HIERARCHICAL REASONING-GUIDED PROGRESSIVE AERIAL VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex aerial visual grounding requires not only understanding what a referring expression describes, but also coordinating how its different semantic components guide visual localization. This challenge is amplified in UAV imagery by small objects, dense same-category instances, and complex spatial layouts. We propose **Hierarchical Reasoning-Guided Aerial Visual Grounding (HRVG)**, a unified framework that organizes semantic guidance throughout a progressive grounding process. The key idea is to align different semantic roles with corresponding visual computations, rather than treating semantic reasoning as an auxiliary language representation. HRVG derives hierarchical scene-, entity-, and relation-level semantic states from language reasoning and progressively integrates them into visual grounding, enabling contextual understanding, candidate grounding, and relational decision-making within a single connected framework. We further introduce **RefScene-20K**, a benchmark comprising 20,000 referring expressions over 8,550 UAV images across relational, ordinal, extremal, multi-target, and no-target tracks, with outputs uniformly represented as potentially empty sets of bounding boxes. Experiments on RefScene-20K demonstrate that HRVG outperforms existing grounding baselines. Further analyses, including semantic decomposition, reasoning-order, and stage-intervention studies, verify the effectiveness of organizing hierarchical semantics and aligning them with progressive visual computations. These results highlight semantic-role alignment as an effective modeling principle for complex aerial visual grounding.

1 INTRODUCTION

Visual grounding aims to localize the visual region referred to by a natural language expression. Recent advances in vision-language models have significantly improved grounding ability by learning direct alignment between text and visual regions. However, existing approaches (Liu et al., 2025; Sun et al., 2025; Wang et al., 2026) generally formulate grounding as a text-object matching problem, where the entire expression is represented as a unified semantic embedding and matched with visual features.

Such a formulation is insufficient for complex referring expressions. Aerial expressions often contain multiple levels of semantic information, including scene context, target entities, and relationships among objects. For example, ”the second red vehicle behind the truck” requires the model to first understand the surrounding scene, identify relevant vehicle candidates, and finally reason about their spatial relationships. Treating the expression as a single representation may obscure these hierarchical semantics and make it difficult to associate different semantic cues with corresponding visual processing stages.

UAV imagery further amplifies this challenge. Compared with conventional images, aerial scenes contain small objects, densely distributed instances of the same category, and complex spatial layouts. Multiple objects may share similar appearance, making target identification dependent on contextual and relational information rather than appearance alone. Existing aerial visual grounding benchmarks (Liu et al., 2025; Sun et al., 2025) have explored positional relations and variable target cardinality, but a unified framework that explicitly connects hierarchical language semantics with progressive visual grounding remains underexplored.

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.

Motivated by these observations, we propose Hierarchical Reasoning-Guided Progressive Aerial Visual Grounding (HRVG), a unified framework that organizes semantic guidance throughout visual localization. HRVG integrates text-side Hierarchical Semantic Reasoning and Reasoning-Vision Alignment to connect different semantic roles with corresponding visual computations. Given a referring expression, HRVG derives scene-, entity-, and relation-level semantic states from language reasoning and progressively incorporates them into the grounding process. The key idea is not semantic decomposition alone, but its coupling with a continuous visual inference process: different semantic roles guide different computations, while intermediate visual representations are propagated across stages. By coordinating contextual understanding, candidate grounding, and relational disambiguation within a single framework, HRVG provides a coherent solution for resolving complex aerial referring expressions.

To systematically evaluate complex aerial grounding, we further introduce RefScene-20K, a reasoning-centric UAV visual grounding benchmark containing relational, ordinal, extremal, multi-target, and no-target expressions. RefScene-20K complements existing benchmarks by providing a unified evaluation setting for diverse semantic requirements and target cardinalities.

Our contributions are summarized as follows:

- **A unified framework for hierarchical reasoning-guided grounding.** We propose HRVG, which models *Text-side Hierarchical Reasoning* and *Reasoning–Vision Alignment* as a single progressive visual grounding framework. This coupling unifies scene-aware encoding, candidate selection, and relational prediction into a connected grounding process.
- **A benchmark for complex aerial referring expressions.** We introduce RefScene-20K, which supports studying diverse referential requirements within a common evaluation setting, with particular attention to ordinal references alongside established expression types.
- **Systematic evaluation of hierarchical semantic guidance.** We evaluate HRVG against representative grounding baselines and investigate its design through component ablations, text-side reasoning analyses, reasoning-order comparisons, and stage-specific interventions. These experiments support the effectiveness of the unified framework and show that the organization and use of intermediate semantics influence the final grounding decisions.

2 Related Work

2.1 Visual Grounding and Referring Expression Comprehension

Visual grounding associates referring expressions with image regions. Structured approaches such as MAttNet (Yu et al., 2018) explicitly model subject, location, and relationship information. Transformer-based methods instead emphasize joint visual–linguistic representation learning: TransVG (Deng et al., 2021) directly regresses the referred box, MDETR (Kamath et al., 2021) learns phrase–region alignment through text-conditioned detection, and Dynamic MDETR (Shi et al., 2024) progressively refines localization through adaptive visual sampling and text-guided decoding. Grounded pretraining further improves localization. GLIP (Li et al., 2022) unifies detection and phrase grounding, while Grounding DINO (Liu et al., 2024) introduces language interaction in feature enhancement, query selection, and decoding.

Intermediate reasoning provides another relevant direction. Chain-of-thought prompting (Wei et al., 2022) generates intermediate reasoning steps, and Grounded Chain-of-Thought (Wu et al., 2026) extends this idea to stepwise recognition and localization of visual cues. Meanwhile, GREC (He et al., 2023) broadens REC to zero, one, or multiple referents using gRefCOCO (Liu et al., 2023). HRVG first generates a chain-of-thought response and extracts scene, entity, and relation fields from it; these fields then guide corresponding visual computations, with intermediate representations and candidate information passed forward.

2.2 Aerial and UAV Visual Grounding

Remote-sensing grounding addresses scale variation and complex backgrounds through multi-scale cross-modal learning. RSVG (Zhan et al., 2023) introduces a benchmark and multi-level feature learning, while LQVG (Lan et al., 2024) uses language queries to aggregate object information from multi-scale visual features. UAV-specific benchmarks further emphasize dense object configurations



Figure 1: Representative examples and statistics of RefScene-20K. Left: six types of referring expression samples. Top-right: pie chart of referring-type distribution. Bottom-right: word cloud of referring expression vocabulary. Red boxes denote targets, orange boxes reference evidence; no-target expressions have empty target sets despite visible scene objects.

and diverse referential requirements. AerialVG (Liu et al., 2025) studies positional relations among visually similar instances and combines hierarchical cross-attention with relation-aware grounding. RefDrone (Sun et al., 2025) covers small targets, contextual expressions, and multi-target or no-target cases, introducing Number GroundingDINO to incorporate target-number information.

Recent work explores richer semantic and candidate-level reasoning. DVGBench (Zhou et al., 2026) evaluates explicit and implicit UAV references, with DroneVG-R1 combining implicit-to-explicit chain-of-thought and reinforcement learning. GrabVG (Wang et al., 2026) performs object-hypothesis search followed by graph-attentive feature binding. Building on these advances, our HRVG focuses on coordinating different semantic roles throughout a unified, progressive localization process. RefScene-20K complements existing benchmarks by explicitly organizing relational, ordinal, and extremal requirements alongside multi-target and no-target prediction, supporting evaluation across diverse semantic conditions and target cardinalities.

3 RefScene-20K Benchmark

RefScene-20K is designed to evaluate complex referring-expression grounding in UAV imagery, with an emphasis on *semantic coverage* and *multi-constraint expressions*. Existing aerial benchmarks have advanced the evaluation of positional relations and variable target cardinality (Liu et al., 2025; Sun et al., 2025). RefScene-20K complements these settings by establishing explicit relational, ordinal, and comparison/extremal evaluation tracks alongside multi-target and no-target prediction. The objective is not to increase sentence length for its own sake, but to evaluate grounding decisions that require the coordinated interpretation of scene context, entity descriptions, and referential conditions.

Scale and Semantic Coverage. We build RefScene-20K on VisDrone2019-DET (Zhu et al., 2021). It contains **20,000 referring expressions** over **8,550 UAV images**. As shown in Fig. 1, the relational, ordinal, and comparison/extremal tracks each contain 5,000 expressions, while the multi-target and no-target tracks contain 3,000 and 2,000 expressions, respectively. Each of the first three tracks accounts for 25% of the benchmark, providing substantial coverage for analyzing different referential requirements. In particular, ordinal references receive a dedicated evaluation track rather than being subsumed under general spatial descriptions. Across these tracks, expressions incorporate object categories and appearance attributes together with scene restrictions, inter-object relations, ordering, and extremal comparisons.

Complex Multi-constraint Expressions. A central design goal is to include expressions in which multiple semantic conditions jointly determine the referent. Consider *"the fifth white car from the left among the cars parked along the road."* This query combines an object category (*car*), an appearance attribute (*white*), a scene restriction (*parked along the road*), an ordering direction (*from*

the left), and an ordinal condition (*fifth*). Recognizing a white car alone does not resolve the query when several instances satisfy the appearance description: the model must identify the relevant group and evaluate the ordering within that group. Similar dependencies arise when an expression combines reference objects, spatial relations, and contextual restrictions. RefScene-20K emphasizes such combinations to test semantic coordination beyond isolated object recognition.

Language Characteristics. Fig. 1 summarizes the vocabulary of the benchmark. It includes category and attribute terms such as *car*, *sedan*, *white*, and *black*; scene terms such as *road*, *intersection*, and *parking*; spatial descriptors such as *left*, *right*, *near*, and *along*; and ordinal terms such as *second*, *third*, and *fourth*. These frequencies illustrate the range of language used in the annotations. Per-expression semantic tags, rather than word frequency alone, support analysis of how different requirements occur together.

Construction and Verification. Target sets and geometric conditions are established before expression generation. Generated expressions are checked against the intended targets through rule-based verification, an independent vision–language checker, and human review, with particular attention to dense scenes, high ordinal ranks, and difficult negatives. This procedure supports traceable target assignments and correction of expressions that do not preserve the intended referents. Detailed source preparation, generation prompts, geometric conventions, verification procedures, and split construction are described in the appendix.

4 METHOD

4.1 PROBLEM FORMULATION

Unlike conventional referring expression comprehension that assumes a single target, complex aerial expressions may correspond to zero, one, or multiple objects. Given a UAV image I and a referring expression T, the goal of visual grounding is to identify the target objects described by the expression. We formulate grounding as a set prediction problem:

$$\widehat{\mathcal{Y}} = f_\theta(I, T) = \{\widehat{b}_m\}_{m=1}^{\widehat{M}}, \tag{1}$$

where f_θ is the grounding model parameterized by θ. Here, $\widehat{b}_m$ is the m-th predicted box and $\widehat{M} \geq 0$ is the predicted target cardinality; in particular, $\widehat{M} = 0$ represents a no-target prediction. Complex aerial expressions contain different semantic roles, including scene context, entities, and relational constraints. For example, "the second white car behind the truck near the building" combines scene, entity, and relational semantics that respectively guide visual scope, candidate identification, and final disambiguation. HRVG models these roles as different computational stages of the same grounding process:

$$T \rightarrow \{R_s, R_e, R_r\} \rightarrow \{H_s, H_e, H_r\} \rightarrow \widehat{\mathcal{Y}}, \tag{2}$$

where R_s, R_e, R_r denote scene-, entity-, and relation-level semantics, and H_s, H_e, H_r denote the corresponding scene-, entity-, and relation-conditioned visual states. The arrows denote the direction of semantic decomposition and progressive visual processing, rather than separate prediction branches. HRVG forms a continuous grounding path where later stages inherit intermediate visual states. The proposed framework is formulated as:

$$H_s = \mathcal{A}_s(H_0, R_s), \quad H_e = \mathcal{A}_e(H_s, R_e), \quad H_r = \mathcal{A}_r(H_e, R_r), \tag{3}$$

where H_0 is the initial visual representation extracted from I; $\mathcal{A}_s$, $\mathcal{A}_e$, and $\mathcal{A}_r$ denote scene-, entity-, and relation-guided transformations, performing scene-aware encoding, candidate construction, and relational disambiguation, respectively. The final prediction is

$$\widehat{\mathcal{Y}} = \mathcal{G}(H_r). \tag{4}$$

Here, $\mathcal{G}$ denotes the grounding output head that maps the final visual state H_r to the predicted box set $\widehat{\mathcal{Y}}$.

4.2 HIERARCHICAL SEMANTIC REASONING

HRVG aligns different semantic roles with corresponding stages of a unified grounding process. Given an expression T, a frozen language model generates hierarchical semantic states through

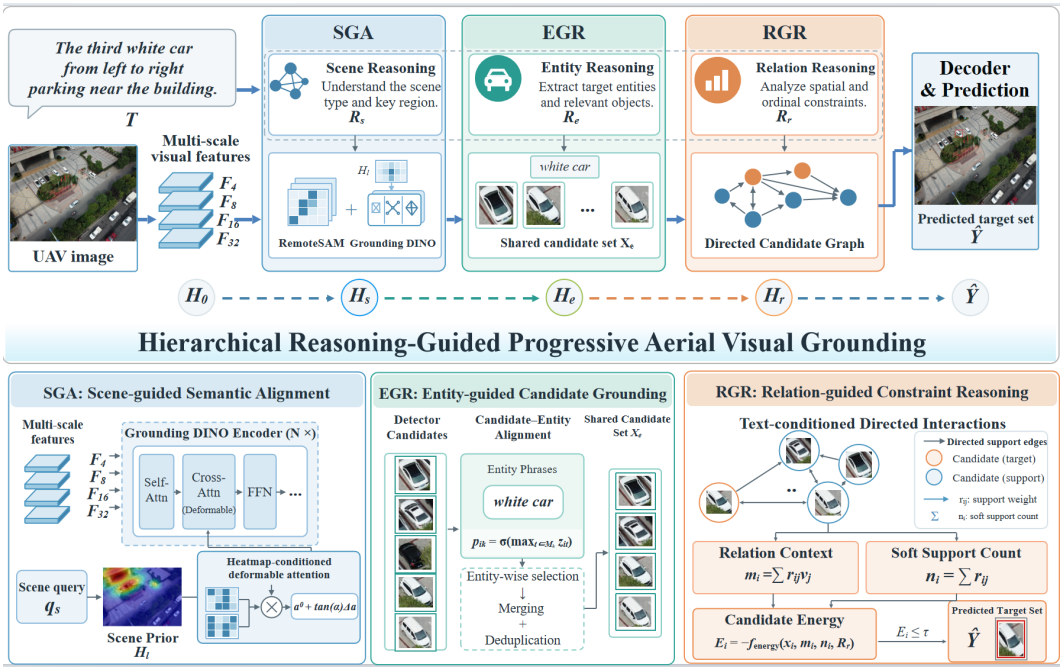

Figure 2: Overview of HRVG. The text-side model first generates a chain-of-thought response and extracts scene-, entity-, and relation-level semantics, which progressively guide corresponding visual computations. By aligning semantic roles with successive grounding stages, HRVG forms a unified progressive visual grounding framework.

chain-of-thought reasoning and structured extraction:

$$(C, J) = R_\phi^{\text{cot}}(T), \qquad (R_s, R_e, R_r) = \mathcal{E}(J). \tag{5}$$

where R_ϕ^{cot} denotes chain-of-thought generation with frozen parameters ϕ, J is the extracted structured output, and $\mathcal{E}$ is a deterministic extraction operator. We instantiate R_ϕ with frozen Qwen3-8B-Instruct and apply one fixed reasoning-and-extraction prompt to every split.

4.3 Reasoning–Vision Alignment

Scene-guided Semantic Alignment. Scene semantics guide visual encoding through a continuous spatial prior. Given multi-scale feature maps F_l from the frozen RemoteSAM (Yao et al., 2025) backbone and a scene-query vector q_s obtained by encoding and pooling R_s, we compute:

$$H_l(u) = \sigma\left(\left\langle \text{norm}(W_F^{(l)} F_l(u)), \text{norm}(W_q^{(l)} q_s) \right\rangle / \tau_l\right). \tag{6}$$

where $l \in \{4, 8, 16, 32\}$ indexes feature scales, u is a spatial location on scale l, and $F_l(u)$ is its feature vector. $W_F^{(l)}$ and $W_q^{(l)}$ are learnable scale-specific projections; $\text{norm}(\cdot)$ denotes ℓ_2 normalization; $\langle \cdot, \cdot \rangle$ is the inner product; $\tau_l > 0$ is a learnable temperature; and $\sigma(\cdot)$ is the sigmoid function. The scalar $H_l(u) \in (0, 1)$ is the scene-relevance response at u, and H_l is the response map at scale l rather than a binary mask. Collectively, these maps contribute to the scene-conditioned state H_s.

Scene-guided semantic alignment (SGA) uses this prior to condition Grounding DINO's encoder attention (Liu et al., 2024). It applies a bounded, zero-initialized residual to the attention logits and thereby reweights sampled visual features without changing sampling locations or hard-masking regions.

Entity-guided Candidate Grounding. After scene-aware encoding, Grounding DINO processes T to produce candidate boxes b_i, feature vectors h_i, confidence scores s_i, and token-level alignment logits. For entity phrases from R_e, these logits are aggregated into candidate–entity alignment vectors $\boldsymbol{p}_i$ (details in Appendix C.3). After entity-wise selection and deduplication, we obtain the

shared candidate set:

$$X_e = \{x_i\}_{i=1}^{N}, \qquad x_i = (b_i, h_i, \boldsymbol{p}_i, s_i). \tag{7}$$

The candidate state X_e preserves both potential targets and reference objects, providing the visual evidence for subsequent relation-guided target selection rather than directly determining the final prediction.

Relation-guided Constraint Reasoning. The relation-guided constraint reasoning (RGR) stage evaluates these candidates under relation semantics R_r by modeling text-conditioned directed interactions between candidate pairs.

Specifically, RGR aggregates relational support and soft counting information to jointly capture *which* candidates support a target and *how many* candidates provide such support. The resulting relation-aware candidate score is computed as:

$$E_i = -f_{\text{energy}}(x_i, m_i, n_i, R_r), \tag{8}$$

where m_i denotes relation context and n_i denotes soft count. The final target set is obtained by:

$$\widehat{\mathcal{Y}} = \{b_i \mid x_i \in X_e,\ E_i \le \tau\}. \tag{9}$$

This unified formulation enables relational, ordinal, and extremal constraints to be handled within the same candidate reasoning process.

4.4 Optimization

The semantic parser remains frozen, and the visual grounding path is optimized with detection, scene, graph, and final-selection objectives:

$$\mathcal{L} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{scene}}\mathcal{L}_{\text{scene}} + \lambda_{\text{graph}}\mathcal{L}_{\text{graph}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}}. \tag{10}$$

The detection objective $\mathcal{L}_{\text{det}}$ combines entity-alignment focal loss with box ℓ_1 and generalized-IoU losses. $\mathcal{L}_{\text{scene}}$ supervises scene responses and cross-scale consistency, $\mathcal{L}_{\text{graph}}$ supervises directed relations, ordering, routing, and soft-count prediction, and $\mathcal{L}_{\text{energy}}$ supervises membership in the ground-truth target set. Reference and comparison objects provide auxiliary supervision but are not treated as final positives; all candidates are negatives for no-target expressions.

Training follows the progressive inference dependency: scene-aware encoding and entity grounding are first stabilized, relation and count supervision are then introduced, and the complete grounding path is finally jointly fine-tuned.

5 Experiments

We evaluate HRVG on RefScene-20K and additionally evaluate it on AerialVG (Liu et al., 2025) and RefDrone (Sun et al., 2025). All models use the official training, validation, and test splits of each benchmark. The evaluated configurations use Swin-T and Swin-B backbones based on the Swin Transformer architecture (Liu et al., 2021).

Metrics. For RefScene-20K, single-target grounding is evaluated by IoU@0.5. Multi-target predictions are matched with ground-truth boxes using one-to-one IoU matching, and we report set-level F_1 averaged over queries. No-target performance is measured by the accuracy of correctly predicting an empty set. Avg. denotes the unweighted mean over the five primary tracks: relational, ordinal, extremal, multi-target, and no-target. For external benchmarks, we follow their original protocols: AerialVG (Liu et al., 2025) uses Top-1/Top-5 accuracy, while RefDrone reports instance- and image-level F_1 and accuracy.

Implementation. HRVG is built on Grounding DINO with Swin-T and Swin-B backbones, initialized from the corresponding pretrained checkpoints. RemoteSAM (Yao et al., 2025) provides multi-scale scene features, while the frozen text-side reasoning and extraction process is executed once per expression and its outputs are cached. Detector inputs preserve aspect ratio, with short-side scales of $\{640, 704, 768, 800, 864\}$ during training and 800 pixels during evaluation, with the long

Table 1: Main results on RefScene-20K. Rel., Ord., and Ext. denote relational, ordinal, and extremal reasoning. "–" indicates that the corresponding track was not reported.

Method	Backbone	Rel.	Ord.	Ext.	Multi F_1	No-target	Avg.
TransVG (Deng et al., 2021)	R-101	21.14	11.36	24.27	–	–	–
Dynamic-MDETR (Shi et al., 2024)	R-50	24.08	10.94	25.83	–	–	–
SimVG-GREC (Dai et al., 2024)	ViT-L	23.71	13.82	26.19	47.88	49.05	**32.13**
MDETR (Kamath et al., 2021)	R-101	26.42	15.37	27.61	52.48	53.17	**35.01**
GLIP-T (Li et al., 2022)	Swin-T	29.18	18.63	28.74	60.92	62.78	**40.05**
Grounding DINO-T (Liu et al., 2024)	Swin-T	33.86	21.74	35.18	86.92	87.70	**53.08**
Grounding DINO-B (Liu et al., 2024)	Swin-B	36.91	20.83	34.57	88.35	87.44	**53.62**
AerialVG (Liu et al., 2025)	Swin-L	30.94	19.27	35.62	–	–	–
NGDINO-T (Sun et al., 2025)	Swin-T	32.74	24.91	40.18	84.63	86.39	**53.77**
NGDINO-B (Sun et al., 2025)	Swin-B	35.68	26.37	37.42	87.96	86.77	**54.84**
OTA-Det-L (Wei et al., 2026)	DINOv3-L	36.48	24.12	39.05	85.96	85.94	**54.31**
HRVG-T	**Swin-T**	**40.82**	**30.17**	**43.58**	**91.74**	**93.04**	**59.87**
HRVG-B	**Swin-B**	**43.96**	**29.21**	**47.38**	**92.18**	**92.02**	**60.95**

Table 2: Ablation of visual grounding components on RefScene-20K. SGE denotes Scene-guided Visual Encoding, EGR denotes Entity-guided Candidate Grounding, RGS denotes Relation-guided Target Selection, and PI denotes Progressive Inheritance.

Variant	SGE	EGR	RGS	PI	Rel.	Ord.	Ext.	Multi F_1	No-target	Avg.
w/o SGE	×	✓	✓	✓	39.14	28.96	41.72	88.84	90.49	57.83
w/o EGR	✓	×	✓	✓	37.84	27.63	40.12	89.26	90.31	57.03
w/o RGS	✓	✓	×	✓	34.63	21.96	35.72	89.43	90.31	54.41
w/o PI	✓	✓	✓	×	32.84	20.73	34.15	88.36	90.37	53.29
HRVG-T	✓	✓	✓	✓	40.82	30.17	43.58	91.74	93.04	59.87

side capped at 1,333 pixels. RemoteSAM inputs are independently letterboxed to 896×896 pixels. At most 60 candidates are retained after entity-wise suppression and cross-entity deduplication. We use AdamW with an effective batch size of 16, an initial learning rate of 1×10^{-4}, and weight decay of 1×10^{-4}. The three training stages last 6, 8, and 6 epochs, respectively, with the learning rate reduced to 5×10^{-5} during joint fine-tuning. The final energy threshold τ_E is selected on the validation set and fixed for testing. Training uses four NVIDIA GeForce RTX 4090 GPUs with PyTorch 2.5.1.

5.1 Main Results on RefScene-20K

Table 1 compares HRVG with representative general-purpose and aerial visual grounding methods on RefScene-20K. The relation, ordinal, and extremal columns measure single-target reasoning performance, while Multi F_1 and No-target measure set prediction and empty-set recognition, respectively. The Avg. score is computed over the five primary tracks and is reported only when all five track scores are available. HRVG performs best overall: HRVG-T reaches 59.87 Avg., 5.03 points above the strongest non-HRVG result, while HRVG-B reaches 60.95 and improves over Grounding DINO-B by 7.33 points. Relative to Grounding DINO-T, HRVG-T gains 6.96, 8.43, and 8.40 points on relational, ordinal, and extremal grounding. Its 91.74 Multi F_1 and 93.04 No-target accuracy further show that the model can preserve multiple valid targets while rejecting unsupported candidate sets.

5.2 Qualitative Progressive Grounding

Fig. 3 visualizes the progressive grounding behavior of HRVG on two representative UAV queries. In the first example, "the bottommost red car among the red cars on the road," scene-level reasoning first restricts the search to the road, entity-level retrieval then collects the red-car candidates, and

Table 3: Analysis of text-side hierarchical reasoning on RefScene-20K. FE denotes full-expression guidance.

Variant	r_s	r_e	r_r	Rel.	Ord.	Ext.	Multi F_1	No-target	Avg.
Full Hierarchical Reasoning	✓	✓	✓	40.82	30.17	43.58	91.74	93.04	**59.87**
w/o Scene Reasoning	×	✓	✓	39.61	29.42	42.36	88.41	90.12	57.98
w/o Entity Reasoning	✓	×	✓	37.84	27.63	40.12	89.26	90.31	57.03
w/o Relation Reasoning	✓	✓	×	34.76	22.18	35.91	90.48	91.26	54.92
Flat Expression Guidance	FE	FE	FE	35.11	23.48	37.02	88.74	89.01	54.67

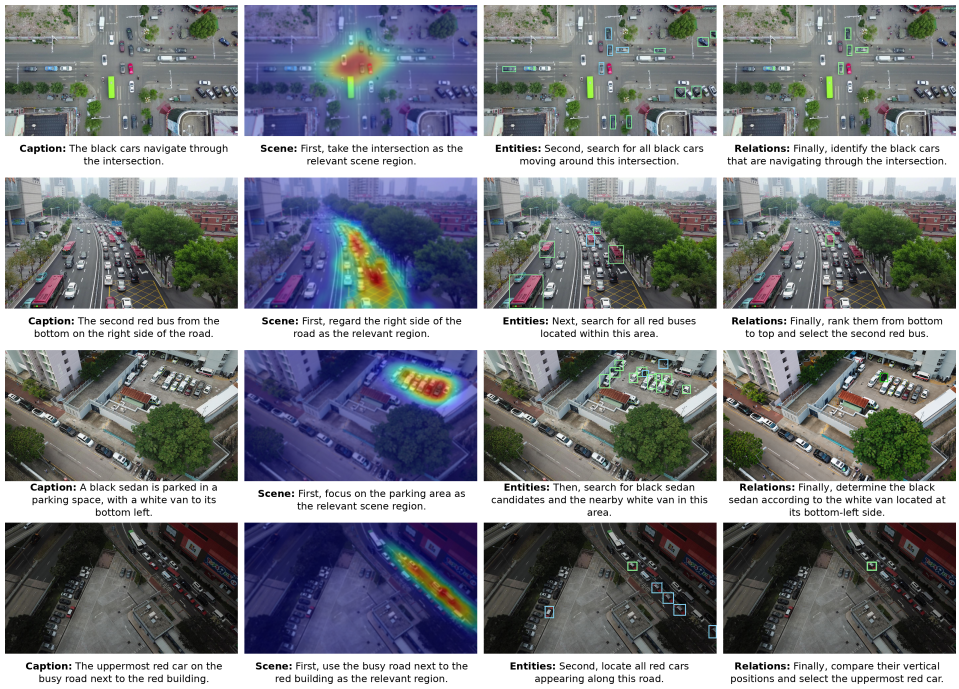

Figure 3: Qualitative visualization of hierarchical progressive grounding. Columns show the input query, scene-level focus, entity-level candidates, and final prediction. Rows illustrate multi-target, ordinal, relational, and extremal grounding examples, respectively.

constraint-guided selection applies the bottommost condition to return a single target. In the second example, "the white van near the buildings," scene reasoning identifies the building neighborhood, entity retrieval preserves all white-van candidates in that region, and the final stage returns the complete matching set. These cases illustrate why intermediate visual evidence must be inherited across stages: the scene mask reduces irrelevant regions, the candidate pool preserves comparison and multi-target instances, and the final constraint operation converts the accumulated evidence into the requested output cardinality.

5.3 Ablation Studies

We analyze the contribution of the visual grounding components and the hierarchical semantic design. Unless otherwise stated, all ablations use the Swin-T configuration and the same training and evaluation protocol as the full model.

Table 2 shows that each component contributes to the unified grounding framework. Removing progressive inheritance (PI) causes the largest Avg. decrease, from 59.87 to 53.29, highlighting the importance of propagating intermediate visual states across stages. Removing relation-guided target selection (RGS) causes the second-largest drop and particularly harms ordinal grounding, whose score decreases from 30.17 to 21.96. SGE and EGR further improve performance by providing scene-aware encoding and entity-level candidate grounding, respectively.

Table 4: Effect of semantic-to-visual stage assignment on RefScene-20K.

Semantic-to-Visual Assignment	Rel.	Ord.	Ext.	Multi F_1	No-target	Avg.
Entity → Scene → Relation	35.92	24.63	38.41	84.72	86.67	54.07
Entity → Relation → Scene	34.11	22.58	36.47	82.96	84.93	52.21
Scene → Entity → Relation	40.82	30.17	43.58	91.74	93.04	**59.87**

Table 5: Stage-wise semantic reasoning ablation on RefScene-20K.

Variant	r_s	r_e	r_r	Rel.	Ord.	Ext.	Multi F_1	No-target	Avg.
Full RefScene-T	✓	✓	✓	40.82	30.17	43.58	91.74	93.04	**59.87**
w/o Scene Reasoning	×	✓	✓	39.61	29.42	42.36	88.41	90.12	57.98
w/o Entity Reasoning	✓	×	✓	37.84	27.63	40.12	89.26	90.31	57.03
w/o Relation Reasoning	✓	✓	×	34.76	22.18	35.91	90.48	91.26	54.92

Table 3 demonstrates that hierarchical semantic decomposition is more effective than applying the full expression uniformly across stages. Flat expression guidance achieves only 54.67 Avg., 5.20 points below the complete hierarchy. Removing relation reasoning reduces Avg. to 54.92 and substantially degrades relational, ordinal, and extremal grounding, showing the importance of relation-level semantics for candidate disambiguation. Scene and entity reasoning provide complementary benefits by guiding visual scope and candidate construction, respectively.

Table 4 demonstrates that semantic roles and visual operations are not interchangeable. The proposed Scene→Entity→Relation assignment reaches 59.87 Avg., outperforming alternative orderings by 5.80 and 7.66 points, respectively. Scene reasoning establishes the relevant visual context, entity reasoning constructs the candidate state, and relation reasoning performs final disambiguation. Reordering these stages disrupts the progressive information flow and reduces both target-level and set-level performance.

The stage-wise ablation in Table 5 directly evaluates the contribution of each semantic state. Removing relation reasoning causes the largest degradation, reducing Avg. from 59.87 to 54.92, while removing entity and scene reasoning decreases it to 57.03 and 57.98, respectively. Relation reasoning is especially important for relational, ordinal, and extremal expressions, whereas scene reasoning mainly affects scene-dependent and set-level predictions. These results verify that the three semantic roles provide complementary guidance within the unified grounding process.

Table 6: Stage-specific semantic guidance ablation on RefScene-20K. In each flat-guidance variant, the designated stage receives the pooled full-expression representation T instead of its specialized reasoning state.

Variant	SGE input	EGR input	RGS input	Rel.	Ord.	Ext.	Multi F_1	No-target	Avg.
Flat guidance at SGE	T	R_e	R_r	39.48	29.02	42.17	90.68	91.24	58.52
Flat guidance at EGR	R_s	T	R_r	38.16	27.85	40.96	90.12	90.74	57.57
Flat guidance at RGS	R_s	R_e	T	35.26	23.17	36.88	90.53	91.01	55.37
HRVG-T	R_s	R_e	R_r	40.82	30.17	43.58	91.74	93.04	**59.87**

6 Conclusion

We presented HRVG, a unified framework that views complex aerial visual grounding as a progressive process where different semantic roles guide corresponding visual computations. Rather than treating reasoning as an auxiliary language representation, HRVG integrates hierarchical semantic guidance with continuous visual state evolution, enabling more coherent grounding under complex contextual and relational constraints. Together with RefScene-20K, which provides a unified evaluation setting for relational, ordinal, extremal, multi-target, and no-target expressions, our work highlights the importance of aligning semantic roles with visual computation for challenging UAV grounding scenarios.

LIMITATIONS

HRVG relies on language-derived semantic states to organize visual grounding. Although these states provide structured guidance, errors in semantic extraction may propagate to subsequent visual computations. In addition, the extra language inference step introduces computational overhead compared with purely end-to-end grounding models. Future work will explore more reliable semantic verification and more efficient joint vision–language inference.

ETHICS AND DATA STATEMENT

RefScene-20K is derived from VisDrone2019-DET and preserves source provenance for all images. The release provides derived referring expressions, target annotations, split information, and annotation-auditing metadata under terms compatible with the original data license; source images are not redistributed beyond the permissions of the original provider. The benchmark contains object-level categories and bounding boxes only, and does not include identity annotations or personal identification tasks. Human review was conducted solely for annotation quality verification. As with other aerial vision benchmarks, models trained on RefScene-20K should be deployed responsibly, considering privacy protection, applicable regulations, and potential misuse in real-world applications.

REPRODUCIBILITY STATEMENT

The main text specifies the model components, evaluation metrics, training resolution, optimizer, and training schedule. The appendix provides the fixed language-model prompt and output schema, validation and fallback procedure, geometric annotation rules, dataset split counts, model equations, loss weights, and efficiency protocol. The same frozen semantic parser and prompt are used for all splits, and its validated outputs are cached by expression identifier. Subject to the VisDrone source-data license, we will release the RefScene-20K annotations and split manifests together with the training and evaluation code, model configurations, cached semantic states, and checkpoints**.We will release our code after the paper is accepted.**

AI USE STATEMENT

The authors used large language models for assisting manuscript preparation, including language polishing, grammar correction, and improving presentation clarity. The authors remain fully responsible for the scientific content, including the research ideas, methodology, dataset construction, experiments, and conclusions. Qwen3-VL-32B-Instruct generated candidate scene descriptions and referring expressions during RefScene-20K construction, and InternVL3.5-38B provided an independent model-based quality assessment.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *arXiv preprint arXiv:2409.17531*, 2024.

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.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized referring expression comprehension, 2023. URL https://arxiv.org/abs/2308.16182. Technical report.

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.

Meng Lan, Fu Rong, Hongzan Jiao, Zhi Gao, and Lefei Zhang. Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024. URL https://ieeexplore.ieee.org/document/10542207.

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.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601. IEEE, 2023.

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*, 2025. URL https://arxiv.org/abs/2504.07836.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2303.05499.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

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. URL https://arxiv.org/abs/2209.13959.

Zhichao Sun, Yepeng Liu, Zhiling Su, Huachao Zhu, Yuliang Gu, Yuda Zou, Zelong Liu, Gui-Song Xia, Bo Du, and Yongchao Xu. Refdrone: A challenging benchmark for referring expression comprehension in drone scenes. *arXiv preprint arXiv:2502.00392*, 2025. URL https://arxiv.org/abs/2502.00392.

Chaowei Wang, Yan Di, Jingjun Sun, Baozhe Liu, Jiaxu Tian, Yuheng Li, Guangqian Guo, and Shan Gao. Grabvg: Graph-attentive binding for visual grounding in uav imagery. *arXiv preprint arXiv:2608.18996*, 2026. URL https://arxiv.org/abs/2608.18996.

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. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33577–33587, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_Grounded_Chain-of-Thought_for_Multimodal_Large_Language_Models_CVPR_2026_paper.html.

Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. Remotesam: Towards segment anything for earth observation. *arXiv preprint arXiv:2505.18022*, 2025. URL https://arxiv.org/abs/2505.18022.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular attention network for referring expression comprehension. *arXiv preprint arXiv:1801.08186*, 2018. URL https://arxiv.org/abs/1801.08186.

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*, 2023. doi: 10.1109/TGRS.2023.3250471. URL https://arxiv.org/abs/2210.12634.

Yue Zhou, Jue Chen, Zilun Zhang, Penghui Huang, Ran Ding, Zhentao Zou, PengFei Gao, Yuchen Wei, Ke Li, Xue Yang, Xue Jiang, Hongxin Yang, and Jonathan Li. Dvgbench: Implicit-to-explicit visual grounding benchmark in uav imagery with large vision-language models. *arXiv preprint arXiv:2601.00998*, 2026. URL https://arxiv.org/abs/2601.00998.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021.

A Evaluation on Additional UAV Grounding Benchmarks

We further assess HRVG on two complementary UAV grounding benchmarks. Table 7 reports results on AerialVG, and Table 8 reports results on RefDrone. Models are trained and evaluated using the official splits of the corresponding benchmark. These experiments evaluate the applicability of HRVG beyond RefScene-20K.

A.1 Evaluation on AerialVG

Table 7 reports Top-1 and Top-5 localization accuracy for representative grounding methods on AerialVG. Zero-shot results are included when available, while the supervised results compare models trained on the official AerialVG training split.

Table 7: Results on AerialVG. Top-1 and Top-5 denote localization accuracy; a dash indicates that the corresponding result was not reported.

Method	Zero-shot Top-1	Zero-shot Top-5	Fine-tuning Top-1	Fine-tuning Top-5
TransVG (Deng et al., 2021)	2.57	3.65	11.53	13.68
Dynamic MDETR (Shi et al., 2024)	2.61	4.63	19.87	29.87
SimVG (Dai et al., 2024)	3.61	4.68	20.94	31.32
Grounding DINO (Liu et al., 2024)	**11.10**	**13.79**	29.36	78.87
AerialVG (Liu et al., 2025)	–	–	50.01	87.00
GrabVG (Wang et al., 2026)	–	–	67.31	89.43
HRVG-T (ours)	–	–	67.42	91.37
HRVG-B (ours)	–	–	**69.28**	**92.14**

Under supervised training on the AerialVG split, HRVG-B achieves 69.28 Top-1 and 92.14 Top-5 accuracy. It exceeds GrabVG by 1.97 and 2.71 points and AerialVG by 19.27 and 5.14 points, respectively. HRVG-T obtains 57.42 Top-1 and 91.37 Top-5. These results indicate that hierarchical semantic guidance improves the ranking of the most relevant candidate while preserving strong candidate recall. HRVG is not evaluated in the zero-shot setting; this experiment therefore supports supervised applicability on AerialVG.

A.2 Evaluation on RefDrone

Table 8 evaluates HRVG on RefDrone using instance-level and image-level F_1 and accuracy. The instance-level metrics measure object-wise grounding quality, whereas the image-level metrics assess the predicted target set for each image.

Table 8: Results on RefDrone. $F_{1,\mathrm{inst}}$ and $\mathrm{Acc}_{\mathrm{inst}}$ are instance-level metrics; $F_{1,\mathrm{img}}$ and $\mathrm{Acc}_{\mathrm{img}}$ are image-level metrics.

Method	Backbone	$F_{1,\mathrm{inst}}$	$\mathrm{Acc}_{\mathrm{inst}}$	$F_{1,\mathrm{img}}$	$\mathrm{Acc}_{\mathrm{img}}$
GLIP-T (Li et al., 2022)	Swin-T	56.92	40.39	41.31	28.88
Grounding DINO-T (Liu et al., 2024)	Swin-T	67.64	51.55	54.54	39.63
NGDINO-T (Sun et al., 2025)	Swin-T	71.11	55.52	56.51	41.20
Grounding DINO-B (Liu et al., 2024)	Swin-B	69.75	53.95	56.95	41.81
NGDINO-B (Sun et al., 2025)	Swin-B	72.51	57.22	57.84	42.54
HRVG-T (ours)	Swin-T	72.03	55.58	56.60	41.28
HRVG-B (ours)	Swin-B	**73.58**	**58.25**	**58.76**	**42.63**

HRVG-B achieves the best performance across all four RefDrone metrics. Relative to NGDINO-B, it improves instance-level F_1 by 1.07 points, instance-level accuracy by 1.03 points, image-level F_1 by 0.92 points, and image-level accuracy by 0.09 points. HRVG-T likewise improves over NGDINO-T on all four metrics, with gains of 0.92, 0.06, 0.09, and 0.08 points, respectively. These consistent gains show that the candidate-set and relation-aware components remain effective beyond RefScene-20K.

Fig. 4 complements the aggregate metrics with representative RefDrone predictions covering relational, ordinal, extremal, multi-target, and no-target expressions. Compared with Grounding DINO and NGDINO, HRVG more consistently follows relational and ordering constraints, recovers the complete target set for the multi-target query, and suppresses unsupported detections for the no-target query.



Figure 4: Qualitative comparison on RefDrone. Columns show the referring expression, predictions from Grounding DINO, NGDINO, and HRVG, and the ground truth. Rows cover relational, ordinal, extremal, multi-target, and no-target cases. Red boxes denote predictions, green boxes denote annotated targets, and inset crops enlarge the relevant regions. For the no-target expression, an empty prediction is correct.

B Dataset Construction Details

This section summarizes the provenance, visual examples, annotation evidence, and construction pipeline of RefScene-20K. The benchmark covers relational, ordinal, comparison/extremal, multi-target, and no-target expressions. The primary track describes the dominant referential requirement, while target cardinality and annotation evidence are recorded independently.

Fig. 5 presents representative visual examples from the RefScene-20K construction set. The samples span traffic, pedestrian, and built environments and illustrate no-target, multi-target, relational, comparison, and ordinal referring expressions.

B.1 Preprocessing and Evidence Representation

Before computing geometric relations, we remove incomplete annotations and correct evident bounding-box errors. Objects that are too small or too heavily occluded to support reliable descriptions are excluded, and each retained image contains at least three valid objects. All expressions associated with the same image are assigned to a single split.

Each record stores the final target set separately from its annotation evidence, including reference boxes, comparison sets, ordering information, and deterministic geometric measurements. This separation makes each target assignment auditable without exposing annotation evidence as model

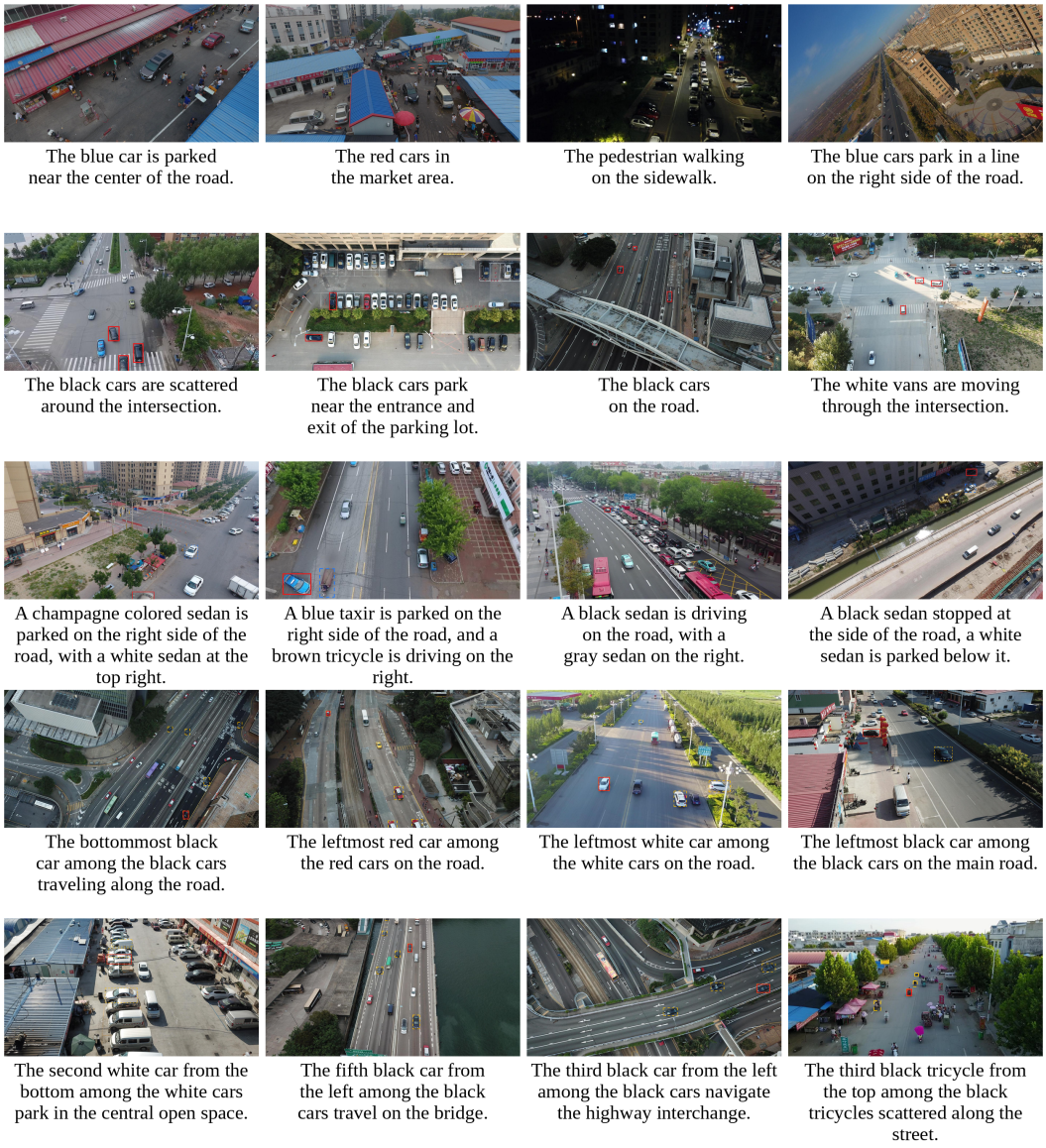

Figure 5: Representative RefScene-20K examples. Each panel shows an aerial scene and its associated referring expression. Red solid boxes mark target objects, blue dashed boxes mark relational reference objects, and yellow dashed boxes mark candidate objects used for comparison or ordinal reasoning. From top to bottom, the rows correspond to no-target, multi-target, relational, comparison, and ordinal expressions.

input. At evaluation time, models return a possibly empty set of boxes. A single-target prediction is correct when its IoU with the ground-truth box is at least 0.5; multi-target predictions are evaluated as sets, and no-target queries are correct only when the predicted set is empty.

B.2 Design Taxonomy

Table 9 summarizes how representative aerial grounding designs structure visual search and which capabilities are explicit in their models or evaluation protocols.

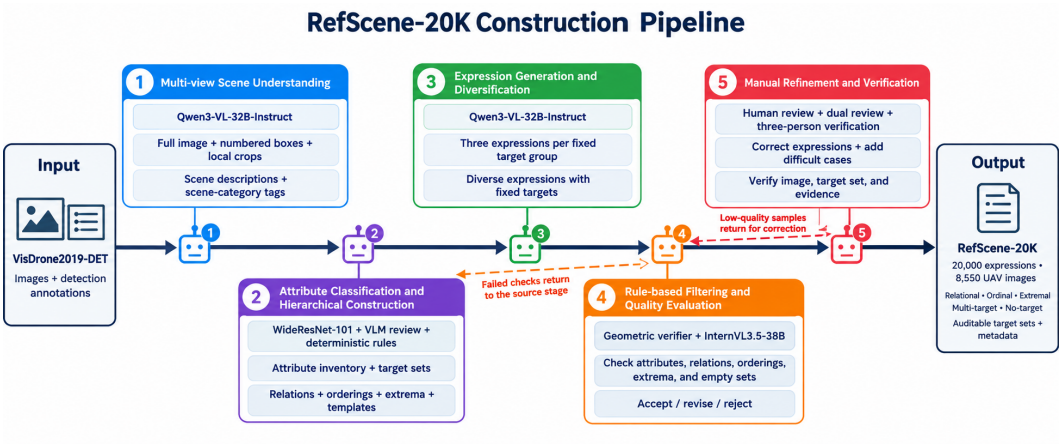

Figure 6: Five-stage construction pipeline for RefScene-20K. VisDrone images and detection annotations are processed through (1) multi-view scene understanding, (2) attribute classification and hierarchical construction, (3) expression generation and diversification, (4) rule-based filtering and independent-model quality evaluation, and (5) manual refinement and verification. Feedback paths route low-quality expressions and inconsistent attributes or templates to earlier stages before final export.

Table 9: Qualitative taxonomy of representative aerial grounding designs. *NR* indicates that a capability is not an explicit mechanism or evaluation focus in the cited description; it does not imply that the model cannot represent that capability. Target cardinality and a candidate's relation support count are distinct quantities.

Design	Scene / visual focus	Candidate structure	Relations	Count / order	Empty set
RSVG/MLCM	Multiscale alignment	Region grounding	*NR*	*NR*	*NR*
AerialVG	Hierarchical attention	Target / auxiliary	Positional	*NR*	*NR*
RefDrone/NGDINO	Small aerial objects	Detection queries	Context in text	Answer count	Explicit
GrabVG	Hypothesis filtering	Sparse graph	Topological	*NR*	*NR*
HRVG (ours)	Scene-conditioned	All-entity pool	Directed context	Relation mass	Threshold

B.3 Five-Stage Annotation Pipeline

Fig. 6 summarizes the closed-loop construction process. Starting from VisDrone images and detection annotations, the pipeline combines multi-view scene understanding, hierarchical attribute and template construction, expression diversification, rule- and model-based filtering, and manual refinement. Feedback paths return low-quality expressions and template or attribute inconsistencies to earlier stages for correction. In the figure, *instruction* denotes a referring expression together with its target set and associated metadata.

Stage 1: Multi-view scene understanding. Qwen3-VL-32B-Instruct (Bai et al., 2025) receives the full image, numbered source boxes, and local crops. It produces descriptions of the visible environment, spatial layout, and object interactions, together with scene-category tags. The released annotations retain only visually verifiable descriptions and tags, excluding unsupported identity claims, off-image knowledge, and free-form model rationales.

Stage 2: Attribute classification and hierarchical construction. A WideResNet-101 color classifier proposes initial attributes, and a vision–language model reviews cases affected by illumination or occlusion. A deterministic rule engine then forms target groups, constructs directed spatial re-

lations, sorts box centers for ordinal labels under a fixed tie policy, and computes extrema from position, area, or distance to a reference. This stage produces a fixed attribute inventory, target sets, and hierarchical templates. A no-target query is permitted only when the completeness of the source annotations can be established.

Stage 3: Expression generation and diversification. Three differently worded expressions are generated for each fixed target group using Qwen3-VL-32B-Instruct. Synonym replacement and varied phrasing increase linguistic diversity while preserving the underlying constraints. Every expression must provide the minimum sufficient conditions for identifying its target group. A changed color, altered cardinality, reversed relation, or ambiguous description triggers regeneration or attribute review rather than an ad hoc change to the target geometry.

Stage 4: Rule-based filtering and quality evaluation. A geometric verifier recomputes target cardinality, ordering, extrema, and empty-set conditions and filters invalid spatial relations or inconsistent attributes. Independently, InternVL3.5-38B evaluates completeness, exclusivity, scene fidelity, and linguistic naturalness. Each candidate is accepted, revised, or rejected, and a failed check returns the sample to the stage that produced the affected field.

Stage 5: Manual refinement and extension. Human reviewers inspect the image, expression, target set, and key evidence, correct low-quality expressions, and add difficult cases where needed. Focused review covers high ordinal ranks, dense candidate sets, tiny or occluded targets, difficult negatives, and disagreements between the two automatic verifiers. Dual review and three-person verification are used for final quality control. Annotation metadata distinguish automatic acceptance, human acceptance, pending review, and rejection.

B.4 Generation Prompt and Geometric Conventions

Expression generation uses the same instruction template for all images. The model receives a scene summary, numbered object boxes, verified attributes, a fixed target set, and the relation records produced in Stage 2. The generation instruction is:

> Write one natural referring expression whose complete set of answers is exactly the supplied target identifiers. Use only the supplied object attributes and relations. The expression must be understandable without object identifiers, must contain the minimum conditions needed to exclude every non-target candidate, and must not mention coordinates, bounding boxes, or annotation metadata. Preserve the supplied target cardinality and relation direction. Return JSON with `expression`, `target_ids`, `reference_ids`, `relation`, and `cardinality`.

Three candidates are sampled for each target set with temperatures 0.3, 0.6, and 0.8. The candidates are subsequently checked against the same deterministic geometry and target set; sampling diversity therefore changes wording but not annotation semantics.

All geometric rules operate on normalized box centers and sizes. A candidate is strictly left or right of a reference when the horizontal center difference is greater than 0.02 image widths, and strictly above or below it when the vertical difference is greater than 0.02 image heights. Two boxes are *near* when their center distance divided by the image diagonal is at most 0.12 and *far* when it is at least 0.30. Overlap requires IoU ≥ 0.10. Ordinal candidates are sorted by center coordinate along the requested axis; centers within 0.01 image widths or heights are tied and are resolved first by the orthogonal coordinate and then by the immutable source annotation index. Extrema use the relevant center coordinate, box area, or reference distance and are accepted only when the best candidate differs from the runner-up by at least 0.02 of the corresponding image dimension or 5% of the scalar value. No-target expressions are created by changing one verified attribute, relation, rank, or cardinality and are retained only when exhaustive rule evaluation over all annotated candidates returns the empty set.

Table 10: Image and expression counts in the RefScene-20K release. Images are disjoint across splits, and each expression is assigned one primary track.

Split	Images	Relational	Ordinal	Comparison/extremal	Multi-target	No-target	Total
Train	5,985	3,500	3,500	3,500	2,060	1,400	13,960
Validation	1,280	750	750	750	450	300	3,000
Test	1,285	750	750	750	490	300	3,040
Total	8,550	5,000	5,000	5,000	3,000	2,000	20,000

B.5 Split Statistics and Quality Audit

The final split is performed at the source-sequence level, and all expressions associated with the same image remain in the same split. Tracks are stratified independently after sequence grouping. Table 10 gives the frozen release counts.

Preprocessing and quality control yield 8,550 eligible source images after box correction and visibility filtering. Expression generation produces 31,842 candidates. Deterministic geometry, cardinality, and attribute checks retain 27,916; independent vision–language assessment retains 23,604 for human review; and semantic deduplication plus adjudication yields the final 20,000.

Every machine-filtered candidate is independently inspected by two reviewers. They agree on the complete target set for 96.2% of candidates and on the primary track for 94.8%, corresponding to Cohen's $\kappa = 0.92$ over the five track labels. A third reviewer adjudicates all 1,684 cases with any disagreement. The final adjudication rejects 2,147 semantically incorrect or ambiguous expressions and removes 1,457 paraphrastic duplicates. A final automatic replay of the geometric rules confirms the stored target set for all 20,000 records.

C Additional Method Details

C.1 Chain-of-Thought Reasoning and Extraction Protocol

The frozen semantic reasoning model uses Qwen3-8B-Instruct with greedy decoding at temperature 0, top-$p = 1$, and a maximum output length of 192 tokens. Each expression is processed independently with the following fixed instruction:

> Reason step by step about the aerial referring expression. First identify only the scene or search-region cues explicitly supported by the expression. Next identify the target and supporting entities and their attributes. Finally, analyze the spatial, ordinal, extremal, cardinality, or empty-set constraints needed to select the answer. Keep this reasoning concise and do not infer image content that is not stated in the expression. After the reasoning, write the marker `FINAL_JSON:` followed by exactly one JSON object using the schema {`"scene": string, "entities": [string], "constraint": {"type": string, "relation": string, "direction": string, "ordinal": integer, "cardinality": integer, "allow_empty": boolean`}}. Include only conditions supported by the expression. Use an empty string or -1 when a field is not specified; do not invent visible objects or image content.

The user message contains only the original expression. We retain the generated reasoning response for auditing, locate the block following `FINAL_JSON:`, and expose only its validated fields to the visual model. The scene string summarizes an explicitly mentioned environment or local search region. The entity list preserves the surface forms of target and supporting-object phrases so that each item can be mapped to a contiguous or punctuation-separated token span. The constraint object records the active relation family and its arguments. Allowed types are `none`, `spatial`, `ordinal`, `extremal`, `multi`, and `no-target`; multiple simultaneous conditions are serialized in their textual order using semicolon-separated relation strings.

We parse the extracted block with a strict JSON decoder and verify field types, token alignment, non-negative specified cardinalities, and consistency between the type and its required arguments. An invalid extraction receives one deterministic repair instruction containing the schema error but no image information. If it remains invalid, the scene field is set to a neutral global context, the complete expression is retained as the entity phrase, and the constraint field is set to unrestricted

selection. The model is frozen and receives no benchmark labels, images, target boxes, or split identifiers. Cached records contain the original expression, the generated reasoning response, the three validated fields, aligned entity-token spans, model version, and a SHA-256 hash of the prompt. This record makes the text-side inputs reproducible and prevents prompt changes between training and evaluation.

The validated strings are encoded by the Grounding DINO text encoder. Mean pooling over valid scene tokens gives the scene query q_s; entity phrases and their token spans define R_e and the sets M_k used for candidate alignment; and mean pooling over constraint tokens gives the constraint query q_c associated with R_r. These states are generated before visual-model training, stored with the expression identifier, and reused in all epochs. The same model, prompt, schema, validation rules, and frozen parameters are used for training, validation, and testing.

C.2 Heatmap-Conditioned Deformable Attention

For detector query index q, attention head h, feature scale (pyramid level) l, and sampling-point index p, let $\boldsymbol{q}_{qh}$ denote the visual query vector for query q and head h, a^0_{qhlp} the pretrained attention logit, u_{qhlp} the sampling location, and v_{qhlp} the visual feature sampled at that location. The scene heatmap H_l is the scene prior defined in the main text and is evaluated at u_{qhlp}; ϕ maps this response to the gating input. SGE updates the attention logit as

$$a_{qhlp} = a^0_{qhlp} + \tanh(\alpha)\, f_{\mathrm{HCDA}}(\boldsymbol{q}_{qh}, v_{qhlp}, H_l(u_{qhlp}))\,, \tag{11}$$

where f_{HCDA} is a learnable low-rank query–value interaction that returns a scalar correction conditioned on $H_l(u_{qhlp})$. The learnable scalar α, initialized to zero, controls the residual strength, and $\tanh(\cdot)$ bounds its effective scale. We parameterize this heatmap-conditioned correction, denoted Δa_{qhlp}, as

$$\Delta a_{qhlp} = \frac{1}{\sqrt{r}} \sum_{m=1}^{r} (U_h^\top \boldsymbol{q}_{qh})_m \, g_m(\phi(H_l(u_{qhlp}))) \; (V_h^\top v_{qhlp})_m. \tag{12}$$

Here, r is the residual projection dimension and m indexes its latent channels. U_h and V_h are learnable, head-specific projections for the query and visual feature, respectively, while $g_m(\cdot)$ is the learnable scalar gate for channel m. Thus, Δa_{qhlp} is a scene-conditioned residual rather than a replacement for the pretrained attention logit. In Eq. 11, we set $f_{\mathrm{HCDA}}(\boldsymbol{q}_{qh}, v_{qhlp}, H_l(u_{qhlp})) = \Delta a_{qhlp}$. The bounded, zero-initialized gate preserves the pretrained behavior at initialization. The residual changes attention weights without changing the pretrained sampling coordinates or hard-masking any location. When an expression contains no explicit local scene condition, the module uses a neutral scene response.

C.3 Entity-guided Candidate Alignment

After scene-aware encoding, Grounding DINO provides alignment logits z_{it} between detector-query index i and text-token index t. For the K entity phrases identified by R_e, let e_k be the k-th phrase and M_k the set of its token indices in T. Candidate–entity alignment is computed as

$$p_{ik} = \sigma\left(\max_{t \in M_k} z_{it}\right), \qquad \boldsymbol{p}_i = [p_{i1}, \ldots, p_{iK}], \tag{13}$$

where $p_{ik} \in (0, 1)$ is the alignment score between candidate i and entity phrase e_k, and $\boldsymbol{p}_i$ collects the K entity-alignment scores for that candidate. Entity-wise selection retains candidates supported by any entity phrase; subsequent merging and deduplication produce the shared pool used by relation-guided reasoning. This preserves both potential targets and supporting or comparison objects instead of making a final target decision at the entity stage.

C.4 Directed Support Field

For an ordered candidate pair (i, j), i denotes the candidate being scored and j a context, reference, or comparison candidate. Following the main-text convention, r_{ij} denotes the directed support from j to i; this support need not equal r_{ji}. The main-text candidate state x_i is mapped to a learned node embedding g_i. Likewise, q_c is the embedding of the active relation or ordering constraint R_r, and

Table 11: Efficiency at the RefScene-20K evaluation resolution. Online latency uses cached semantic states and includes image preprocessing, visual inference, and final target-set selection.

Method	Trainable params (M)	Frozen visual params (M)	GFLOPs	Peak memory (GB)	Online latency (ms)
Grounding DINO-T	172.3	0.0	282	5.1	76.4
HRVG-T	191.8	96.4	337	6.7	94.8
Grounding DINO-B	250.6	0.0	421	7.2	105.6
HRVG-B	270.1	96.4	476	8.8	124.7

Δb_{ij} is the relative geometric descriptor of context candidate j with respect to scored candidate i, containing center offset, scale ratio, IoU, and direction. We compute

$$z_{ij}^{\mathrm{rel}} = f_{\mathrm{rel}}(g_i, g_j, \Delta b_{ij}, q_c), \tag{14}$$

$$z_{ij}^{\mathrm{ord}} = \widetilde{z}_{ij} - \widetilde{z}_{ji}, \tag{15}$$

$$\boldsymbol{\pi}(q_c) = \mathrm{softmax}(W_r q_c), \tag{16}$$

$$r_{ij} = \sigma(\pi_1 z_{ij}^{\mathrm{rel}} + \pi_2 z_{ij}^{\mathrm{ord}}), \quad r_{ii} = 0. \tag{17}$$

Here, z_{ij}^{rel} is the general relation logit, while $\widetilde{z}_{ij}$ is the preliminary directional-ordering logit before antisymmetrization. Their difference z_{ij}^{ord} therefore changes sign when the pair is reversed. W_r is a learnable mixing matrix and $\boldsymbol{\pi}(q_c) = (\pi_1, \pi_2)$ gives the constraint-dependent mixture weights, which sum to one. σ is the logistic sigmoid, so r_{ij} is a probability-like directed support edge; $r_{ii} = 0$ removes self-support. The unrestricted branch models general relations, whereas the antisymmetric branch provides an ordering bias. Independent sigmoid edges avoid the unit-sum constraint of a neighbor softmax and retain approximate count information. In the main-text notation, the candidate-level soft support count is $n_i = \sum_{j \neq i} r_{ij}$. When an auxiliary constraint-level count prior is used, $\widehat{n}_c = \mathrm{Softplus}(\mathrm{MLP}(q_c))$ is its predicted expected count; it is distinct from the candidate-level quantity n_i.

For completeness, $\psi(\cdot)$ is a learnable embedding of a candidate representation and $\epsilon > 0$ prevents division by zero in the normalized support summary

$$m_i = \frac{\sum_{j \neq i} r_{ij} \psi(x_j)}{\sum_{j \neq i} r_{ij} + \epsilon}, \qquad n_i = \sum_{j \neq i} r_{ij}. \tag{18}$$

The vector m_i captures which candidates support i, whereas n_i is their soft support count. Candidate consistency is scored by

$$E_i = -f_{\mathrm{energy}}(x_i, m_i, n_i, R_r), \qquad \widehat{\mathcal{Y}} = \{b_i \mid x_i \in X_e,\ E_i \leq \tau_E\}. \tag{19}$$

Here, f_{energy} is a learnable scalar scoring function, and the validation-selected threshold τ_E retains exactly the boxes whose energies indicate sufficient consistency. These summaries and energies constitute H_r and implement the output mapping $\mathcal{G}$.

D Additional Experiments

D.1 Additional Ablation Results

The following experiment complements the core ablations in the main paper by replacing one stage-specific state at a time with the pooled expression.

D.2 Evaluation Reporting and Qualitative Analysis

The text-side reasoning and extraction process runs once per expression and is cached, whereas the visual grounding modules execute online. We measure efficiency on a single NVIDIA GeForce RTX 4090 with batch size one, FP16 visual inference, and the same 800-pixel short side and 1,333-pixel long-side cap used in evaluation. Latency is averaged over 500 expressions after 50 warm-up iterations; peak memory is measured with CUDA synchronization. FLOPs include the online visual path but exclude the cached language model. Table 11 reports the mean of three runs.

Relative to its Grounding DINO counterpart, HRVG adds 19.5M trainable parameters and 55 GFLOPs. Its online latency increases by 18.4 ms for Swin-T and 19.1 ms for Swin-B, while retaining

throughput of 10.55 and 8.02 expressions per second, respectively. The frozen Qwen3-8B-Instruct parser contains 8.2B parameters and, when executed in BF16 on the same GPU, uses 16.9 GB peak memory and averages 1.18 s per expression. Sequential uncached end-to-end latency is therefore 1.27 s for HRVG-T and 1.30 s for HRVG-B. In the standard evaluation protocol, parsing is performed once and its approximately 1.8-KB structured record is cached, so repeated training epochs and online visual evaluation incur only the latencies reported in Table 11.

Fig. 3 and 4 provide complementary qualitative evidence. The former traces scene response, entity retrieval, and constraint-based selection across multi-target, ordinal, relational, and extremal queries; the latter compares final predictions on RefDrone and includes a no-target case. Together, they cover all five RefScene-20K tracks and illustrate intermediate evidence propagation and behavior on an additional UAV grounding benchmark.