (2025). Moreover, the elevated viewpoint weakens perspective and depth cues while bringing a larger number of objects into view simultaneously. 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 Deng et al. (2021); Wei et al. (2026), while relation-based approaches incorporate spatial configurations among instances Liu et al. (2025). 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 discrim-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

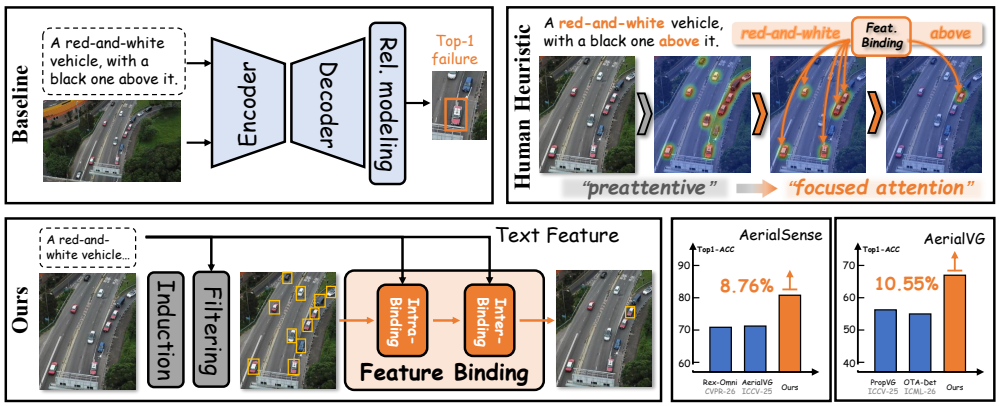

Figure 1: **Comparison of visual grounding paradigms.** Existing methods rely on global decoding or dense relational modeling, often leaving visually similar hypotheses insufficiently distinguished Liu et al. (2025). Human visual search first narrows the search through preattentive guidance and then binds discriminative evidence with focused attention Wolfe (1994); Treisman & Gelade (1980). GrabVG follows this progressive process through hypothesis induction, hypothesis filtering, and graph-attentive feature binding.

inative 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 Wolfe (1994), while Feature Integration Theory posits that focused attention is required to bind distributed visual features into coherent object representations Treisman & Gelade (1980). 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. **Contributions.** 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. Early proposal-based methods rank detected regions using appearance, location, and contextual cues Yu et al. (2016; 2018), while one-stage and transformer-based approaches directly regress the referred region from cross-modal representations Yang et al. (2019c); Liao et al. (2020); Deng et al. (2021); Kamath et al. (2021); Zhu et al. (2022). Subsequent studies improve visual–language correspondence through language-guided feature refinement Ye et al. (2022); Su et al. (2023), dynamic visual sampling Shi et al. (2024), language-irrelevant token removal Su et al. (2024), and decoupled multimodal fusion Dai et al. (2024). Grounded pre-training, proposal-driven grounding, and small-object refinement further extend this paradigm Li et al. (2022); Liu et al. (2024); Dai et al. (2025); Goto et al. (2025). However, most operate on dense multimodal features or generic region proposals without 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 Zhan et al. (2023); Kuckreja et al. (2024); Wei et al. (2026), relational reasoning Liu et al. (2025); Zhou et al. (2026), and semantic–geometric modeling Zhang et al. (2026a); Chai et al. (2026). ProVG further adopts a progressive *survey–locate–verify* strategy that sequentially injects global, relational, and attribute cues into dense visual features Li et al. (2026). 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 Huang et al. (2026); Zhang et al. (2026b). These approaches offer explicit rationales and improved low-shot reasoning, but require extra supervision or rollout-based optimization and incur autoregressive inference overhead. GrabVG instead performs fixed-depth differentiable reasoning over sparse candidate graphs without generating textual reasoning trajectories.

Relation Modeling for Visual Grounding. Early methods compare target proposals with surrounding objects or decompose expressions into subject, location, and relation components Yu et al. (2016; 2018). Later methods use language-guided, cross-modal, multi-step, or scene-graph reasoning Wang et al. (2019); Yang et al. (2019a;b; 2020). AerialVG explicitly captures positional relations among aerial objects Liu et al. (2025), 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: $\mathbf{F}_{img} = \{\mathbf{F}_{img}^{s}\}_{s=1}^{S} = \mathcal{E}_v(\mathbf{I})$ and $\mathbf{F}_{txt} = \mathcal{E}_l(\mathbf{T})$. A DETR-style proposal generator $\mathcal{F}_{ind}$ Carion et al. (2020); Zhu et al. (2020) 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: $\mathcal{H}^0 \xrightarrow{\mathcal{F}_{filt}} \mathcal{H}^f \xrightarrow{\mathcal{F}_{app}} \mathcal{H}^a \xrightarrow{\mathcal{F}_{topo}} \mathcal{H}^g$. Here, Text-Aware Hypothesis Filtering $\mathcal{F}_{filt}$ scores the initial hypotheses using $\mathbf{F}_{txt}$ and retains K foreground

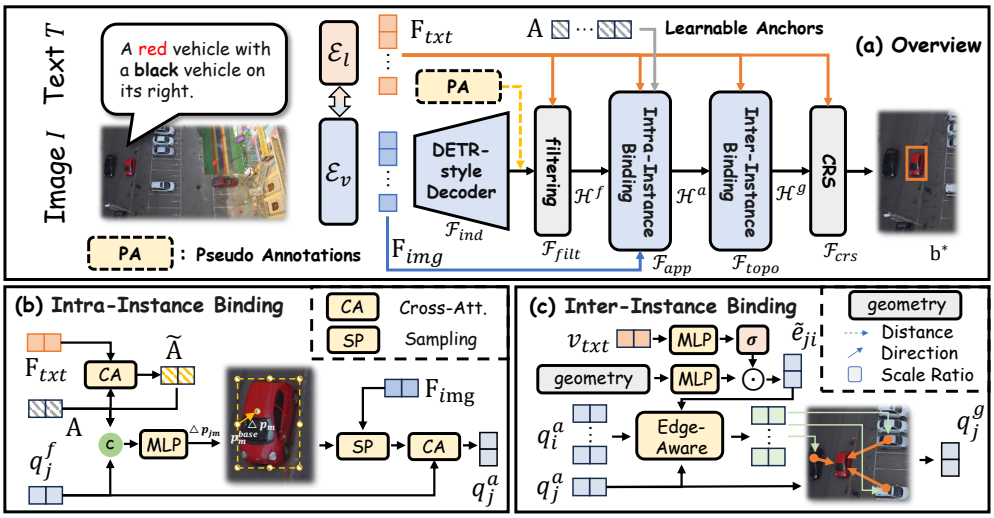

Figure 2: **Overall architecture of GrabVG.** (a) The framework induces and filters expression-aligned hypotheses, then applies intra-instance appearance binding, inter-instance topological binding, and CRS for final target selection. (b) Appearance binding samples language-guided local features from each hypothesis and its surroundings. (c) Topological binding propagates language-conditioned geometric evidence over a sparse graph.

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 $\mathcal{H}^g = \{(\mathbf{q}_j^g, \mathbf{b}_j)\}_{j=1}^K$. We adopt the Contrastive-based Refer Scoring (CRS) module from PropVG Dai et al. (2025) 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: $s_j^{ref} = \mathcal{F}_{crs}(\mathbf{q}_j^g, \mathbf{F}_{txt})$, $\mathbf{b}^* = \mathbf{b}_{j^*}$, and $j^* = \arg\max_j s_j^{ref}$.

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 Hinton et al. (2015) and employ an external teacher pipeline $\mathcal{F}_{tea}$ Liu et al. (2025; 2024) to generate an expression-aligned pseudo-annotation set offline, $\mathcal{Y}_{tea} = \left(\{\overline{\mathbf{b}}_r\}_{r=1}^R, r^*\right) = \mathcal{F}_{tea}(\mathbf{I}, \mathbf{T})$, where $\overline{\mathbf{b}}_r$ 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. Further teacher details are provided in the supplementary material.

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 $\{\overline{\mathbf{b}}_r\}_{r=1}^R$ 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 Vaswani et al. (2017) injects linguistic context into the proposal features for relevance estimation:

$$\begin{aligned} \mathbf{U} &= \mathrm{LN}\big(\mathbf{Q}^0 + \mathrm{MHA}(\mathbf{Q}^0, \mathbf{F}_{txt}, \mathbf{F}_{txt})\big)\,, \\ \overline{\mathbf{Q}} &= \mathrm{LN}(\mathbf{U} + \mathrm{FFN}(\mathbf{U}))\,, \qquad \mathbf{s}^f = \sigma(\mathbf{W}_f \overline{\mathbf{Q}}). \end{aligned} \tag{1}$$

Here, $\mathrm{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes a multi-head attention block. The top-K scoring hypotheses are retained: $\Omega = \mathrm{TopK}(\mathbf{s}^f, K) = \{\omega_1, \dots, \omega_K\}$ and $\mathcal{H}^f = \{(\mathbf{q}^0_{\omega_j}, \mathbf{b}^0_{\omega_j})\}_{j=1}^K \triangleq \{(\mathbf{q}^f_j, \mathbf{b}_j)\}_{j=1}^K$. The auxiliary features $\overline{\mathbf{Q}}$ are used only to estimate filtering scores, while the selected original decoder features $\{\mathbf{q}^0_{\omega_j}\}_{j=1}^K$ 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:

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

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:

$$\begin{aligned} \mathbf{F}_j^{loc} &= \mathrm{Sample}\big(\mathbf{F}_{img}, \{\mathbf{p}_{jm}\}_{m=1}^M\big)\,, \\ \mathbf{q}_j^a &= \mathrm{MHA}\left(\mathbf{q}_j^f, [\mathbf{q}_j^f; \mathbf{F}_j^{loc}], [\mathbf{q}_j^f; \mathbf{F}_j^{loc}]\right). \end{aligned} \tag{3}$$

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,

$\mathcal{N}(j) = \mathrm{SymKNN}_k(\mathbf{c}_j) \cup \{j\}$. 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: $\widetilde{\mathbf{e}}_{ji} = \mathrm{MLP}_{geo}(\mathbf{r}_{ji}) \odot \sigma(\mathrm{MLP}_{txt}(\mathbf{v}_{txt}))$. 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:

$$
\begin{aligned}
a_{ji}^{\ell} &= \frac{1}{\sqrt{d}} \left(\mathbf{W}_q^{\ell} \mathbf{z}_j^{(\ell-1)} \right)^{\top} \left(\mathbf{W}_k^{\ell} \mathbf{z}_i^{(\ell-1)} + \mathbf{W}_{ek}^{\ell} \widetilde{\mathbf{e}}_{ji} \right) + \left(\mathbf{w}_{eb}^{\ell} \right)^{\top} \widetilde{\mathbf{e}}_{ji}, \\
\alpha_{ji}^{\ell} &= \mathrm{Softmax}_{i \in \mathcal{N}(j)} \left(a_{ji}^{\ell} \right), \\
\mathbf{m}_j^{\ell} &= \sum_{i \in \mathcal{N}(j)} \alpha_{ji}^{\ell} \left(\mathbf{W}_v^{\ell} \mathbf{z}_i^{(\ell-1)} + \mathbf{W}_{ev}^{\ell} \widetilde{\mathbf{e}}_{ji} \right), \\
\mathbf{z}_j^{(\ell)} &= \mathrm{GraphBlock}^{\ell} \left(\mathbf{z}_j^{(\ell-1)}, \mathbf{m}_j^{\ell} \right).
\end{aligned}
\tag{4}
$$

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 $\mathcal{H}^g = \{(\mathbf{q}_j^g, \mathbf{b}_j)\}_{j=1}^K$. Further implementation details of both binding modules are provided in the supplementary material.

3.4 Training Objective

GrabVG is jointly optimized with proposal-induction, hypothesis-filtering, grounding, and auxiliary segmentation objectives: $\mathcal{L} = \lambda_{det}\mathcal{L}_{det} + \lambda_{filt}\mathcal{L}_{filt} + \lambda_{ref}\mathcal{L}_{ref} + \lambda_{mask}\mathcal{L}_{mask}$. The proposal-induction loss $\mathcal{L}_{det}$ follows DETR Carion et al. (2020) and combines foreground classification, ℓ_1 box regression, and generalized IoU losses Rezatofighi et al. (2019). The grounding loss $\mathcal{L}_{ref}$ and auxiliary segmentation loss $\mathcal{L}_{mask}$ follow PropVG Dai et al. (2025). The filtering loss $\mathcal{L}_{filt}$ combines binary relevance supervision from bipartite-matched pseudo boxes with a retention penalty for pruning the referred hypothesis, while $\mathcal{L}_{ref}$ performs final referent discrimination over the retained hypotheses. Complete target definitions and loss details are provided in the supplementary material.

4 Experiments

4.1 Setup

Datasets. We evaluate GrabVG on **AerialVG** Liu et al. (2025) and the visual-grounding subset of **AerialSense** Guo et al. (2025). 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 Dai et al. (2025), UAV-specific visual grounding methods Liu et al. (2025); Wei et al. (2026), and zero-shot grounding methods Jiang et al. (2026).

Implementation Details. Following PropVG Dai et al. (2025), we adopt the BEiT-3 Base Wang et al. (2023) backbone and a three-layer deformable proposal decoder. Images are resized with the

Table 1: **Comparison with state-of-the-art methods on AerialVG.** "*" denotes hyperparameter optimization, and bold marks the best result. Top1 and Top5 correspond to Acc@0.5 and Top-5 Acc@0.5, respectively.

Method	Top1 (%)	Top5 (%)	mIoU (%)
Zero-Shot Methods			
ReX-Omni Jiang et al. (2026)	28.37	30.09	-
Supervised Methods			
TransVG Deng et al. (2021)	11.53	13.68	-
D-MDETR Shi et al. (2024)	19.87	29.87	-
G-DINO Liu et al. (2024)	29.36	78.87	-
AerialVG Liu et al. (2025)	50.03	87.00	-
PropVG Dai et al. (2025)	49.40	-	49.38
PropVG* Dai et al. (2025)	56.76	74.05	46.76
OTA-Det Wei et al. (2026)	54.90	-	-
GrabVG (Ours)	**67.31**	**89.43**	**53.34**

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

Method	Type	Top1 (%)	mIoU (%)
Large Multimodal Models (Zero-Shot)			
ReX-Omni Jiang et al. (2026)	Zero-shot	71.47	-
Domain-Specific Models			
AerialVG Liu et al. (2025)	Supervised	71.58	-
PropVG* Dai et al. (2025)	Supervised	61.30	67.76
GrabVG (Ours)	Supervised	**80.34**	**72.00**

aspect ratio preserved and padded to 586×586. 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. Complete optimization and pseudo-annotation settings are provided in the supplementary material.

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

Table 3: **Component-wise ablation on AerialVG.** The baseline is hyperparameter-optimized PropVG trained on the original ground-truth annotations. $\mathcal{Y}_{tea}$ 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. $\mathcal{F}_{topo}$ uses a symmetric k-NN graph.

Baseline	GrabVG Components				Test		Val	
	$\mathcal{Y}_{tea}$	$\mathcal{F}_{app}$	$\mathcal{F}_{topo}$	$\mathcal{F}_{filt}$	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

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. More detailed analyses of the effectiveness and design choices of individual components are provided in the supplementary material.

Inference Efficiency. Table 4 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.

Table 4: **Inference-speed comparison.** GrabVG, its component variants, and existing baselines are evaluated under the same setup on a single NVIDIA RTX 4090 GPU.

Method	FPS (img/s)
Existing Baselines	
AerialVG Liu et al. (2025)	13.65
OTA-Det Wei et al. (2026)	32.00
GrabVG Variants	
$\mathcal{F}_{ind}$	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

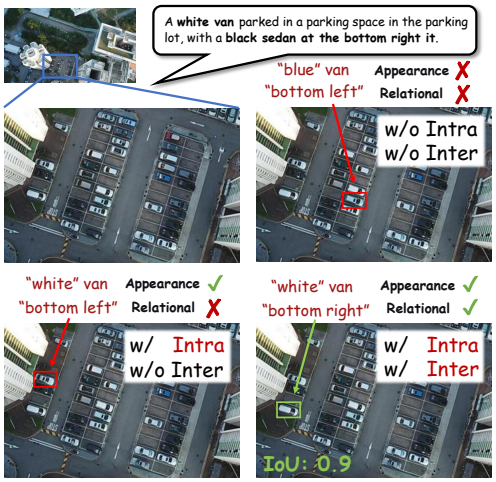

Figure 3: **Qualitative ablation of progressive feature binding.** Appearance binding resolves attribute ambiguity, while topological binding resolves spatial ambiguity.

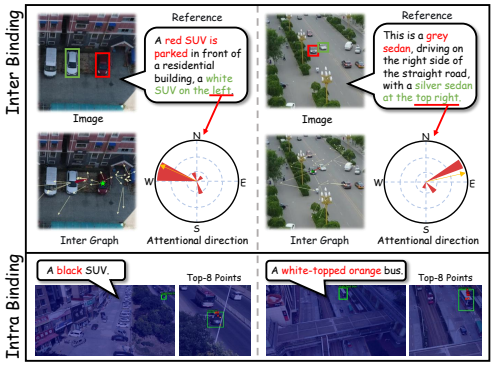

Figure 4: **Visualization of Graph-Attentive Feature Binding.** Top: sparse graph edges and directional attention distributions, where thick red arrows indicate the highest-attended edges. Bottom: top-8 adaptive sampling points highlighting appearance-relevant regions.

4.4 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 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.

5 Conclusion

GrabVG separates expression-aligned candidate construction from appearance- and topology-based selection in crowded UAV imagery. Distillation-guided proposal induction and text-aware filtering first establish a compact hypothesis space, after which adaptive sampling and sparse graph attention integrate intra-instance appearance and inter-instance topology. This design suppresses background queries and distant interactions while retaining the fine-grained cues needed to distinguish nearby, visually similar objects. GrabVG reaches $67.31\%/80.34\%$ Top1 on AerialVG/AerialSense, exceeding the corresponding baselines by $10.55/8.76$ points. The ablations confirm complementary gains from the four components, indicating that candidate screening and structured local reasoning provide a favorable accuracy–efficiency trade-off for UAV grounding. Future work will study scene-adaptive retention and neighborhoods.

AI USE STATEMENT

Generative AI tools were used only to assist with code generation and English-language polishing. All generated code and edited text were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

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*, pp. 213–229. Springer, 2020.

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.

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*, pp. 42213–42224, 2026.

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*, volume 40, pp. 2805–2813, 2026.

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.

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*, pp. 7058–7068, 2025.

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

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*, pp. 21231–21242, 2025.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Shiqi Huang, Shuting He, and Bihan Wen. Rsground-r1: Rethinking remote sensing visual grounding through spatial reasoning. *arXiv preprint arXiv:2601.21634*, 2026.

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*, pp. 25472–25483, 2026.

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*, pp. 1780–1790, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

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*, pp. 27831–27840, 2024.

Hengyou Li, Xinyan Liu, and Guorong Li. A benchmark for uav-view natural language-guided tracking. *Electronics*, 13(9):1706, 2024.

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.

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*, pp. 10965–10975, 2022.

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*, pp. 10880–10889, 2020.

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*, pp. 5177–5187, 2025.

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 *European conference on computer vision*, pp. 38–55. Springer, 2024.

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*, pp. 658–666, 2019.

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.

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*, pp. 10857–10866, 2023.

Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. Scanformer: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13449–13458, 2024.

Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

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.

Peng Wang, Qi Wu, Jiewei 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*, pp. 1960–1968, 2019.

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*, pp. 19175–19186, 2023.

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.

Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

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*, pp. 3974–3983, 2018.

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.

Sibei 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*, pp. 4145–4154, 2019a.

Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4644–4653, 2019b.

Sibei 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*, pp. 9952–9961, 2020.

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*, pp. 4683–4693, 2019c.

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*, pp. 15502–15512, 2022.

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

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

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.

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*, pp. 14335–14345, 2026a.

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*, pp. 0–0, 2019.

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, 2026b.

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.

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*, pp. 598–615. Springer, 2022.

Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.

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.

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

OVERVIEW

This supplementary material provides additional methodological details, experimental analyses, and qualitative results that complement the main paper. Its organization is summarized as follows.

Appendix A: Additional Method and Training Details. This appendix presents the pseudo-annotation generation process, the implementation details of the two binding modules, and the complete training targets and loss settings.

Appendix B: Additional Experimental Details and Ablations. This appendix reports supplementary optimization settings and analyzes the filtering strategy, query budget, binding-module configurations, and different hypothesis-prior sources.

Appendix C: Additional Qualitative Results. This appendix provides qualitative analyses of the two binding modules and comparisons with existing visual grounding methods.

A ADDITIONAL METHOD AND TRAINING DETAILS

A.1 PSEUDO-ANNOTATION GENERATION

Pseudo annotations are generated offline from the training split using either (i) an end-to-end grounding teacher such as AerialVG Liu et al. (2025), or (ii) an LLM–open-vocabulary pipeline. In the latter, Qwen3 Yang et al. (2025) extracts the object category from the expression after removing attributes and relations (e.g., "a red vehicle with a white vehicle on its left" becomes "vehicle"), and Grounding-DINO Liu et al. (2024), SAM3 Carion et al. (2025), or a compatible model predicts category-relevant boxes. After NMS, the retained boxes form the pseudo foreground set. The ground-truth referent is appended only when no prediction satisfies the matching criterion. For both pipelines, the corresponding instance masks are generated by providing each target bounding box to SAMKirillov et al. (2023) as a box prompt. Teachers are not used during validation or inference.

A.2 BINDING IMPLEMENTATION DETAILS

Intra-Instance Appearance Binding. The formulation is given in the main paper. We use $M = 48$ anchors distributed on three box-relative square rings with scale factors $[0.5, 1.0, 1.2]$. The inner, boundary, and outer rings cover the candidate interior and immediate context. Features are bilinearly sampled from four visual levels, concatenated, and projected to the model dimension. The original candidate feature is included among the keys and values of local attention to preserve decoder semantics when sampled evidence is unreliable. Both text conditioning and local aggregation use multi-head attention with residual, layer-normalization, and feed-forward updates.

Inter-Instance Topological Binding. For edge $j \leftarrow i$, we define the relative displacement, distance, normalized direction, and log-scale difference as

$$\begin{aligned} \Delta\mathbf{c}_{ji} &= \mathbf{c}_i - \mathbf{c}_j, \qquad d_{ji} = \|\Delta\mathbf{c}_{ji}\|_2, \\ \mathbf{u}_{ji} &= \frac{\Delta\mathbf{c}_{ji}}{d_{ji} + \epsilon}, \quad \Delta\boldsymbol{\rho}_{ji} = \left[\log\frac{w_i}{w_j}; \log\frac{h_i}{h_j}\right]. \end{aligned} \tag{5}$$

The raw geometric descriptor is then

$$\mathbf{r}_{ji} = \left[\Delta\mathbf{c}_{ji}; d_{ji}; \log(d_{ji} + \epsilon); \mathbf{u}_{ji}; \Delta\boldsymbol{\rho}_{ji}\right]. \tag{6}$$

Using masked mean pooling, the sentence feature and language-conditioned edge representation are computed as

$$\begin{aligned} \mathbf{v}_{txt} &= \frac{\sum_l m_l \mathbf{t}_l}{\sum_l m_l + \epsilon}, \\ \mathbf{e}_{ji} &= \mathrm{MLP}_{geo}(\mathbf{r}_{ji}), \\ \mathbf{g}_{txt} &= \sigma(\mathrm{MLP}_{txt}(\mathbf{v}_{txt}))\,, \\ \widetilde{\mathbf{e}}_{ji} &= \mathbf{e}_{ji} \odot \mathbf{g}_{txt}. \end{aligned} \tag{7}$$

The edge-aware attention equations follow the main paper. We use eight heads, a 64-dimensional edge representation, and graph blocks composed of output projection, residual connection, layer normalization, and an FFN. After L_g layers, the output gate is

$$\begin{aligned} \mathbf{g}_j^{out} &= \sigma\Big(\mathrm{MLP}_{out}([\mathbf{q}_j^a; \mathbf{z}_j^{(L_g)}])\Big)\,, \\ \mathbf{q}_j^g &= \mathbf{q}_j^a + \mathbf{g}_j^{out} \odot \Big(\mathbf{z}_j^{(L_g)} - \mathbf{q}_j^a\Big)\,. \end{aligned} \tag{8}$$

A.3 Exact Training Targets and Weights

DETR matching and proposal losses follow the main paper, including auxiliary losses from intermediate decoder layers. For hypothesis filtering, let $\mathcal{M}_b$ denote the bipartite matching between induced hypotheses and pseudo annotations for sample b. A hypothesis is treated as a positive filtering target if it is matched to any pseudo box:

$$y_{b,i}^f = \mathbb{I}\left[\exists r,\ (i,r) \in \mathcal{M}_b\right]. \tag{9}$$

Let i_b^r be the hypothesis matched to the referred pseudo box and Ω_b the retained top-K set. The retention loss is averaged only over samples in which the referent is pruned:

$$\begin{aligned} \mathcal{B}_{miss} &= \{b \mid i_b^r \notin \Omega_b\}, \\ \ell_b^{keep} &= -\log\Big(s_{b,i_b^r}^f + \epsilon\Big)\,, \\ \mathcal{L}_{keep} &= \begin{cases} \dfrac{1}{|\mathcal{B}_{miss}|} \displaystyle\sum_{b \in \mathcal{B}_{miss}} \ell_b^{keep}, & |\mathcal{B}_{miss}| > 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{10}$$

The complete filtering loss combines binary relevance supervision and referent retention:

$$\mathcal{L}_{filt} = \lambda_{bce}\,\mathrm{BCE}(\mathbf{s}^f, \mathbf{y}^f) + \lambda_{keep}\mathcal{L}_{keep}. \tag{11}$$

For retained hypothesis j with original index ω_{bj},

$$y_{bj}^{ref} = \mathbb{I}[\omega_{bj} = i_b^r], \qquad \mathcal{L}_{ref} = \mathrm{BCE}(\mathbf{s}^{ref}, \mathbf{y}^{ref}). \tag{12}$$

When the referent is not retained, all retained hypotheses receive negative referring labels and $\mathcal{L}_{keep}$ raises the missed referent's filtering score. We use $\lambda_{det} = 0.025$ and $\lambda_{filt} = \lambda_{ref} = \lambda_{mask} = 1$. The classification, ℓ_1, and generalized-IoU terms in $\mathcal{L}_{det}$ are weighted by 1, 5, and 2; the BCE and retention terms in $\mathcal{L}_{filt}$ are weighted by 0.3 and 1.

B Additional Experimental Details and Ablations

B.1 Additional Optimization Details

The dataset splits, metrics, architecture, input resolution, and query budgets are reported in the main paper. The maximum text length is 64. The models are trained using Adam for 36 epochs 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 then decayed by 0.1; on AerialVG, decay occurs at epochs 24 and 32.

Table 5: **Pipeline design choices.** All variants use the same AerialVG-derived pseudo-annotation supervision for $\mathcal{F}_{ind}$. The table compares learned filtering placement with NMS-based candidate screening.

Pipeline Strategy	Acc@0.5 (%)	Top-5 (%)
$\mathcal{F}_{ind}$ + $\mathcal{F}_{topo}$ (No Intra)	65.10	88.02
$\mathcal{F}_{ind}$ + $\mathcal{F}_{app}$ + $\mathcal{F}_{topo}$ + 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 6: **Query-budget control experiment on AerialVG.** All variants provide 128 hypotheses to the appearance and topological binding modules. The comparison disentangles the effects of text-aware filtering and the initial proposal-query budget.

Initial Queries	Filtering	Binding Queries	Acc@0.5 (%)	Top-5 (%)
128	None	128	66.18	88.53
128	$\mathcal{F}_{filt}$	128	66.60	87.01
256	$\mathcal{F}_{filt}$	128	**67.31**	**89.43**

B.2 Pipeline Design Analysis

Hypothesis-Filtering Strategy. Table 5 investigates the placement and frequency of **Text-Aware Hypothesis Filtering** $\mathcal{F}_{filt}$. All rows use the same Distillation-Guided Proposal Induction trained with AerialVG-derived pseudo annotations; only the post-induction candidate-processing strategy changes. Applying the filter both before and after **Intra-Instance Appearance Binding** slightly reduces Acc@0.5 to 66.75%. Compared with the default single early filtering strategy at 67.31%, this degradation suggests that repeated pruning may remove ambiguous but potentially correct hypotheses before they can be disambiguated through topological reasoning. Replacing the learned filter with NMS gives a similar Acc@0.5 but a higher Top-5 value, showing that localization accuracy and candidate-set recall respond differently to the screening strategy.

Query Budget versus Text-Aware Filtering. Table 6 disentangles the effects of text-aware filtering and the initial proposal-query budget while fixing the number of hypotheses processed by the subsequent binding modules to 128. With 128 initial queries, introducing $\mathcal{F} * filt$ improves Acc@0.5 from 66.18% to 66.60%, although Top-5 accuracy decreases from 88.53% to 87.01%. This result suggests that the auxiliary filtering supervision provides a modest benefit to top-1 referent discrimination, but cannot compensate for the limited coverage of a small initial hypothesis set. In contrast, generating 256 hypotheses and selecting 128 of them using $\mathcal{F} * filt$ achieves 67.31% Acc@0.5 and 89.43% Top-5 accuracy, outperforming the 128-query filtered variant by 0.71 and 2.42 percentage points, respectively. Since both configurations pass the same number of hypotheses to appearance and topological binding, the improvement cannot be attributed to a larger downstream computational budget. Instead, a broader initial hypothesis pool improves candidate coverage, while text-aware filtering preserves the most relevant hypotheses for subsequent fine-grained binding. These results indicate that generating a sufficiently diverse hypothesis set before filtering is more effective than directly reducing the initial query budget.

B.3 Detailed Analysis of Feature Binding

Intra-Instance Appearance Binding. Table 7 analyzes the spatial layout of the adaptive sampling anchors in $\mathcal{F}_{app}$. To isolate appearance binding, these experiments use Distillation-Guided Proposal Induction followed by $\mathcal{F}_{app}$, without $\mathcal{F}_{filt}$ or $\mathcal{F}_{topo}$; this is the same appearance-only configuration reported in the component ablation of the main paper. We compare single-, double-, and triple-ring layouts to study hypothesis-proximal coverage. The triple-ring configuration with scale factors $[0.5, 1.0, 1.2]$ achieves the best Acc@0.5 of 65.59%. Extending the outer ring to $1.2\times$ the hypothesis

Table 7: **Ablation on Intra-Instance Appearance Binding.** All rows use the induction-plus-appearance configuration with AerialVG-derived pseudo annotations; filtering and topological binding are disabled.

Anchor Topology	Scales	Acc@0.5 (%)	Top-5 (%)
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 8: **Ablation on Inter-Instance Topological Binding.** The upper part compares relational graph structures, while the lower part evaluates the depth of the default KNN graph. All experiments use AerialVG-derived hypothesis priors. All topology comparisons use the same proposal induction, filtering, and appearance-binding configuration; only the relational graph structure is changed.

Relational Topology	Layers	Acc@0.5 (%)	Top-5 (%)
Topology Strategy			
Dense (AerialVG-style)	4	63.98	87.07
KNN ($k = 6$)	4	**67.31**	**89.43**
Delaunay Triangulation	4	66.21	88.65
Graph Depth with KNN ($k = 6$)			
KNN	2	64.34	88.32
KNN	4	67.31	**89.43**
KNN	6	**67.96**	88.51
KNN	8	63.79	87.58

box allows the model to capture immediate surrounding evidence that can distinguish visually similar UAV targets.

Inter-Instance Topological Binding. Table 8 examines the graph topology and propagation depth of $\mathcal{F}_{topo}$ while keeping proposal induction, hypothesis filtering, and appearance binding unchanged. In the topology comparison, dense all-pairs attention achieves only 63.98% Acc@0.5, suggesting that unrestricted interaction introduces distracting information from distant or weakly relevant hypotheses. Both sparse alternatives perform substantially better. Delaunay triangulation reaches 66.21% Acc@0.5, whereas the KNN graph with $k = 6$ achieves 67.31%, outperforming it by 1.10 percentage points. This result suggests that restricting message passing to a fixed number of nearby candidates provides a more effective local relational structure for crowded UAV scenes.

The depth analysis further shows that increasing the number of graph-attention layers from two to four improves Acc@0.5 from 64.34% to 67.31% and Top-5 accuracy from 88.32% to 89.43%. Six layers yield the highest Acc@0.5 of 67.96%, but reduce Top-5 accuracy to 88.51%. Increasing the depth to eight layers causes a marked decline in both metrics, indicating that excessive propagation may lead to redundant feature mixing or over-smoothing among neighboring candidates. We therefore use four layers in all main experiments, as this configuration provides the best Top-5 accuracy and a favorable balance between localization performance and graph depth.

Robustness to Hypothesis Priors. GrabVG can use proposal supervision from different sources. Table 9 evaluates original training annotations, LLM–Grounding-DINO detections, LLM–SAM3 proposals, AerialVG-derived pseudo annotations, and an oracle setting with all ground-truth boxes. Relative to the corresponding source baseline, adding the GrabVG refinement and binding stages improves Acc@0.5 by 2.18 to 4.61 percentage points. For the first four rows, the GrabVG result

Table 9: **Robustness to hypothesis priors on AerialVG (test).** All entries report Acc@0.5. Source Baseline denotes the result before adding the GrabVG refinement stack. With GrabVG includes filtering and both binding modules for rows 1–4. LLM refers to Qwen3 Yang et al. (2025). * In the oracle setting, all ground-truth boxes are supplied as hypotheses and $\mathcal{F}_{filt}$ is omitted.

Prior Source	Source Baseline	With GrabVG	Absolute Gain (pp)
1. Training GT Boxes	56.76	58.94	**+2.18**
2. LLM + Grounding-DINO Liu et al. (2024)	56.83	60.06	**+3.23**
3. LLM + SAM3 Carion et al. (2025)	56.97	60.36	**+3.39**
4. AerialVG-Derived Liu et al. (2025)	62.84	67.31	**+4.47**
5. Oracle (All GT Boxes)*	69.62	74.23	**+4.61**

Table 10: **Effect of pseudo-annotation sources on AerialSense.** All other model and training configurations are kept unchanged.

Pseudo-Annotation Source	Acc@0.5 (%)	mIoU (%)
AerialVG-derived	64.04	69.52
LLM + Grounding-DINO	**80.34**	**72.00**

includes filtering and both binding modules. In the oracle row, perfect ground-truth hypotheses are supplied directly and $\mathcal{F}_{filt}$ is omitted, so the comparison isolates the contribution of feature binding under ideal proposal coverage.

Effect of Pseudo-Annotation Sources on AerialSense. We further examine the effect of different pseudo-annotation sources on AerialSense. As shown in Table 10, replacing the AerialVG-based teacher with the LLM–Grounding-DINO pipeline improves Acc@0.5 from 64.04% to 80.34% and mIoU from 69.52% to 72.00%.

The two teacher pipelines produce substantially different raw prediction distributions. Before the referent-coverage correction described in Section A.1, the AerialVG-based teacher retains four predictions per expression, whereas the LLM–Grounding-DINO pipeline produces a scene-adaptive number of retained predictions ranging from 1 to 62, with an average of 6.80. In either pipeline, the final pseudo-annotation set may contain one additional box when the ground-truth referent must be appended. Thus, the AerialVG-derived supervision is nearly fixed-cardinality after correction, while the Grounding-DINO-derived supervision remains strongly variable-cardinality. This variation exposes the proposal generator to more diverse candidate densities and ambiguity levels. Differences in referent coverage, localization quality, and domain compatibility may also contribute to the performance gap. These results motivate the use of Grounding-DINO-derived pseudo annotations for the main AerialSense experiments.

C Additional Qualitative Results

Attribute and Relation Cases. Figure 5 qualitatively compares the two binding modules under daytime and nighttime conditions. Without either module, the model is easily distracted by candidates with incorrect appearance and spatial relations. Intra-Instance Appearance Binding helps identify candidates with the correct attribute or category, but may still select an instance with an incorrect relative position. Inter-Instance Topological Binding further resolves this ambiguity by modeling the geometric relations among neighboring hypotheses. These examples demonstrate that the two modules are complementary: appearance binding improves target recognition, while topological binding supports accurate relational reasoning.

Comparison with Existing Methods. Figure 6 compares GrabVG with PropVGDai et al. (2025) and AerialVGLiu et al. (2025) in crowded UAV scenes. The referring expressions involve visually similar vehicles and multiple spatial constraints, which cause both baselines to select nearby distractors or unrelated regions. In contrast, GrabVG consistently localizes the correct referent with high IoU across different object scales and scene layouts. These results suggest that the constrained hypothesis

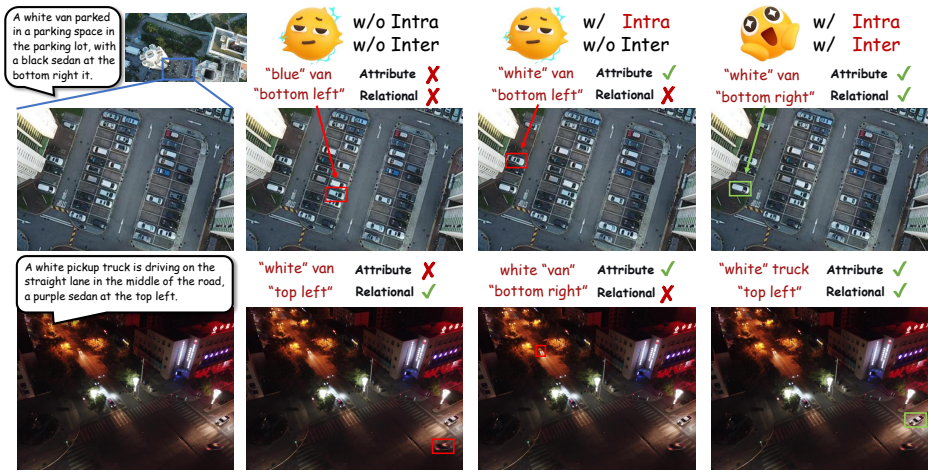

Figure 5: **Grounding under attribute and relational variation.** The examples contrast spatial descriptions such as "bottom left" and "bottom right" and include both intra-class and inter-class distractors in UAV imagery.

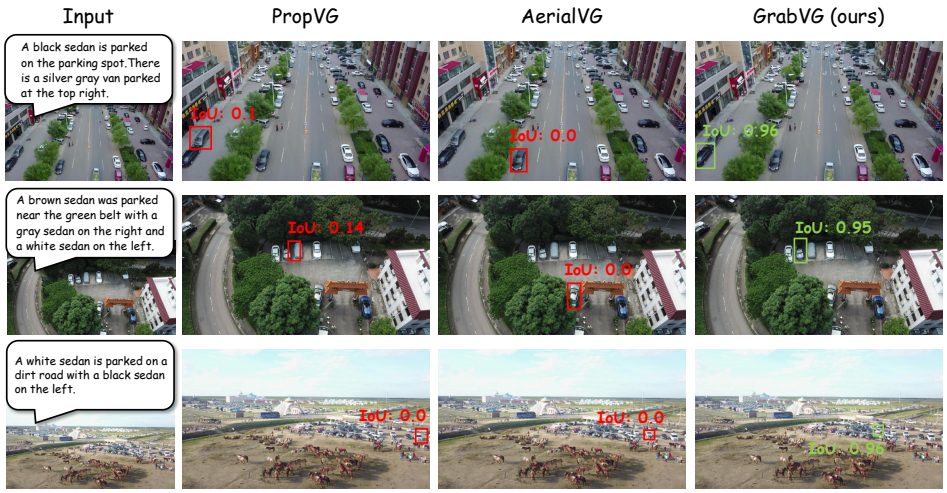

Figure 6: **Qualitative comparison on AerialVG.** From left to right: the referring expression and image, followed by predictions from PropVG, AerialVG, and GrabVG. Green boxes denote correct predictions with high IoU, and red boxes denote localization errors.

space and complementary appearance–topology binding enable more reliable discrimination among densely distributed candidates.