REASONING-DETR: CONTEXTUAL AND RELATIONAL QUERY LEARNING FOR AERIAL DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Object detection has progressed rapidly from two stage and one stage frameworks to Transformer detectors. Among them, DETR formulates detection as set prediction and allows object queries to exchange contextual information through attention. However, its direct application to aerial imagery remains limited. Tiny objects are easily suppressed by complex backgrounds, while their number and spatial distribution vary considerably across scenes. A fixed query budget therefore cannot accommodate different search demands, and interactions based on feature similarity may introduce unreliable background or neighboring evidence into ambiguous queries. To address these issues, we propose **Reasoning-DETR**, which couples adaptive object search with query reasoning based on uncertainty. First, a dynamic query allocator guided by scene context combines regional density prediction with Scene Memory to determine the number and spatial distribution of supplementary queries. Second, the query reasoner retrieves category prototypes for ambiguous queries and selectively aggregates evidence from reliable nearby objects. A reliability gate suppresses relational messages that conflict with the current query representation. Both modules are integrated into the original RT-DETR pipeline without requiring additional scene or relation annotations. Experiments on VisDrone and AI-TOD-V2 demonstrate consistent improvements in aerial and tiny-object detection, with improved query coverage and recognition of weak objects at a modest computational overhead.

1 INTRODUCTION

Object detection aims to localize and categorize objects, forming a fundamental perception capability for autonomous driving, surveillance, robotics, and UAV systems. From two-stage frameworks such as Faster R-CNN to one-stage and Transformer-based detectors, substantial progress has been made in generic object detection (Ren et al., 2015; Wang et al., 2024a; Zhao et al., 2024; Peng et al., 2025). Among them, RT-DETR combines an efficient hybrid encoder with high-quality query initialization and achieves strong end-to-end performance against YOLO detectors (Zhao et al., 2024). Nevertheless, progress on generic benchmarks does not transfer directly to aerial imagery, where objects are tiny, densely clustered, occluded, and unevenly distributed against complex backgrounds (Li et al., 2023; Xu et al., 2023; Cai et al., 2024; Hu et al., 2025; Ji et al., 2026). These characteristics weaken object evidence and make aerial detection substantially more challenging.

Existing aerial object detection methods mainly follow three directions. **(I) Region-based detection** identifies object-dense regions and processes them at a finer resolution to reduce the influence of large background areas (Huang et al., 2022; Ge et al., 2022; Xu et al., 2022c; Meethal et al., 2023). **(II) Feature enhancement** employs additional prediction heads, attention mechanisms, multi-scale fusion, or enlarged receptive fields to preserve the visual evidence of small objects (Zhu et al., 2021a; Zhao et al., 2023; Li et al., 2023; Cai et al., 2024; Sun et al., 2025). More recently, **(III) End-to-end aerial detection** adapts DETR-like architectures to aerial imagery through density estimation, dynamic query generation, frequency enhancement, and feature-query interaction (Huang et al., 2024a; Hu et al., 2025; Zhang et al., 2025; Ji et al., 2026). Despite this progress, existing methods primarily discover and recognize objects from visual features in the current image. They make limited use of scene-dependent occurrence patterns and inter-object relations, which are particularly valuable when small objects provide weak or ambiguous evidence.

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.

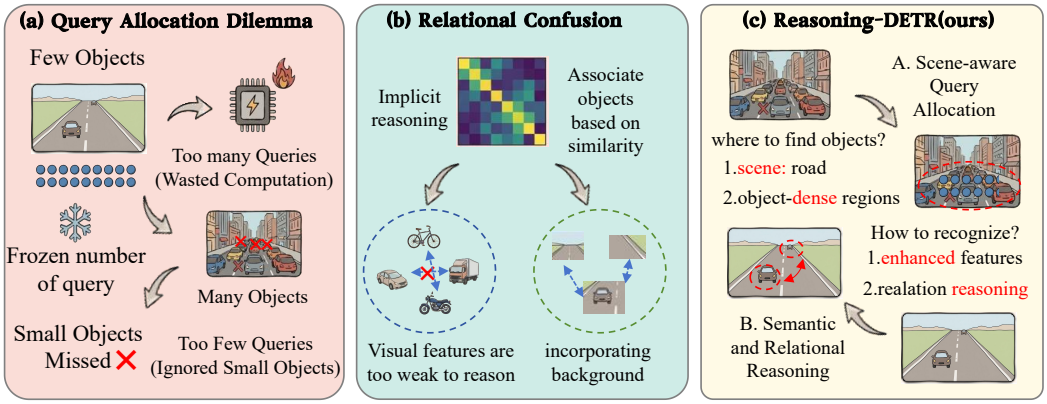

Figure 1: **Motivation of Reasoning-DETR.** (a) A fixed query budget under-serves crowded scenes while wasting queries in sparse ones. (b) Similarity driven interaction can mix weak object queries with confusing categories or background features. (c) Reasoning-DETR couples scene- and density-aware search with semantic and reliable-neighbor reasoning.

DETR formulates object detection as end-to-end set prediction with a Transformer decoder and object queries (Carion et al., 2020; Zhu et al., 2021b). Within the decoder, self-attention enables queries to exchange object-level information (Chen et al., 2023), while cross-attention retrieves visual evidence through content and positional interactions (Meng et al., 2021; Liu et al., 2022). This architecture provides a natural basis for modeling scene context and inter-object dependencies, and RT-DETR further improves its efficiency through multi-scale encoding and high-quality query initialization (Zhao et al., 2024). However, aerial imagery presents substantial variations in object number and spatial distribution, which cannot be adequately accommodated by a fixed query budget (Huang et al., 2024a; Hu et al., 2025). More critically, small objects often produce weak and semantically ambiguous query representations. During unrestricted interaction, these queries may attend to visually similar background responses or receive messages from unreliable neighboring queries, thereby amplifying rather than resolving their recognition ambiguity. As illustrated in Fig. 1, these limitations lead to the pivotal research question that motivates our work:

How can an aerial detector reason about where to search for potential objects and which semantic and contextual evidence is reliable for recognition?

To answer this question, we introduce **Reasoning-DETR**, which organizes aerial detection as a coupled process of adaptive search and evidence-grounded recognition. It comprises two synergistic components: a **scene-aware query allocator**, which combines regional density evidence with learned scene patterns to adapt the number and spatial distribution of supplementary queries; and a **semantic and relational query reasoner**, which identifies uncertain queries, retrieves their plausible category prototypes, and selectively aggregates evidence from reliable nearby queries. The former determines where additional search capacity is required, whereas the latter separates category-level ambiguity from relation-level evidence aggregation, preventing weak candidates from indiscriminately absorbing contextual information. The refined queries subsequently interact with multi-scale image features in the final decoder layer, allowing the updated representations to be further refined using current visual features. Reasoning-DETR thus couples adaptive query discovery with uncertainty-aware refinement while preserving the end-to-end detection pipeline.

Our contributions are threefold. (1) **Scene-aware query allocation.** We propose a dynamic query allocator that combines regional density with Scene Memory. Unlike the fixed query budget in RT-DETR, it adapts both the number and spatial distribution of supplementary queries to each aerial scene. (2) **Uncertainty-aware query reasoning.** We introduce category prototypes to clarify ambiguous queries and selectively propagate evidence from reliable local neighbors. Unlike implicit query interaction in DETR, the proposed reasoner explicitly controls what relational evidence is used. Both processes are learned end to end without additional semantic or relation annotations. (3) **Comprehensive evaluation.** Experiments on VisDrone and AI-TOD-V2 demonstrate consistent improvements in aerial and tiny-object detection with competitive efficiency, while further analyses examine the complementary contributions of the two modules.

2 Related Work

Aerial object detection. Existing aerial detectors mainly follow three directions. **(I) Region-based detection** identifies object-intensive regions for fine-scale processing. ClusDet generates and normalizes cluster proposals (Yang et al., 2019), DMNet uses density maps to guide regional cropping (Li et al., 2020), and UFPMP-Det packs foreground regions into compact proxy images (Huang et al., 2022). **(II) Feature enhancement** strengthens tiny-object representations through contextual and multi-scale modeling, including global context-enhanced sparse convolution in CEASC (Du et al., 2023), adaptive large kernels in LSKNet (Li et al., 2023), dynamic coarse-to-fine assignment in DCFL (Xu et al., 2023), and multi-scale kernel integration in PKINet (Cai et al., 2024). **(III) End-to-end aerial detection** adapts Transformer detectors to dense tiny objects. DQ-DETR generates dynamic queries from predicted counts and density maps (Huang et al., 2024a), while Dome-DETR jointly adjusts density-oriented features and queries (Hu et al., 2025). Despite their effectiveness, scene-dependent occurrence patterns and semantic object relations remain insufficiently explored.

DETR and query learning. DETR-related studies can be grouped according to how object queries are represented, learned, and initialized. **(I) Spatial query formulation** improves the interaction between queries and image features. Deformable DETR restricts cross-attention to sparse sampling locations (Zhu et al., 2021b), Conditional DETR separates content and spatial attention (Meng et al., 2021), and DAB-DETR represents positional queries as dynamically refined anchor boxes (Liu et al., 2022). **(II) Query optimization** focuses on stabilizing set prediction. DN-DETR introduces denoising queries (Li et al., 2022a), DINO combines contrastive denoising with mixed query selection (Zhang et al., 2023), and Group DETR employs group-wise one-to-many assignment during training (Chen et al., 2023). **(III) Image-adaptive query initialization** selects query content and reference boxes from encoder features. RT-DETR uses uncertainty-minimal encoder features to initialize a fixed number of high-quality queries (Zhao et al., 2024), while DQ-DETR and Dome-DETR adapt the query budget using predicted object density (Huang et al., 2024a; Hu et al., 2025). In contrast, Reasoning-DETR combines current regional density with accumulated scene experience and subsequently evaluates the reliability of relations used to refine uncertain queries.

Contextual and relational reasoning for detection. Context-aware detection mainly follows three directions. **(I) Visual relation modeling** captures dependencies among proposals or features through appearance, geometry, or global interactions (Hu et al., 2018; Xu et al., 2019). **(II) Structured context modeling** organizes evidence using object graphs or memory banks, as exemplified by Graph R-CNN (Yang et al., 2018) and Context R-CNN (Beery et al., 2020). **(III) Semantic representation learning** aligns object features with category-level concepts to improve recognition, including region-language approaches such as GLIP (Li et al., 2022b) and Grounding DINO (Liu et al., 2024). Unlike methods that propagate context globally or rely on external language inference, Reasoning-DETR uses the detector's own category prototypes to refine uncertain queries and restricts relation propagation to reliable spatial neighbors.

3 Methodology

3.1 Problem Statement

Given a UAV detection dataset $\mathcal{D} = \{(I_i, Y_i)\}_{i=1}^{N}$, where $Y_i = \{(b_{ij}, c_{ij})\}_{j=1}^{n_i}$ contains the bounding boxes and categories of objects in image I_i, we aim to improve the coverage and recognition of small objects without additional scene or relation annotations. We adopt RT-DETR (Zhao et al., 2024) as the base detector because it combines end-to-end set prediction with efficient multi-scale feature interaction, providing a favorable balance between accuracy and inference efficiency.

Overview. As illustrated in Fig. 2, our framework consists of two components. Scene-aware dynamic query allocation combines regional density prediction with Scene Memory to determine where supplementary queries should be placed and how many are required. Uncertainty-aware query reasoning then retrieves category prototypes for ambiguous supplementary queries and refines them using reliable neighboring queries. The resulting queries are processed by the final RT-DETR decoder layer. The framework is optimized using the standard detection objective together with the regional density loss.

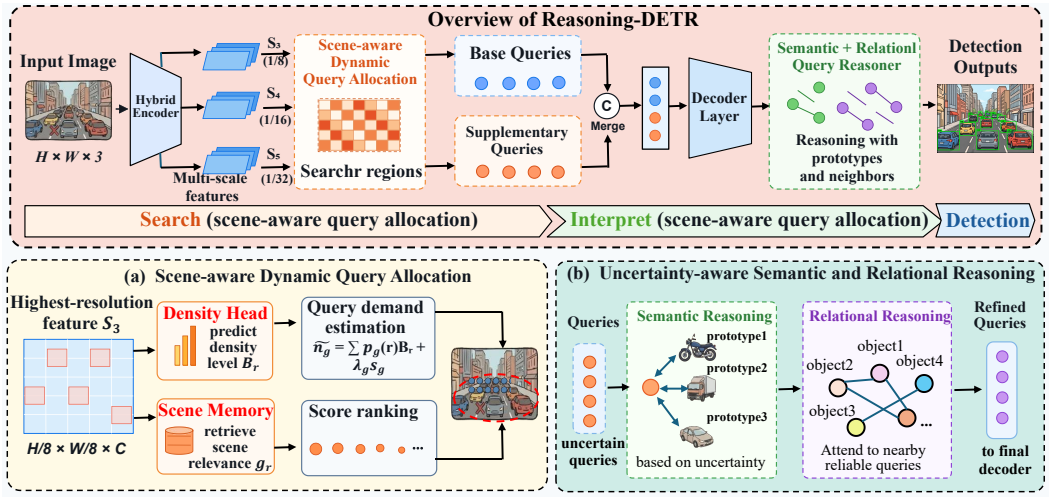

Figure 2: **Overview of Reasoning-DETR.** The upper panel presents the complete detection pipeline, where supplementary queries complement the original RT-DETR queries. (a) Regional density and Scene Memory jointly determine query demand and guide query placement toward crowded, scene-relevant regions. (b) Category prototypes refine uncertain queries, while reliable neighboring queries provide local relational evidence for final detection.

3.2 Scene-Aware Dynamic Query Allocation

RT-DETR initializes decoder queries by globally selecting the Top-K encoder proposals (Zhao et al., 2024). This score-based competition is unfavorable to aerial small objects, whose weak responses can be suppressed by salient objects or background before entering the decoder. To recover overlooked objects without indiscriminately enlarging Top-K, we retain the original queries and introduce a capacity-limited supplementary query pool, which is dynamically allocated across regions according to current object density and learned scene priors.

Overlapping region construction. Let $\{S_3, S_4, S_5\}$ denote the multi-scale features produced by the RT-DETR Hybrid Encoder. All three levels are retained for decoder interaction, while query planning is performed only on the highest-resolution feature S_3. We place M equally sized windows along each spatial dimension of S_3, with the boundary windows aligned to the feature-map borders and adjacent windows overlapping by a ratio ρ. This produces $G = M^2$ regional features, denoted by $\{\mathcal{R}_g\}_{g=1}^{G}$. The pooled representation of each region is used for scene retrieval, whereas its spatial feature is retained for density prediction and query initialization.

Scene memory retrieval. Humans rarely search an entire scene uniformly; instead, they prioritize environments in which the object is more likely to appear. Inspired by contextual memory (Beery et al., 2020) and prototype-based representation (Snell et al., 2017), we construct a Scene Memory that summarizes regional environments frequently associated with objects. For each region containing at least one ground-truth object center, we obtain a scene representation:

$$c_g = \mathrm{Norm}(\mathrm{GAP}(\mathcal{R}_g)), \tag{1}$$

where GAP and Norm denote global average pooling and feature normalization respectively. The collected representations are clustered to initialize K_s scene prototypes $\{M_k\}_{k=1}^{K_s}$. During training, each positive regional representation is assigned to its nearest prototype, which is updated through exponential moving average:

$$k^* = \arg\max_k \cos(c_g, M_k), \qquad M_{k^*} \leftarrow \mathrm{Norm}(\mu M_{k^*} + (1-\mu)c_g), \tag{2}$$

where k^* identifies the most similar memory prototype and μ is the momentum coefficient. The memory is updated after retrieval during training and remains fixed at inference time. Given an input region, we retrieve its L most similar scene prototypes, denoted by $\mathcal{T}_g$, and compute its scene relevance as

$$s_g = \frac{1}{L} \sum_{k \in \mathcal{T}_g} \max(0, \cos(c_g, M_k)). \tag{3}$$

A high s_g indicates that the current region resembles environments in which objects frequently occurred during training, and should therefore receive additional search attention.

Density-aware query planning. While Scene Memory provides a prior on where objects are likely to appear, the number of queries required by a region depends on its current object density. Density-aware aerial detectors have shown that regional object counts can provide useful cues for adapting query construction (Huang et al., 2024a; Hu et al., 2025). We estimate this density using a lightweight prediction head f_d. It applies two 3×3 convolutions to $\mathcal{R}_g$, followed by global average and max pooling, and predicts a distribution over R density levels:

$$\mathbf{p}_g = \mathrm{Softmax}(f_d(\mathcal{R}_g)), \qquad \mathbf{p}_g = [p_g(0), \ldots, p_g(R-1)]. \tag{4}$$

Here, $p_g(r)$ denotes the probability that region $\mathcal{R}_g$ belongs to density level r. Density supervision is derived directly from the detection annotations. We count the ground-truth centers located in each region,

$$N_g = \sum_j \mathbb{I}(\mathrm{center}(b_j) \in \mathcal{R}_g), \tag{5}$$

and map N_g to a density label y_g according to predefined count intervals. Here, b_j is the j-th ground-truth box and $\mathbb{I}(\cdot)$ is the indicator function. The density head is trained with a class-balanced cross-entropy loss:

$$\mathcal{L}_{\mathrm{den}} = -\frac{1}{G} \sum_{g=1}^{G} \alpha_{y_g} \log p_g(y_g), \tag{6}$$

where α_{y_g} compensates for the imbalance between different density levels. We associate each density level r with a monotonically increasing base query budget B_r. The expected query demand of region $\mathcal{R}_g$ is then jointly determined by its density distribution and scene relevance:

$$\widetilde{n}_g = \sum_{r=0}^{R-1} p_g(r) B_r + \lambda_s s_g, \tag{7}$$

where λ_s controls the contribution of Scene Memory. Together, the density estimate and scene relevance determine the supplementary query demand of each region. We then reuse the foreground scores produced by the RT-DETR encoder classification head to initialize these queries, ranking candidate locations within each region. The continuous demands are converted into integer regional budgets and globally constrained by $K_{\max}$, with implementation details provided in the Appendix.

3.3 Uncertainty-Aware Semantic and Relational Reasoning

Supplementary queries improve the coverage of potential objects, but their representations are often dominated by weak foreground evidence and complex backgrounds. Directly propagating features according to visual similarity may further amplify such ambiguity, since an uncertain query can be associated with irrelevant regions or unreliable object queries. We therefore decompose query refinement into two consecutive processes: uncertainty-aware semantic reasoning first identifies plausible category hypotheses, and reliability-guided relational reasoning then supplements the uncertain representations with selected local context.

Uncertainty-aware semantic reasoning. Reasoning is performed after the penultimate decoder layer, where each query has obtained a preliminary category prediction and a relatively stable reference box. Let $s_i(c)$ denote the sigmoid score of query q_i for category c. We normalize the scores across categories and measure the prediction uncertainty using entropy:

$$p_i(c) = \frac{s_i(c)}{\sum_{c'=1}^{C} s_i(c') + \varepsilon}, \qquad u_i = -\frac{1}{\log C} \sum_{c=1}^{C} p_i(c) \log p_i(c), \tag{8}$$

where C is the number of categories and ε ensures numerical stability. A large u_i indicates that the current visual representation cannot clearly distinguish among several categories. We select the supplementary queries with the highest uncertainty as the semantic reasoning set $\mathcal{Q}_u$, while confident queries retain their original representations.

The weight vectors of the decoder classification head provide discriminative directions for different categories and can therefore serve as compact category representatives (Qi et al., 2018). We normalize these vectors and treat them as category prototypes, denoted by $\mathcal{E} = \{e_c\}_{c=1}^{C}$. For each uncertain query, the categories with the highest preliminary scores form a candidate set $\mathcal{T}_i$, and their prototypes are collected as $E_i = \{e_c \mid c \in \mathcal{T}_i\}$. The query then retrieves category-level evidence from these competing hypotheses:

$$h_i^{\text{sem}} = \text{MHA}(q_i, E_i, E_i), \qquad q_i^{\text{sem}} = \text{LN}\left(q_i + W_s h_i^{\text{sem}}\right), \tag{9}$$

where MHA denotes multi-head attention (Vaswani et al., 2017) and W_s projects the retrieved semantic evidence into the query space. Instead of prematurely assigning the query to its highest-scoring category, this operation preserves several plausible category hypotheses and introduces the discriminative directions required to distinguish them.

Reliability-guided relational reasoning. Category prototypes describe what an uncertain object may be, but its refined category hypothesis should also be consistent with the current scene. Unlike global proposal interaction or densely connected object graphs (Hu et al., 2018; Yang et al., 2018), we perform relational reasoning only between the uncertain supplementary query and a small set of reliable neighboring queries. The reliability of query q_j is measured by

$$\rho_j = (1 - u_j) \max_c s_j(c), \tag{10}$$

which is high only when the query has both a confident foreground prediction and a concentrated category distribution. For each uncertain supplementary query q_i^{sem}, we select the spatially nearest reliable original queries according to their reference-box centers, forming its neighborhood $\mathcal{N}_i$. We first retain original queries satisfying $\rho_j > \tau_r$ as reliable candidates and then select the eight spatially nearest candidates to form $\mathcal{N}_i$.The relevance of neighbor q_j to query q_i^{sem} is computed through local attention:

$$\alpha_{ij} = \text{Softmax}_{j \in \mathcal{N}_i}\left(\frac{(W_q q_i^{\text{sem}})^\top (W_k q_j)}{\sqrt{D}} + b_{ij}\right), \tag{11}$$

where D is the query dimension and b_{ij} is a learnable relative-position bias derived from their reference boxes, following the use of geometric cues in object relation modeling (Hu et al., 2018). The first term measures whether the neighboring object is compatible with the semantic hypotheses of the uncertain query, while the second encodes their spatial relationship. The resulting relational message is

$$m_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij} W_v q_j. \tag{12}$$

Because nearby objects are not necessarily informative for the current target, the relational message is incorporated through an adaptive gate:

$$\gamma_i = \sigma\left(W_g[q_i^{\text{sem}}; m_i]\right), \qquad \widehat{q}_i = \text{LN}\left(q_i^{\text{sem}} + \gamma_i \odot W_o m_i\right), \tag{13}$$

where $[\cdot;\cdot]$ denotes feature concatenation and γ_i controls the contribution of neighborhood evidence in each feature channel. Consequently, compatible relational cues are preserved, whereas messages that conflict with the query's semantic representation are suppressed.

The refined supplementary queries are combined with the original queries and passed to the final decoder layer. In this layer, cross-attention updates the refined queries using multi-scale image features before the final classification and box predictions. The complete framework is trained with

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{den}} \mathcal{L}_{\text{den}}, \tag{14}$$

where $\mathcal{L}_{\text{det}}$ is the original RT-DETR detection objective and $\mathcal{L}_{\text{den}}$ supervises regional density prediction. Both reasoning processes are optimized end-to-end without additional semantic or relation annotations.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate **Reasoning-DETR** on two complementary benchmarks. **VisDrone2019-DET** (Zhu et al., 2018) evaluates detection in real UAV imagery with substantial

Model	Publication	Input Size	Params (M)	GFLOPs	AP	AP_{50}
Real-time Object Detectors						
YOLOv8-L (Jocher et al., 2023)	Software 2023	640	43.7	165.2	26.1	42.7
YOLOv9-M (Wang et al., 2024b)	ECCV 2024	640	20.1	76.8	25.2	42.0
YOLOv10-L (Wang et al., 2024a)	NeurIPS 2024	640	24.4	120.3	26.3	43.1
Object Detectors for UAV Imagery						
QueryDet (Yang et al., 2022)	CVPR 2022	2400	33.9	212	28.3	48.1
ClusDet (Yang et al., 2019)	ICCV 2019	1000	30.2	207	26.7	50.6
DCFL (Xu et al., 2023)	CVPR 2023	1024	36.1	157.8	–	32.1
HIC-YOLOv5 (Tang et al., 2024)	ICRA 2024	640	9.4	31.2	26.0	44.3
End-to-end Object Detectors						
DETR (Carion et al., 2020)	ECCV 2020	1333	60.0	187	24.1	40.1
Deformable DETR (Zhu et al., 2021b)	ICLR 2021	1333	40.0	173	27.1	42.2
Sparse DETR (Roh et al., 2022)	ICLR 2022	1333	40.9	121	27.3	42.5
RT-DETR (Zhao et al., 2024)	CVPR 2024	640	32.9	108	29.7	50.6
Reasoning-DETR (Ours)	–	640	34.0	118	**33.5**	**54.2**

Table 1: Comparison with representative real-time, UAV-oriented, and end-to-end object detectors on VisDrone. Input Size denotes the primary input resolution (single-side notation).

variations in altitude, viewpoint, background, object scale, and scene density. It contains 10,209 images; we train on the 6,471-image training set and report results on the 548-image validation set. To further examine whether the proposed reasoning mechanism remains effective when visual evidence becomes extremely weak, we additionally evaluate on **AI-TOD-V2** (Xu et al., 2022b), a benchmark designed for tiny-object detection. Following the corresponding evaluation protocols, we report COCO-style AP together with AP_{50} and AP_{75} when available. For AI-TOD-V2, we additionally report the scale-specific AP_{vt}, AP_t, AP_s, and AP_m metrics.

Comparison methods. On VisDrone, we compare against representative real-time detectors, UAV-oriented detectors, and end-to-end Transformer detectors, including YOLOv8 (Jocher et al., 2023), YOLOv9 (Wang et al., 2024b), YOLOv10 (Wang et al., 2024a), QueryDet (Yang et al., 2022), ClusDet (Yang et al., 2019), HIC-YOLOv5 (Tang et al., 2024), DETR (Carion et al., 2020), Deformable DETR (Zhu et al., 2021b), Sparse DETR (Roh et al., 2022), and RT-DETR (Zhao et al., 2024). On AI-TOD-V2, we further compare with methods specialized for tiny objects and strong end-to-end detectors, including NWD (Xu et al., 2022b), RFLA (Xu et al., 2022a), DNTR (Huang et al., 2024b), DINO-DETR (Zhang et al., 2023), DQ-DETR (Huang et al., 2024a), D-FINE (Peng et al., 2025), and Dome-DETR (Hu et al., 2025). Together, these baselines cover real-time, UAV-oriented, tiny-object, and end-to-end detection paradigms.

Implementation details. We adopt COCO-pretrained RT-DETR-L with $K_b = 300$ base queries. On VisDrone, models are trained at 640×640 for 75 epochs on two RTX 3090 GPUs with a batch size of 6. We use at most $K_{\max} = 500$ supplementary queries and divide the highest-resolution feature into overlapping 5×5 regions (25% overlap). Regional counts are quantized into five intervals, $[0, 20)$, $[20, 40)$, $[40, 60)$, $[60, 80)$, and $[80, +\infty)$, with budgets $\{4, 16, 32, 48, 64\}$. Scene Memory uses 64 prototypes with top-4 retrieval, $\mu = 0.95$, and $\lambda_s = 8$. Each uncertain query retains its top-3 category hypotheses and interacts with eight reliable neighboring queries. On AI-TOD-V2, we follow the 800×800 trainval-to-test protocol. Results without * are taken from the corresponding original papers, while * denotes methods re-implemented by us with four feature levels. Our model uses the same input resolution and COCO-pretrained initialization as the re-implemented DETR baselines.

4.2 Comparison with State-of-the-Art Methods

Results on VisDrone. As shown in Table 1, Reasoning-DETR achieves the highest AP among the listed methods and consistently improves over the RT-DETR baseline under our training setting. Its clear advantage over the RT-DETR baseline demonstrates that the improvement does not merely arise from the underlying detector, but from adapting query allocation and refining uncertain candidates. Specifically, adaptive query allocation increases the search capacity in regions containing

Method	Publication	AP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
Non-end-to-end Object Detectors								
Cascade R-CNN (Cai & Vasconcelos, 2018)	CVPR 2018	15.1	34.2	11.2	0.1	11.5	26.7	38.5
DetectoRS w/ NWD (Xu et al., 2022b)	ISPRS 2022	20.8	49.3	14.3	6.4	19.7	29.6	38.3
DetectoRS w/ RFLA (Xu et al., 2022a)	ECCV 2022	24.8	55.2	18.5	9.3	24.8	30.3	38.2
DNTR (Huang et al., 2024b)	TGRS 2024	26.2	56.7	20.2	12.8	26.4	31.0	37.0
End-to-end Object Detectors								
DETR-DC5 (Carion et al., 2020)	ECCV 2020	10.4	32.5	3.9	3.6	9.3	13.2	24.6
Deformable DETR* (Zhu et al., 2021b)	ICLR 2021	18.2	59.7	10.2	5.8	16.8	25.2	34.0
DINO-DETR* (Zhang et al., 2023)	ICLR 2023	25.6	61.3	17.4	12.7	25.3	31.7	39.4
DQ-DETR* (Huang et al., 2024a)	ECCV 2024	30.2	68.6	22.3	15.3	30.5	36.5	44.6
D-FINE* (Peng et al., 2025)	ICLR 2025	31.3	66.7	24.9	15.3	31.0	38.4	46.6
Dome-DETR (Hu et al., 2025)	ACM MM 2025	34.6	69.2	32.0	19.0	35.6	39.8	47.3
Reasoning-DETR (Ours)		**36.3**	**70.5**	**32.9**	**20.4**	**37.1**	**41.0**	**47.9**

Table 2: Comparison with representative object detectors on AI-TOD-V2. All models are trained on the trainval split and evaluated on the test split with 800×800 input resolution. * denotes re-implementation with four feature levels

.

dense or weakly responding objects, while uncertainty-aware refinement enhances the representations of ambiguous candidates using semantic and relational information. The consistent improvement in detection performance suggests that these two components complement each other in addressing missed detections and recognition ambiguity. These results answer RQ1 by confirming that the proposed search and reasoning framework is effective for detecting weak and densely distributed objects in complex aerial scenes.

Results on AI-TOD-V2. As shown in Table 2, Reasoning-DETR achieves the highest reported performance among the compared methods across all metrics. The improvements in AP, AP_{50}, and AP_{75} demonstrate stronger overall detection performance while maintaining accurate localization under different IoU thresholds. The scale-specific metrics AP_{vt}, AP_t, AP_s, and AP_m evaluate very tiny, tiny, small, and medium objects, respectively. The clear gains on AP_{vt} and AP_t indicate improved detection of objects with extremely weak visual evidence, while the consistent improvements on AP_s and AP_m show that the benefit extends across different object scales. Together with the semantic and relational reasoning module, these results suggest that Reasoning-DETR improves both tiny-object discovery and the recognition of ambiguous candidates.

Computational efficiency. As shown in Table 1, Reasoning-DETR introduces only a modest increase in parameters and GFLOPs over the RT-DETR baseline. The additional cost remains limited because supplementary queries are allocated according to regional demand, and semantic and relational reasoning are performed only for uncertain queries within small local neighborhoods. Moreover, the method reuses the original encoder, decoder, and detection heads without introducing a larger backbone or additional high resolution branches. These designs allow Reasoning-DETR to retain competitive computational efficiency while making more effective use of its query budget.

4.3 Ablation Study and Visualization

Component ablation. To disentangle the effect of query number from the contribution of the proposed modules, Table 3 first compares 300 and 500 Top-K queries without introducing any proposed component, and then progressively adds Scene Memory, density prediction, semantic reasoning, and relational reasoning. Increasing the Top-K query number from 300 to 500 improves AP only from 29.7 to 30.7, indicating that simply enlarging the query set provides limited gains. Introducing Scene Memory further raises AP to 31.2 by guiding supplementary queries toward scene-relevant regions, while density prediction improves AP to 32.0 by adapting the regional query demand to the current object distribution. Semantic reasoning then increases AP to 32.7 by refining ambiguous queries with category prototypes, and relational reasoning further improves the performance to 33.5 AP and 54.2 AP_{50} by selectively incorporating evidence from reliable neighboring queries. Overall, the complete model improves the original 300-query baseline by 3.8 AP, showing that the performance

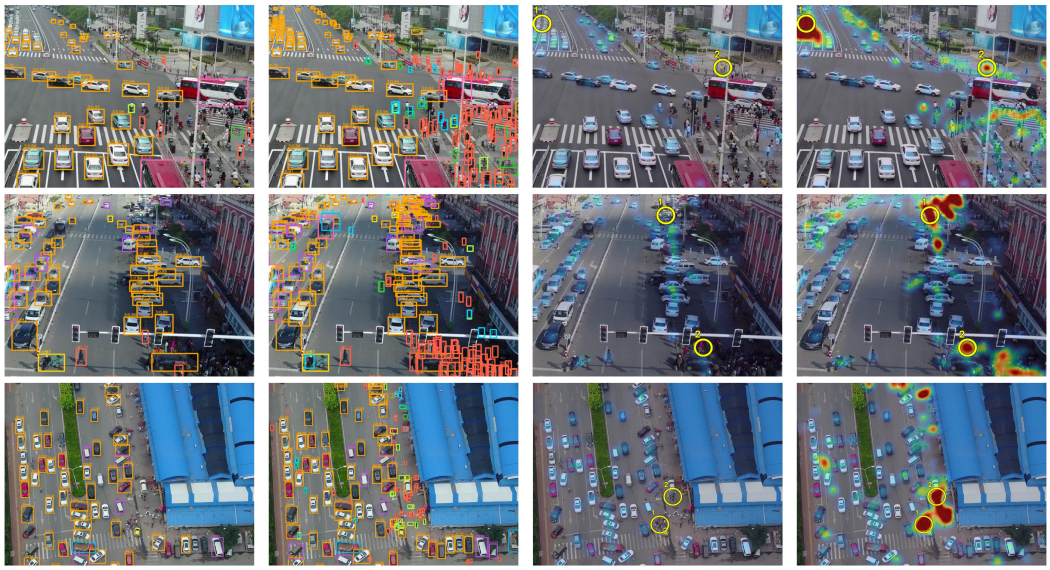

Figure 3: **Qualitative comparison.** Detection results of (a) DETR and (b) RT-DETR, and response heatmaps (c) before and (d) after dynamic query allocation.

Configuration	AP_{50}	AP	Configuration	AP_{50}	AP
Top-K only ($n_q = 300$)	50.6	29.7	+ Density Prediction	52.6	32.0
Top-K only ($n_q = 500$)	51.5	30.7	+ Semantic Reasoning	53.4	32.7
+ Scene Memory	51.8	31.2	+ Relation Reasoning	**54.2**	**33.5**

Table 3: Ablation study on VisDrone. Increasing the query number alone yields limited improvement, while progressively introducing the proposed components leads to consistent gains.

gain mainly comes from the proposed adaptive query allocation and uncertainty-aware reasoning mechanisms rather than simply increasing the number of queries.

Visulization. Figure 3 compares detection results and response heatmaps in representative aerial scenes. DETR in (a) misses many small objects because their weak features receive insufficient attention, while RT-DETR in (b) recovers more targets but still struggles in crowded regions and complex backgrounds. Reasoning-DETR produces more complete detections for densely distributed vehicles and isolated tiny objects, while reducing responses to visually distracting background structures. The comparison between the heatmaps in (c) and (d) further shows that our responses are more concentrated around object-relevant regions instead of being diffusely distributed across the scene. This behavior reflects the complementary roles of the two modules: scene-aware query allocation improves the coverage of regions containing potential objects, while semantic and relational reasoning helps distinguish ambiguous candidates using category prototypes and reliable neighboring queries. The qualitative results therefore suggest that the performance gain arises from both improved object discovery and more reliable query refinement.

5 Conclusion

We presented Reasoning-DETR, which improves aerial detection through a coupled process of contextual query allocation and uncertainty-aware query reasoning. Regional density and Scene Memory adapt the search capacity to each scene, while category prototypes and reliable local neighbors refine supplementary queries with weak visual evidence. Their complementary roles jointly address insufficient query coverage and ambiguous object representations in challenging aerial scenes. Both components remain integrated into the end-to-end RT-DETR pipeline and require no additional scene or relation annotations. Experiments on VisDrone and AI-TOD-V2 demonstrate the effectiveness of the proposed framework across aerial and extreme tiny-object detection settings.

AI USE STATEMENT

In the preparation of this manuscript, we used generative AI tools, specifically ChatGPT-5.6 Luna, to assist with language polishing, including the correction of grammatical errors and improvement of clarity and readability. Generative AI was not used to generate or modify figures or images, or to alter the technical content or methodology of this work. All research ideas, methodological decisions, interpretations, and experiments were developed and conducted independently by the authors. All AI-assisted text was carefully reviewed by the authors, and we take full responsibility for the final content of this work, including all text, figures, and experimental results.

REFERENCES

Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13075–13085, 2020.

Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27706–27716, 2024.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group DETR: Fast DETR training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6633–6642, 2023.

Bowei Du, Yecheng Huang, Jiaxin Chen, and Di Huang. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zuhao Ge, Lizhe Qi, Yuzheng Wang, and Yunquan Sun. Zoom-and-reasoning: Joint foreground zoom and visual-semantic reasoning detection network for aerial images. *IEEE Signal Processing Letters*, 29:2572–2576, 2022. doi: 10.1109/LSP.2022.3229638.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, 2018.

Zhangchi Hu, Peixi Wu, Jie Chen, Huyue Zhu, Yijun Wang, Yansong Peng, Hebei Li, and Xiaoyan Sun. Dome-DETR: DETR with density-oriented feature-query manipulation for efficient tiny object detection. In *Proceedings of the ACM International Conference on Multimedia*, pp. 101–110, 2025. doi: 10.1145/3746027.3754861.

Yecheng Huang, Jiaxin Chen, and Di Huang. UFPMP-Det: Toward accurate and efficient object detection on drone imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1026–1033, 2022. doi: 10.1609/aaai.v36i1.19986.

Yi-Xin Huang, Hou-I Liu, Hong-Han Shuai, and Wen-Huang Cheng. Dq-detr: Detr with dynamic query for tiny object detection. In *European Conference on Computer Vision*, pp. 290–305. Springer, 2024a. doi: 10.1007/978-3-031-73116-7_17.

Yi-Xin Huang et al. A denoising fpn with transformer r-cnn for tiny object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024b.

Linxia Ji, Li Yan, Hong Xie, Tengfei Wang, Longze Zhu, Hao Wu, and Duonan Wang. RDA-DETR: Rotation-aware decoupled oriented object detection in aerial images with agent query. *ISPRS Journal of Photogrammetry and Remote Sensing*, 240:476–491, 2026. doi: 10.1016/j.isprsjprs.2026.07.017.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. URL https://github.com/ultralytics/ultralytics.

Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density map guided object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel network for remote sensing object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16794–16805, 2023.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oMI9PjOb9Jl.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, 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.

Akhil Meethal, Eric Granger, and Marco Pedersoli. Cascaded zoom-in detector for high resolution aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2046–2055, 2023.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3631–3640, 2021. doi: 10.1109/ICCV48922.2021.00363.

Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-fine: Redefine regression task of detrs as fine-grained distribution refinement. In *International Conference on Learning Representations*, 2025.

Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5822–5830, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. In *International Conference on Learning Representations (ICLR)*, 2022.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Huixin Sun, Runqi Wang, Yanjing Li, Linlin Yang, Shaohui Lin, Xianbin Cao, and Baochang Zhang. SET: Spectral enhancement for tiny object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4713–4723, 2025. doi: 10.1109/CVPR52734.2025.00444.

Shiyi Tang, Shu Zhang, and Yini Fang. Hic-yolov5: Improved yolov5 for small object detection. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6614–6619, 2024. doi: 10.1109/ICRA57147.2024.10610273.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOv10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems*, 2024a.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision (ECCV)*, 2024b. doi: 10.1007/978-3-031-72751-1_1.

Chang Xu, Jinwang Wang, Wen Yang, and Huai Yu. Rfla: Gaussian receptive field based label assignment for tiny object detection. In *European Conference on Computer Vision*, 2022a.

Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:79–93, 2022b. doi: 10.1016/j.isprsjprs.2022.06.002.

Chang Xu, Jian Ding, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Dynamic coarse-to-fine learning for oriented tiny object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7318–7328, 2023.

Hang Xu, Chenhan Jiang, Xiaodan Liang, Liang Lin, and Zhenguo Li. Reasoning-RCNN: Unifying adaptive global reasoning into large-scale object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jintao Xu, Yali Li, and Shengjin Wang. AdaZoom: Towards scale-aware large scene object detection. *IEEE Transactions on Multimedia*, 25:4598–4609, 2022c. doi: 10.1109/TMM.2022.3178871.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13668–13677, 2022. doi: 10.1109/CVPR52688.2022.01330.

Fan Yang, Heng Fan, Peng Chu, Erik Blasch, and Haibin Ling. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8311–8320, 2019.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *European Conference on Computer Vision*, pp. 670–685, 2018.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023.

Huaxiang Zhang, Hao Zhang, Kai Liu, Zhongxue Gan, and Guo-Niu Zhu. UAV-DETR: Efficient end-to-end object detection for unmanned aerial vehicle imagery. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 15143–15149, 2025. doi: 10.1109/IROS60139.2025.11246176.

Qi Zhao, Binghao Liu, Shuchang Lyu, Chunlei Wang, and Hong Zhang. TPH-YOLOv5++: Boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer. *Remote Sensing*, 15(6):1687, 2023. doi: 10.3390/rs15061687.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024.

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

Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2778–2788, 2021a. doi: 10.1109/ICCVW54120.2021.00312.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021b.

Algorithm 1 Aerial Object Detection with Reasoning-DETR

Input: training image and annotations (I, Y); base-query number K_b; maximum supplementary budget $K_{\max}$; Scene Memory $\mathcal{M}$; category prototypes $\mathcal{E}$; decoder layers L.
Output: detection set $\widehat{Y}$.

Extract multi-scale features $\{F_l\}_{l=1}^{S} = \mathrm{Encoder}(I)$.
Select K_b base queries $\mathcal{Q}_b$ from encoder candidates.
Partition the highest-resolution feature into regions $\{\mathcal{R}_g\}_{g=1}^{G}$.
for $g = 1$ to G **do**
    Predict density probabilities $\mathbf{p}_g = \mathrm{Softmax}(f_d(\mathcal{R}_g))$.
    Retrieve top-T scene prototypes and compute relevance s_g by Eq. equation 3.
    Compute demand $\widetilde{n}_g = \sum_r p_g(r) B_r + \lambda_s s_g$.
    Select regional foreground candidates to initialize $\mathcal{Q}_s^g$.
end for
Normalize $\{\widetilde{n}_g\}$ if $\sum_g \widetilde{n}_g > K_{\max}$; set $\mathcal{Q}_s = \bigcup_g \mathcal{Q}_s^g$.
Decode $\mathcal{Q}_b \cup \mathcal{Q}_s$ through the first $L-1$ decoder layers.
Compute normalized entropy u_i for every $q_i \in \mathcal{Q}_s$ by Eq. equation 8.
for each uncertain supplementary query q_i **do**
    Retrieve top-3 category prototypes and obtain q_i^{sem}.
    Select eight nearby base queries with high reliability $\rho_j = (1 - u_j) \max_c s_j(c)$.
    Aggregate the local message $m_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij} W_v q_j$.
    Gate the message: $\widehat{q}_i = \mathrm{LN}(q_i^{\mathrm{sem}} + \gamma_i \odot W_o m_i)$.
end for
Update $\mathcal{Q}_b \cup \widehat{\mathcal{Q}}_s$ with the final decoder layer L and predict $\widehat{Y}$.
Optimize $\mathcal{L} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{den}} \mathcal{L}_{\mathrm{den}}$.

A APPENDIX

Appendix overview. The appendix provides complementary material organized around the two stages of Reasoning-DETR: adaptive query discovery and uncertainty-aware query refinement. Section B first presents the end-to-end algorithm and discusses the design of its allocation and refinement components. Section C then provides additional sensitivity, allocation, component, and computational analyses. Section D relates these measurements to qualitative observations. Finally, Section E outlines the current scope and several directions for future work.

B ALGORITHMIC ANALYSIS AND DESIGN INTUITION

This section makes the overall computational flow explicit before examining its individual design choices. We first specify how supplementary queries are allocated and refined within the original detection pipeline. We then discuss budget implementation, the intuition behind regional allocation, the bounded contribution of memory and relational messages, and the additional computational cost.

End-to-end algorithm. Algorithm 1 summarizes the two proposed modules and their placement in RT-DETR. Regional density and Scene Memory determine the demand for supplementary queries, which are initialized from encoder candidates. After the first $L-1$ decoder layers, category prototypes and reliable nearby queries refine uncertain supplementary representations. The final decoder layer then uses multi-scale visual features for another standard query update and produces detection predictions. The method requires no scene, relation, or language annotations.

Operational view. Figure 4 illustrates how the two modules interact with the original detection pipeline. The allocator acts before decoding and changes where the model spends its supplementary query budget. The reasoner acts after preliminary decoder predictions become available and changes only the selected uncertain query representations. The remaining decoder layer continues ordinary visual-feature interaction, so the two proposed modules extend rather than replace the existing RT-DETR decoder.

Implementation details. The VisDrone configuration exactly follows the main paper: $K_b = 300$, a 5×5 overlapping regional partition with 25% overlap, density intervals $[0, 20)$, $[20, 40)$, $[40, 60)$,

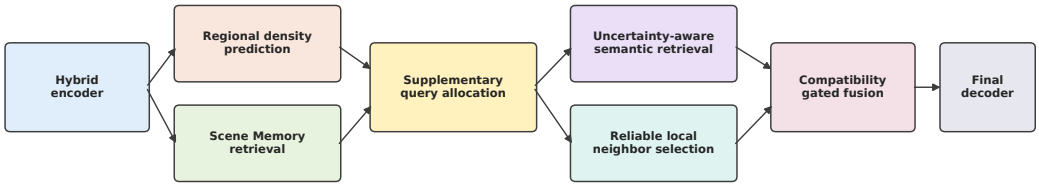

Figure 4: **Algorithmic flow of Reasoning-DETR.** Regional density and Scene Memory allocate supplementary queries; uncertainty-aware semantic and relational reasoning refines ambiguous candidates; the final decoder layer further updates the resulting queries using image features.

$[60, 80)$, and $[80, +\infty)$, and budgets $\{4, 16, 32, 48, 64\}$. We cap supplementary queries at $K_{\max} = 500$. Scene Memory contains 64 prototypes, uses top-4 retrieval and momentum $\mu = 0.95$, and is weighted by $\lambda_s = 8$. Each uncertain query retains three semantic hypotheses and interacts with eight reliable neighboring base queries. These settings are fixed for all VisDrone ablations unless explicitly varied.

Query budget discretization. For each region, the continuous demand $\widetilde{n}_g$ is first converted into a non-negative integer budget. When the summed regional budget exceeds $K_{\max}$, we proportionally rescale the regional demands to satisfy the image-level query limit. Since adjacent regions overlap, candidate encoder locations selected by multiple regions are globally deduplicated, and duplicated slots are filled by the next highest-scoring unused candidates.

Why adaptive allocation improves coverage. Let O_g be the unknown number of detectable objects in region g, and let n_g denote its assigned supplementary queries. For a fixed per-query hit probability p_g, a simple lower-bound model gives regional miss probability $P_{\text{miss}}(g) \leq (1 - p_g)^{n_g}$. Thus, the marginal coverage gain is

$$\Delta_g(n_g) = p_g(1 - p_g)^{n_g}, \tag{15}$$

which decreases with n_g. Under a global budget, allocating the next query to the region with the largest estimated Δ_g is therefore preferable to uniform allocation. Density prediction estimates demand through O_g, while Scene Memory improves the estimate of p_g by identifying object-favorable contexts. Their combination approximates a diminishing-return resource allocation rule and explains why gains are largest in crowded regions without assigning the maximum budget to every image.

Robustness of scene-memory retrieval. For normalized regional feature c_g and prototypes M_k, the scene score averages positive cosine similarities over top-T memories. Since cosine similarity is bounded in $[-1, 1]$, $s_g \in [0, 1]$. If the regional feature is perturbed from c_g to $c_g + \delta$ and remains nonzero, cosine similarity is locally Lipschitz; consequently $|s_g(c_g + \delta) - s_g(c_g)| \leq C\|\delta\|_2$ for a finite C. Top-T averaging further reduces dependence on any single prototype. This bounded contribution is important: memory can increase search attention in plausible regions but cannot create an unbounded query demand.

Reliability-gated message stability. The relational update is a residual transformation $\widehat{q}_i = \mathrm{LN}(q_i^{\text{sem}} + \gamma_i \odot W_o m_i)$ with $\gamma_i \in (0, 1)^D$. Before layer normalization, its deviation from semantic-only reasoning satisfies

$$\|\gamma_i \odot W_o m_i\|_2 \leq \|W_o\|_2 \|m_i\|_2. \tag{16}$$

The gate therefore cannot amplify a relational message beyond the ungated projection and may suppress incompatible dimensions independently. Reliability filtering also constrains the neighborhood to queries with concentrated foreground predictions. Together these mechanisms reduce negative transfer relative to unrestricted self-attention, especially when weak aerial targets are surrounded by background candidates.

Computational complexity. For G regions, M memory prototypes, K_u uncertain queries, C classes, and K_n selected neighbors, allocation costs $\mathcal{O}(GM)$ for memory retrieval plus lightweight regional prediction. Semantic reasoning costs $\mathcal{O}(K_u C)$ for prototype scoring, and local relation reasoning costs $\mathcal{O}(K_u K_n D)$. In practice, $G = 25$, $M = 64$, and $K_n = 8$, while reasoning is restricted to the uncertain subset. This is substantially cheaper than constructing dense relations among all decoder queries, whose pairwise interaction grows quadratically with the enlarged query count.

Training and inference consistency. Density labels are derived only from the same bounding-box annotations used by the detector. Scene Memory is updated from positive regional features during training and frozen as learned model state during evaluation. Category prototypes reuse classifier weights, and reference boxes determine local neighborhoods. Hence every auxiliary signal is internal to the detector, and inference requires neither ground-truth counts nor external metadata.

C Additional Experimental Results

This section complements the main-paper comparisons with analyses at three levels. First, parameter sensitivity on UAVDT examines the effect of the reliability threshold and scene-memory weighting. Second, repeatability and pre-decoder coverage measurements characterize the stability of the reported results and the contribution of query allocation. Third, component-level and budget analyses examine uncertainty-aware refinement and the computation–accuracy trade-off. We keep these measurements separate because pre-decoder coverage, final detection AP, and efficiency quantify different aspects of the system.

Sensitivity analysis of τ and α on UAVDT. Here, α is a normalized version of the Scene Memory coefficient λ_s, used only to present the sensitivity analysis on a common scale. We additionally use the UAVDT validation split to test whether the reliability threshold τ and scene-memory weight α require narrow tuning. Table 4 and Fig. 5 show a broad optimum around $\tau = 0.45$ and $\alpha = 0.50$. Performance changes by at most 1.4 AP over the full grid and by only 0.5 AP in the central neighborhood, indicating that the method is not sensitive to a single exact operating point. The selected pair is used for this auxiliary UAVDT experiment; the VisDrone setting remains the one reported in the main paper.

Single-parameter decomposition. Holding $\alpha = 0.50$, a low threshold admits ambiguous neighbors and a high threshold discards useful context. Holding $\tau = 0.45$, excessive memory weighting overemphasizes prior scene patterns, whereas a very small weight approaches density-only allocation. The asymmetric degradation is expected: relational noise mainly affects uncertain-query refinement, while excessive scene weighting affects where the finite query budget is spent.

Result statistical analysis. To assess repeatability, we train the full VisDrone model with five random seeds while retaining the protocol of the main paper. Table 6 reports a mean AP of 33.50 with a standard deviation of 0.10; the corresponding AP_{50} mean is 54.20. The main-paper result, 33.5 AP and 54.2 AP_{50}, therefore coincides with the multi-run mean rather than an unusually favorable seed.

Confidence intervals and significance. Using a Student-t interval over the five runs gives a 95% confidence interval of approximately $[33.38, 33.62]$ AP. The lower bound remains well above the 29.7 AP RT-DETR baseline reported in Table 1.

Query allocation diagnosis. Table 7 evaluates query coverage before decoder refinement. Uniformly adding queries improves coverage, but density-aware placement is more efficient. Combining density with Scene Memory gives the highest overall and small-object coverage, consistent with the progressive main-paper ablation from 29.7 AP for Top-K only to 32.0 AP after both allocation cues are included. This analysis characterizes pre-decoder search behavior and complements, rather than replaces, the final detection AP comparisons in the main paper.

Reasoning diagnosis under a consistent VisDrone protocol. Table 8 expands the final stages of the component ablation in the main paper. The AP values preserve the exact cumulative trajectory

$\tau \backslash \alpha$	0.25	0.50	0.75	1.00
0.30	24.7	25.0	24.9	24.5
0.40	25.1	25.5	25.4	25.0
0.45	25.3	**25.8**	25.6	25.2
0.50	25.0	25.5	25.3	24.9
0.60	24.6	25.0	24.8	24.4

Table 4: **Sensitivity of UAVDT validation AP to reliability threshold τ and scene-memory weight α.** The best setting is highlighted.

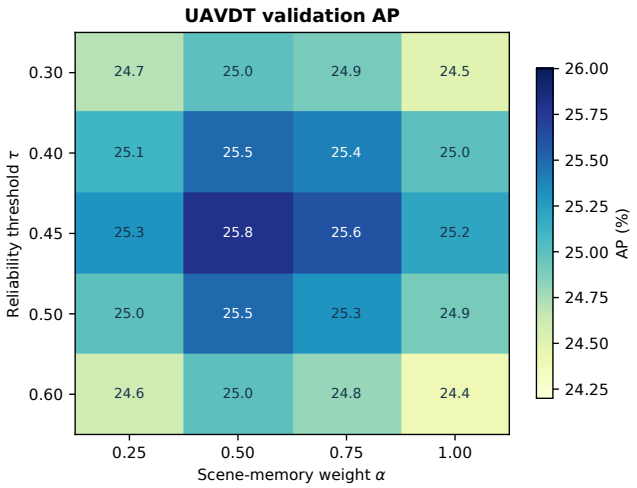

Figure 5: **Two-dimensional parameter sensitivity on UAVDT.** The smooth central plateau supports the robustness observed in Table 4.

Setting	AP	AP_{50}	AP_{75}	AR_{100}
$\tau = 0.30, \alpha = 0.50$	25.0	44.8	24.1	38.0
$\tau = 0.45, \alpha = 0.25$	25.3	45.2	24.5	38.4
$\tau = 0.45, \alpha = 0.50$	**25.8**	**45.9**	**25.1**	**39.1**
$\tau = 0.45, \alpha = 1.00$	25.2	45.0	24.3	38.3
$\tau = 0.60, \alpha = 0.50$	25.0	44.7	24.0	37.8

Table 5: **Detailed UAVDT metrics for representative parameter settings.** AP_{50}, AP_{75}, and recall follow the same trend as overall AP.

reported there: 32.0 after density and Scene Memory, 32.7 after semantic reasoning, and 33.5 after relation reasoning. Uncertain-query accuracy increases while negative log-likelihood decreases, showing that the final AP gain accompanies improved confidence calibration rather than only more detections.

Supplementary query budget. Table 9 examines how the maximum number of supplementary queries affects detection accuracy and computation. Increasing $K_{\max}$ improves AP substantially up to the adopted budget of 500, while the additional gain at 700 is small relative to the increased GFLOPs. The zero-budget row corresponds to the 29.7 AP RT-DETR baseline in Table 1. This analysis clarifies why we use a bounded supplementary pool rather than simply maximizing the query count.

Efficiency decomposition. The complete model has 34.0M parameters and 118 GFLOPs, matching Table 1. Most extra computation comes from decoding the adaptive supplementary pool; the allo-

Run	AP	AP_{50}	AP_{75}	AR
Seed 0	33.4	54.1	34.2	49.8
Seed 1	33.5	54.2	34.4	50.0
Seed 2	33.6	54.3	34.5	50.1
Seed 3	33.4	54.1	34.3	49.9
Seed 4	33.6	54.3	34.4	50.1
Mean $\pm$ std.	33.50 ± 0.10	54.20 ± 0.10	34.36 ± 0.11	49.98 ± 0.13

Table 6: **Five-run statistical analysis on the VisDrone validation set.** The mean AP and AP_{50} exactly match the headline values in Table 1.

Allocation strategy	GT coverage	Small GT coverage	Pre-decoder AR
Original RT-DETR Top-K	86.9	78.4	66.7
Uniform supplementary queries	89.5	82.1	68.8
Global Top-K supplementary queries	90.2	82.7	69.4
Regional density only	90.8	84.3	70.5
Density + Scene Memory	**91.7**	**86.1**	**71.8**

Table 7: **Effect of allocation before the decoder on VisDrone.** Coverage is the percentage of ground-truth objects with at least one selected query in their local neighborhood.

Configuration	AP	AP_{50}	Acc_u	NLL_u
Density + Scene Memory	32.0	52.6	42.1	1.39
+ Semantic prototypes	32.7	53.4	47.0	1.25
+ Reliable local relations, no gate	33.1	53.8	48.6	1.20
+ Reliability gate	**33.5**	**54.2**	**50.2**	**1.14**

Table 8: **Detailed reasoning analysis on VisDrone.** Acc_u and NLL_u are computed over the uncertain supplementary-query subset.

cator and reasoner together contribute only 1.5 GFLOPs. Restricting reasoning to uncertain queries and eight neighbors avoids dense pairwise reasoning over all queries.

Cross-dataset consistency. On AI-TOD-V2, Reasoning-DETR reaches 36.3 AP, 70.5 AP_{50}, and 32.9 AP_{75}, with 20.4 AP_{vt} for very tiny objects. These values are copied from Table 2. Together with the VisDrone small-object coverage analysis and the auxiliary UAVDT sensitivity experiment, they indicate that improvements are not tied to one density distribution or one IoU threshold.

D Additional Visualizations

The following visual analyses connect the quantitative results to the behavior of the two proposed modules. We first examine where supplementary queries are allocated across scenes of different densities, then discuss difficult cases and the scope of the visual evidence. These observations are intended to illustrate model behavior alongside the quantitative experiments, rather than to establish that the internal representations are intrinsically interpretable.

Query allocation across scene density. Figure 6 shows that supplementary-query usage rises with scene density rather than remaining fixed. The associated increases in small-object coverage and pre-decoder recall are strongest in crowded images, where competition in global Top-K selection is most severe. This pattern is consistent with the diminishing-return allocation intuition discussed in Section B.

Challenging visual cases. The visualizations also help delineate scenarios in which the proposed operations have limited information to exploit. For extremely small or severely blurred objects, allocating additional queries does not necessarily produce a discriminative feature representation. Under heavy occlusion, nearby predictions can be uncertain as well, leaving fewer candidates for the

$K_{\max}$	Maximum total queries	AP	AP_{50}	GFLOPs
0	300	29.7	50.6	108.0
100	400	30.4	51.2	110.2
200	500	31.7	51.6	112.3
300	600	32.6	52.9	114.4
500	800	**33.5**	**54.2**	118.0
700	1000	33.6	54.3	122.1

Table 9: **Sensitivity to the maximum supplementary-query budget on VisDrone.** The adopted budget captures most of the available gain.

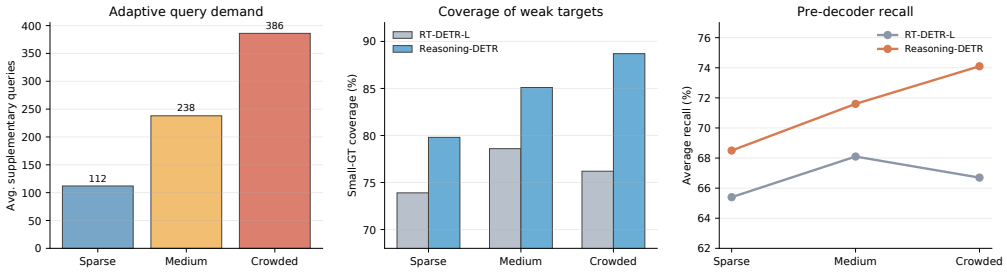

Figure 6: **Supplementary-query allocation on VisDrone.** The allocator assigns more capacity to crowded scenes and improves coverage and recall before decoder refinement.

Model variant	Params (M)	GFLOPs	AP	AP_{50}
RT-DETR-L	32.9	108.0	29.7	50.6
+ Adaptive query allocation	33.5	116.5	32.0	52.6
+ Semantic and relational reasoning	34.0	118.0	**33.5**	**54.2**

Table 10: **Accuracy and computational decomposition at** 640×640**.** The first and last rows reproduce the main-paper baseline and final result.

reliability-filtered neighborhood. For uncommon scene layouts, a learned memory may offer weaker contextual guidance than it does on familiar patterns. These examples explain why query placement and contextual refinement complement the underlying image evidence rather than substitute for it.

Relation to the quantitative analyses. The visual results should be read together with the measurements in Section C. Confusion matrices and precision-related curves characterize the resulting predictions, while the spatial visualizations illustrate the placement of supplementary queries. The pre-decoder coverage analysis provides the direct quantitative test of improved candidate discovery; the component ablation separately measures the contribution of semantic and relational refinement. This separation avoids attributing a specific causal mechanism to a visualization alone.

E Limitations and Future Work

Reasoning-DETR is designed to improve how a finite query budget is allocated and how uncertain supplementary queries use category and local-context information. Its scope is therefore tied to the quality of the available visual and contextual evidence. When objects occupy very few pixels or are heavily occluded, additional queries and neighboring representations may contain insufficient information for reliable refinement. Likewise, Scene Memory reflects the training distribution and may provide less useful guidance for scenes that differ markedly from those observed during training. These cases motivate further study rather than implying that the proposed components can recover information absent from the image.

Several extensions follow naturally from the current design. An adaptive memory-update strategy could explore how to maintain useful scene prototypes under changing scene distributions, while more selective query pruning could reduce the extra decoding cost introduced by supplementary queries. A further direction is to investigate when semantic and relational information should be used jointly or independently, particularly when nearby predictions are themselves uncertain. We leave these directions for future work and do not claim improvements from them in the present experiments.