Correlate-and-Compress: Enabling Multimodal Large Language Models to Organize Evidence for Generative Grounding

Anonymous ACL submission

Abstract

Multimodal large language models can perform referring expression grounding by directly generating bounding box coordinates, but they remain unreliable in high-resolution UAV imagery, where tiny targets are easily diluted by redundant background tokens. We identify this problem as a visual evidence bottleneck: small-object cues may become inaccessible before language-side reasoning. We propose Q-REX, a query-deferred evidence organization framework for generative grounding. Instead of modifying the visual encoder, adding detection heads, or relying on iterative crop-and-zoom search, Q-REX operates directly on native MLLM visual tokens. Its Correlate-and-Compress module constructs a spatial-semantic graph to route tokens into compact object-like foreground evidence, while masked background aggregation compresses unselected tokens into contextual memory for subsequent language-conditioned decoding. The referring expression is introduced only during autoregressive decoding, where the MLLM disambiguates among the organized evidence and generates the target box. Experiments on SkyFind show that Q-REX achieves 65.35 Acc@0.5 and 52.85 mIoU on the test set, outperforming representative task-specific grounding methods and the direct Qwen3-VL-2B baseline. Ablations and visualizations further show that CnC preserves object-level evidence and forms compact candidate clusters for high-resolution UAV grounding. Code: https://anonymous.4open.science/r/D18EAA578/

1 Introduction

Referring Expression Grounding (REG) aims to localize objects described by natural language expressions within an image (Mao et al., 2016; Yu et al., 2016, 2018). In recent years, the emergence of Multimodal Large Language Models (MLLMs) has enabled generative REG, where bounding boxes are serialized into coordinate tokens and generated through a language generation interface (Chen et al., 2023b; Wang et al., 2024; Bai et al., 2025a). This paradigm circumvents the need for task-specific detection heads and provides a unified framework for visual reasoning and grounding.

Despite these advances, generative grounding still faces significant challenges when handling high-resolution Unmanned Aerial Vehicle (UAV) images. UAV scenarios typically feature top-down views covering vast areas, causing the referred objects to often be extremely small and frequently hidden among dense, homogeneous distractors. Meanwhile, given a limited visual token budget, the critical cues for small objects may occupy only a handful of tokens, whereas large homogeneous regions—such as roads, roofs, vegetation, and water bodies—dominate the overall visual representation. Consequently, the fine-grained information relied upon by language-conditioned grounding is highly susceptible to being suppressed, diluted, or engulfed by background features before entering the language decoder. We refer to this failure mode as the "visual evidence bottleneck".

Existing methods mainly improve visual coverage through high-resolution encoding, crop-and-zoom search, or token reduction (Guo et al., 2024; Li et al., 2024; Wang et al., 2024; Ge et al., 2024; Bai et al., 2025a; Shang et al., 2025; Zhang et al., 2025; Luo et al., 2025). However, high-resolution or local-search strategies may increase cost or fragment global layout, while generic token compression may discard weak target cues or context needed for relational reasoning. Thus, the key challenge is not simply to retain more visual information, but to organize visual evidence into a decoder-accessible context that supports language-conditioned disambiguation and coordinate generation.

To address this challenge, we propose Q-REX, a Correlate-and-Compress framework that reorga-

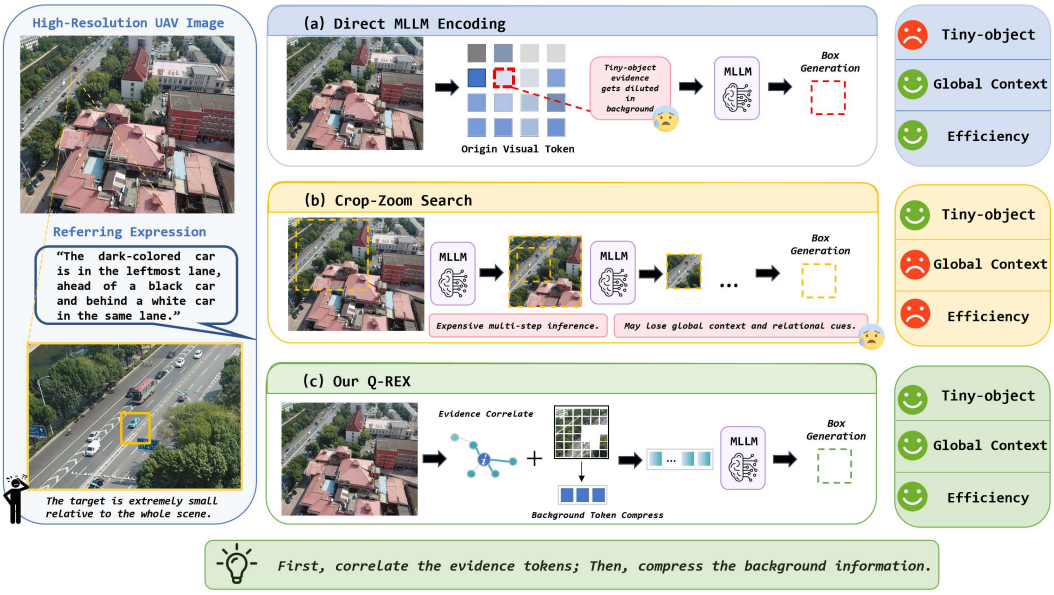

Figure 1: Motivation of Q-REX. Tiny target cues in high-resolution UAV imagery are easily diluted by redundant background tokens. Q-REX addresses this visual evidence bottleneck by reorganizing Qwen3-VL native visual tokens into compact foreground evidence and compressed background context for generative grounding.

nizes the native visual tokens of an MLLM for generative grounding in high-resolution UAV imagery. Instead of directly selecting the referred object at an early stage, Q-REX first constructs a spatial-semantic graph over native visual tokens to identify compact object-like evidence candidates, and compresses the remaining tokens into a lightweight background context memory. The graph routing stage is not designed to determine the final referred instance, but to preserve high-recall small-object candidate clusters together with surrounding scene context. The referring expression is then used by the MLLM during autoregressive decoding to disambiguate among the organized visual evidence and generate the target bounding box. In this way, Q-REX improves the accessibility of small-object cues under a constrained token budget while retaining the relational context required for high-resolution UAV grounding, as illustrated in Figure 1.

The main contributions of this paper are summarized as follows:

- We identify the visual evidence bottleneck in generative grounding for high-resolution UAV imagery and propose Q-REX, an evidence organization framework based on the native visual tokens of MLLMs.
- We introduce the Correlate-and-Compress module that routes visual tokens through spatial-semantic graph reasoning and aggregates unselected tokens into a fixed-size context memory, thereby enabling the coexistence of compact object-like evidence and global scene context under a limited visual token budget.
- Experiments on the SkyFind benchmark demonstrate that Q-REX outperforms direct MLLM-based grounding methods and representative task-specific baselines. Ablation studies and visualization results further demonstrate that the compact organization of native visual evidence facilitates more reliable localization of small objects in high-resolution scenarios.

2 Related Work

2.1 Referring Expression Grounding

Referring Expression Grounding (REG) aims to localize the object specified by a natural-language expression. Early methods align expressions with candidate regions using region proposals, modular reasoning, or task-specific matching networks (Mao et al., 2016; Yu et al., 2016, 2018). Recent approaches further formulate grounding as end-to-end localization or open-vocabulary detection, where models such as MDETR, GLIP, and Grounding DINO learn text-conditioned region alignment through detection-oriented objectives (Kamath

et al., 2021; Li et al., 2022; Liu et al., 2024).

Another line of work recasts grounding as a generative task, where spatial coordinates are serialized as textual tokens and predicted through the language interface of an MLLM, as in Shikra (Chen et al., 2023b). This formulation removes task-specific detection heads and is closely related to our setting. However, most REG methods are designed for natural images or detection-oriented grounding, and rarely address visual-evidence accessibility when referred objects occupy only a few tokens in high-resolution UAV scenes. Our work therefore focuses on pre-decoding evidence organization for generative MLLM grounding, where the central issue is how to construct a compact visual context that remains accessible to the language decoder.

2.2 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) provide a unified interface for visual reasoning, instruction following, and spatial output generation. Recent models such as Qwen-VL and InternVL have demonstrated strong grounding ability by generating bounding-box coordinates through language decoding (Wang et al., 2024; Bai et al., 2025a). To process high-resolution inputs, existing MLLMs commonly adopt image tiling, dynamic resolution encoding, or hierarchical visual encoders (Guo et al., 2024; Wang et al., 2024; Ge et al., 2024). These designs improve visual coverage, but their tokens remain organized by resolution, patch layout, or encoder structure rather than grounding-oriented evidence accessibility. Q-REX complements such resolution scaling by reorganizing native visual tokens before language-side decoding, without modifying the visual encoder or adding task-specific detection heads.

2.3 Small-Object Grounding

Small-object grounding in high-resolution UAV imagery is challenging because targets often exhibit weak appearance cues, occupy only a few visual tokens, and appear among dense distractors (Wang et al., 2026; Sun et al., 2025). Successful grounding therefore requires both fine-grained object evidence and global relational context, including scene layout, relative position, and surrounding structures.

Existing approaches mainly improve small-object grounding through local visual search or visual-token reduction. Cropping and zooming methods, such as V*, ZoomEye, and PIZA, recover local details by revisiting candidate regions, but often require multi-step inference and may fragment the global context needed for relational expressions (Wu and Xie, 2024; Shen et al., 2025; Goto et al., 2025). Token pruning and sparsification methods, such as LLaVA-PruMerge and SparseVLM, reduce visual redundancy, but generic compression may discard weak target cues or remove contextual tokens useful for localization (Shang et al., 2025; Zhang et al., 2025). Q-REX instead organizes native visual tokens before decoding, using spatial-semantic graph routing to preserve compact object-like candidates and masked background aggregation to retain contextual memory, after which the MLLM performs language-conditioned disambiguation.

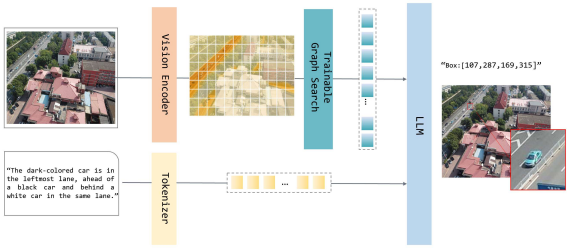

Figure 2: Overview of Q-REX. Q-REX operates directly on Qwen3-VL native visual tokens, selecting compact object-like foreground evidence through query-agnostic graph routing and aggregating unselected tokens into background context. The referring expression is introduced only during final autoregressive box generation.

3 Methodology

3.1 Generative Grounding with Native Visual Tokens

Given an image I and a referring expression Q, referring expression grounding aims to predict the bounding box of the referred object. We denote the target box as $b = (x_1, y_1, x_2, y_2)$, where (x_1, y_1) and (x_2, y_2) are the top-left and bottom-right coordinates, respectively. Following the generative grounding paradigm, the bounding box is serialized as a text sequence and predicted autoregressively by a multimodal large language model (MLLM), without introducing a task-specific detection head.

3.2 Q-REX Overall Framework

Q-REX is a vision-native evidence organization framework built on top of Qwen3-VL. It operates directly on the native visual tokens produced by the frozen visual encoder of the base MLLM. Given

an input image I, we first obtain the visual-token sequence:

$$Z = E_{\text{vis}}(I), \quad Z = \{z_i\}_{i=1}^{N}, \tag{1}$$

where E_{vis} denotes the frozen Qwen3-VL visual encoder, $z_i \in \mathbb{R}^d$ is the i-th native visual token, and N is the number of visual tokens after the native visual-token merging of the base model. Each token is associated with a 2D image-space coordinate p_i according to its position in the visual grid.

Q-REX applies a *Correlate-and-Compress* (CnC) module to reorganize the native token sequence into two complementary parts:

$$E_{\text{fg}}, E_{\text{bg}} = \text{CnC}(Z, P), \tag{2}$$

where $P = \{p_i\}_{i=1}^{N}$, E_{fg} denotes compact object-like foreground evidence, and E_{bg} denotes compressed background context. The concatenation $[E_{\text{fg}}; E_{\text{bg}}]$ is used as the compact visual evidence sequence passed to the language decoder, following the original visual-token interface of the base MLLM.

Importantly, CnC is intentionally query-agnostic. It does not estimate token relevance from the referring expression and does not use any external text–vision scoring module. Instead, it exploits intrinsic spatial adjacency and visual-token similarity to discover coherent object-like foreground candidates, while compressing redundant background tokens into a lightweight contextual memory. The referring expression Q is introduced only during the final autoregressive decoding stage, where the MLLM performs language-conditioned disambiguation and generates the target box:

$$\hat{b} = \text{MLLM}_\theta(Q; [E_{\text{fg}}; E_{\text{bg}}]). \tag{3}$$

3.3 Correlate-and-Compress over Native Visual Tokens

The Correlate-and-Compress (CnC) module converts native visual tokens into a compact evidence sequence. It consists of two components: query-agnostic visual graph routing for identifying object-like token candidates, and masked background aggregation for compressing the remaining tokens into contextual memory.

Spatial-Semantic Graph Construction. Given native visual tokens $Z = \{z_i\}_{i=1}^{N}$ and their grid coordinates $P = \{p_i\}_{i=1}^{N}$, we construct a sparse spatial-semantic graph $G = (V, E)$:

$$E = E_{\text{sp}} \cup E_{\text{sem}}. \tag{4}$$

Here, E_{sp} connects neighboring tokens on the visual grid to preserve local geometric continuity, while E_{sem} links each token to its top-m most similar tokens in a projected feature space, where similarity is computed by cosine similarity after a lightweight linear projection of z_i. Since the graph is built over native merged visual tokens rather than raw patches, the number of nodes remains manageable. Sparse top-m retrieval further keeps graph storage at $O(N(k_{\text{sp}}+m))$, where k_{sp} is the number of spatial neighbors.

Trainable Graph Routing. Each node is initialized by fusing its visual feature with positional information:

$$h_i^0 = \text{MLP}([z_i; \text{PE}(p_i)]). \tag{5}$$

We then perform L layers of message passing over the sparse graph:

$$h_i^{l+1} = \text{FFN}\left(h_i^l + \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^l W_h h_j^l\right), \tag{6}$$

where $\mathcal{N}(i)$ denotes the spatial-semantic neighbors of token i, and α_{ij}^l is computed by a softmax attention over the local graph neighborhood using the projected query and key features of h_i^l and h_j^l. The message-passing cost is $O(LN(k_{\text{sp}} + m)d_g)$, where d_g is the graph hidden dimension. A linear routing head then produces visual-only routing logits:

$$\ell_i = \text{Linear}(h_i^L). \tag{7}$$

These logits are used to preserve compact visual evidence candidates, rather than to resolve the final referent.

We select the top-K_{fg} tokens according to the visual-only routing logits. Let $\mathcal{I}_{\text{fg}}$ denote the hard Top-K index set:

$$\mathcal{I}_{\text{fg}} = \text{TopK}(\{\ell_i\}_{i=1}^{N}, K_{\text{fg}}), \tag{8}$$

$$x_{\text{hard},i} = \mathbf{1}[i \in \mathcal{I}_{\text{fg}}]. \tag{9}$$

To optimize the discrete routing operation, we use a straight-through estimator:

$$x_{\text{soft}} = \sigma(\ell/\tau), \tag{10}$$

$$x_{\text{st}} = x_{\text{hard}} - \text{sg}(x_{\text{soft}}) + x_{\text{soft}}, \tag{11}$$

where $\text{sg}(\cdot)$ stops gradients and τ is the temperature; the hard mask fixes the forward evidence

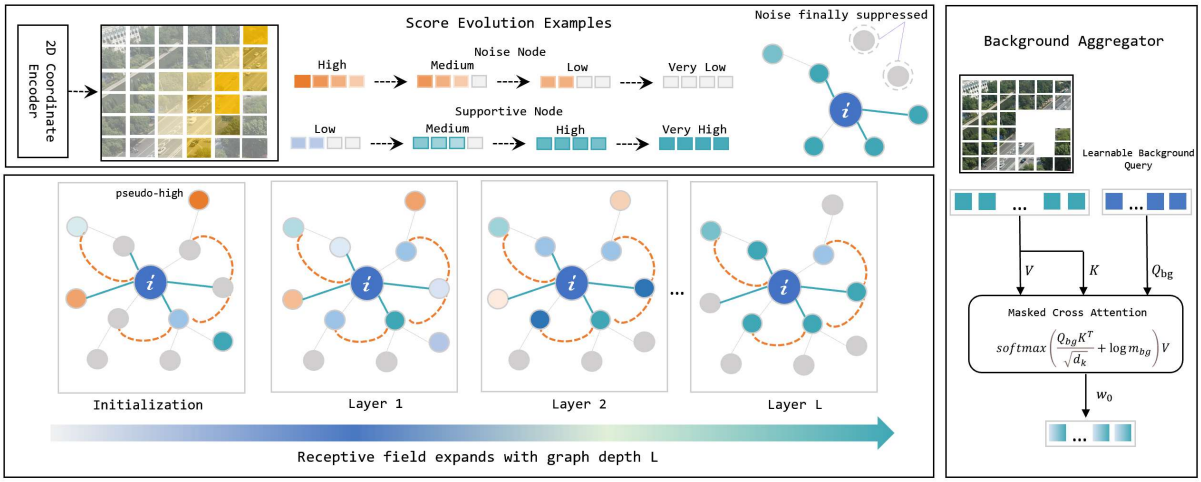

Figure 3: Query-agnostic graph routing in the Correlate-and-Compress module. Spatial edges preserve local geometric continuity, while semantic edges connect visually similar tokens beyond local neighborhoods. Through multi-layer message passing, coherent object-like token clusters are enhanced and selected as compact foreground evidence without using query-conditioned token scores.

budget, while x_{soft} provides a differentiable path for backward gradients.

We apply the straight-through mask to keep the gradient path from the selected evidence tokens to the routing logits:

$$\tilde{z}_i = x_{\text{st},i} z_i, \tag{12}$$

$$E_{\text{fg}} = \text{Order}\left(\{\tilde{z}_i \mid i \in \mathcal{I}_{\text{fg}}\}, P\right). \tag{13}$$

Context Memory Aggregation. Selected evidence alone may be insufficient for UAV grounding, where relational descriptions often depend on surrounding roads, buildings, and scene layouts. We therefore aggregate unselected tokens into a fixed-size background-context memory. Let $m_{\text{bg}} = 1 - x_{\text{st}}$ denote the straight-through background mask. Given K_{bg} learnable background queries U_{bg}, masked cross-attention is computed as:

$$A_{\text{bg}} = \text{softmax}\left(\frac{U_{\text{bg}} K_F^{\top}}{\sqrt{d}} + \log(m_{\text{bg}} + \epsilon)\right), \tag{14}$$

$$E_{\text{bg}} = A_{\text{bg}} V_F, \tag{15}$$

where $K_F = ZW_K$, $V_F = ZW_V$, and ϵ is a small constant for numerical stability. The mask is broadcast to all background queries, so selected foreground tokens are suppressed when forming the background memory, while gradients are propagated through the soft routing surrogate in the backward pass. This encourages E_{bg} to summarize complementary scene context while allowing the final grounding objective to supervise the graph router.

3.4 Training Strategy

Q-REX is trained in a parameter-efficient manner. We freeze the visual encoder and the backbone parameters of the base MLLM, and update only the Correlate-and-Compress module and the LoRA parameters inserted into the language decoder. The hard Top-K routing is optimized with the straight-through estimator described above. During the forward pass, CnC provides a fixed-length hard evidence sequence, while during the backward pass, gradients from both the foreground evidence path and the background memory path are propagated to the routing logits through the soft surrogate x_{soft}.

Given the selected evidence tokens E_{fg} and the contextual memory E_{bg}, the model autoregressively predicts the target bounding box as a token sequence. We use the standard language modeling loss on the answer tokens:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T_b} \log p_\theta(y_t \mid y_{<t}, Q, E_{\text{fg}}, E_{\text{bg}}), \tag{16}$$

where $\{y_t\}_{t=1}^{T_b}$ denotes the tokenized ground-truth bounding box sequence. Prompt tokens are masked out during loss computation. No ranking loss or IoU regression loss is introduced; the visual evidence router is optimized through the final generative grounding objective.

Method	Visual Feature	Text Feature	Val		Test	
			Acc@0.5	mIoU	Acc@0.5	mIoU
FAOA (Yang et al., 2019)	DarkNet-53	BERT	23.10	10.21	13.08	6.91
RSC (Yang et al., 2020)	DarkNet-53	BERT	29.61	16.38	17.59	9.01
RefTR (Li and Sigal, 2021)	ResNet-101	BERT	31.22	15.80	22.68	13.46
TransVG (Deng et al., 2021)	ResNet-101	BERT	35.49	20.56	22.00	11.90
VGTR (Du et al., 2022)	ResNet-101	LSTM	35.30	21.99	20.16	10.55
VLTVG (Yang et al., 2022)	ResNet-101	BERT	30.29	14.40	23.52	13.32
SeqTR (Zhu et al., 2022)	DarkNet-53	GRU	37.49	24.45	25.74	12.57
QRNet (Ye et al., 2022)	Swin-S	BERT	33.90	22.10	26.21	11.22
SimREC (Luo et al., 2023)	CSPDarkNet-53	LSTM	31.17	22.09	21.50	12.15
PolyFormer (Liu et al., 2023b)	Swin-B	BERT	42.84	25.50	31.01	16.44
AerialREC (Wang et al., 2026)	Swin-B	LSTM	45.21	29.78	38.13	20.38
Q-REX (Ours)	Dynamic ViT	Qwen3-VL	**53.15**	**45.61**	**65.35**	**52.85**

Table 1: Performance comparison with task-specific referring expression grounding methods on the SkyFind validation and test sets.

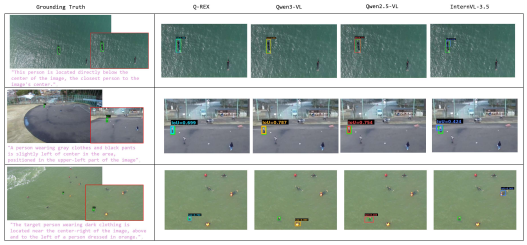

Figure 4: Qualitative grounding examples on SkyFind. Q-REX localizes small targets inDecoupled Weight Decay Regularization high-resolution UAV scenes where the referred objects are surrounded by dense distractors.

4 Experiments

4.1 Experimental Settings

Dataset and Metrics. We evaluate Q-REX on SkyFind (Wang et al., 2026), a large-scale UAV grounding benchmark with diverse scenes, augmented expressions, and a test set from distinct data sources. SkyFind-Train contains approximately 157k expression-box pairs. Following the official preprocessing protocol, we construct multi-scale training samples and group images into four resolution buckets: 768×512, 1280×736, 1344×1024, and 1920×1088. Images are resized with bicubic interpolation according to their assigned buckets, and bounding boxes are scaled accordingly. Evaluation is conducted on SkyFind-Val with 5,000 samples and SkyFind-Test with 16,546 samples. All predictions are mapped back to the original image coordinates. We report Accuracy@0.5 (Acc@0.5) and mean IoU (mIoU).

Model Configuration. We implement Q-REX based on Qwen3-VL-2B-Instruct. The native visual encoder is kept frozen, and LoRA adapters (Hu et al., 2022) with rank $r = 16$ and scaling factor $\alpha = 32$ are inserted into the language decoder. For the Correlate-and-Compress module, the default CnC evidence budget is set to 96 tokens in the main experiments, with $K_{\text{fg}} = 72$ selected evidence tokens and $K_{\text{bg}} = 24$ contextual memory tokens. The graph routing depth is set to $L = 2$. We further analyze the effect of different inference-time evidence budgets in Section 4.3.

Training Protocol. Training is conducted in two stages. In the first stage, we warm up the LoRA adapters and the graph router with a constant learning rate of 1×10^{-4}. In the second stage, we jointly optimize the CnC module and LoRA parameters with a cosine learning rate schedule decaying from 3×10^{-5} to 1×10^{-6}. The model outputs bounding boxes autoregressively in a standard JSON format. We use the AdamW optimizer (Loshchilov and Hutter, 2019) and bfloat16 mixed precision. All experiments are conducted on two NVIDIA A800 80GB GPUs with Distributed Data Parallel (DDP). Additional implementation details are provided in the appendix C.

Method	Visual Feature	Text Feature	SkyFind Test	
			Acc@0.5	mIoU
Grounding-DINO (Liu et al., 2024)	Swin-B	BERT	15.16	8.25
UNINEXT (Yan et al., 2023)	ConvNeXt-L	BERT	12.70	7.40
MiniGPT-V2 (Chen et al., 2023a)	EVA-CLIP-G	LLaMA	18.80	13.64
LLaVA (Liu et al., 2023a)	OpenCLIP-L	Vicuna	19.12	15.75
CogVLM (Wang et al., 2023)	EVA02-CLIP-E	LLaMA	23.67	16.99
Qwen2.5-VL-4B (Bai et al., 2025b)	Dynamic ViT	Qwen2.5	21.10	19.33
Qwen3-VL-2B (Bai et al., 2025a)	Dynamic ViT	Qwen3-VL	53.25	45.87
InternVL3.5-4B (Wang et al., 2025)	InternViT+ViR	InternLM3	64.45	52.85
Q-REX (ours)	Dynamic ViT	Qwen3-VL	**65.35**	**52.85**

Table 2: Performance comparison with open-vocabulary grounding models and general-purpose MLLMs on the SkyFind test set.

4.2 Main Results and Comparison

Main results on SkyFind. Table 1 compares Q-REX with representative task-specific grounding methods on SkyFind. Q-REX achieves the best performance on both validation and test sets, substantially outperforming the strongest prior base-

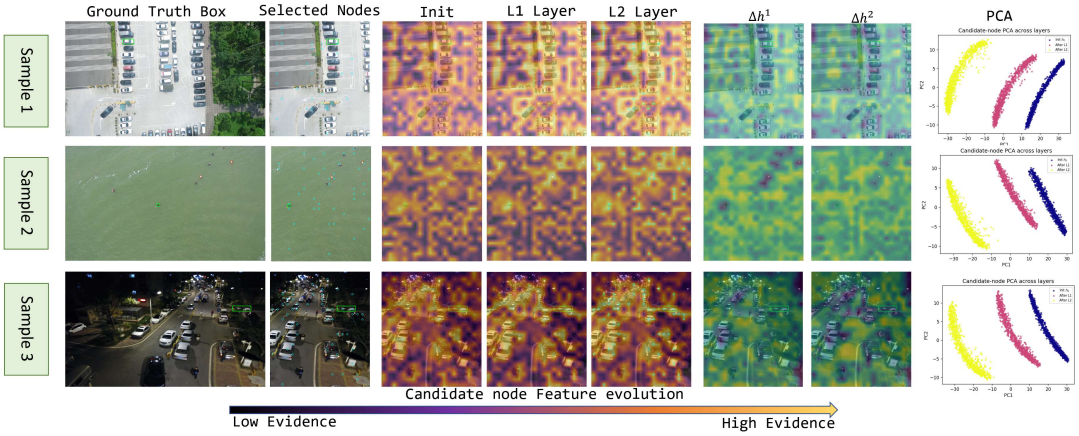

Figure 5: Layer-wise visualization of query-agnostic graph routing. Heatmaps show the evolution of visual-only routing responses across graph layers, while the feature-space visualization illustrates how message passing shifts native tokens toward more coherent object-like clusters.

line AerialREC. Since the compared methods vary in architecture, backbone, and pretraining data, the results are intended to reflect practical grounding performance rather than architecture-only gains. Overall, Q-REX shows that compact evidence organization improves the visual context available for language-conditioned box generation in high-resolution UAV grounding.

Comparison with general-purpose MLLMs. Table 2 compares Q-REX with open-vocabulary grounding models and general-purpose MLLMs. Using the same Qwen3-VL-2B generative interface, Q-REX clearly improves over the direct Qwen3-VL baseline, suggesting that the improvement comes from evidence organization before coordinate generation. It also remains competitive with larger MLLMs while using a compact CnC evidence sequence.

Qualitative results. As shown in Figure 4, Q-REX better localizes small targets under dense distractors and complex backgrounds, especially for expressions involving relative position or scene context.

4.3 Ablation and Analysis

Unless otherwise specified, all ablation variants follow the same SkyFind training protocol, backbone, data split, optimizer, schedule, and LoRA configuration as Q-REX. Only the analyzed component is changed. Thus, the results isolate the effect of visual evidence organization rather than differences in fine-tuning recipes.

Components		Val		Test	
CnC	LoRA	Acc@0.5	mIoU	Acc@0.5	mIoU
		36.92	32.84	53.25	45.87
✓		42.00	37.66	56.50	47.07
	✓	43.70	38.25	61.18	50.44
✓	✓	**53.15**	**45.61**	**65.35**	**52.85**

Table 3: Component ablation of CnC and LoRA on SkyFind. CnC improves visual evidence organization, while LoRA adapts the decoder to the generative grounding format.

Component ablation. Table 3 evaluates the contributions of CnC and LoRA. Both components improve over the direct Qwen3-VL-2B baseline, and their combination performs best. This confirms that CnC and LoRA provide complementary gains rather than simply benefiting from decoder fine-tuning: CnC organizes visual evidence before decoding, while LoRA adapts the decoder to a generative grounding format.

Does graph routing preserve object-level evidence? Since CnC selects evidence without using the referring expression, we examine whether the selected tokens still preserve potential target objects. Table 4 shows that CnC achieves consistently higher object coverage than random selection, uniform sampling, feature-norm selection, and an MLP selector trained under the same protocol. This suggests that graph routing retains a high-recall set of object-like candidates rather than simply selecting visually salient tokens.

Figure 6 further shows that selected tokens are not randomly scattered over large homogeneous regions, but form compact clusters around object-like

Method	Avg. Cov.	Cov@0.5	Center-Hit	Overlap-Hit
Random-K	7.77	5.29	7.72	26.14
Uniform-K	7.29	4.29	7.25	29.95
Feature-norm	52.78	34.28	61.23	80.19
MLP Selector ($L = 0$)	65.90	62.89	70.49	92.80
CnC Selector	**88.64**	**94.71**	**92.34**	**99.86**

Table 4: Compact target coverage analysis on SkyFind-Val. Cov@0.5 measures the percentage of samples where selected foreground token cells cover at least 50% of the ground-truth box. Full coverage results are provided in Appendix A.1.

regions. The cluster-level analysis indicates that 75.12% of selected-node clusters overlap with the ground-truth box, suggesting that CnC organizes native visual tokens into compact target-related evidence rather than isolated high-response fragments.

Effect of spatial-semantic graph construction. We ablate the spatial and semantic edges in the graph selector. Removing either edge type leads to clear performance degradation, while the full spatial-semantic graph performs best. This shows that local spatial continuity and non-local semantic affinity are complementary for coherent evidence organization. Detailed results are provided in Appendix A.2.

Effect of graph routing depth. We study the number of message-passing layers. Moderate routing depth improves grounding performance, while overly deep routing degrades the result, likely due to over-smoothing and excessive background propagation. We therefore use $L = 2$ by default. Full results are shown in Appendix A.3. We further compare inference efficiency with strong MLLM baselines under the same evaluation setting. Q-REX increases average latency only from 2.56s to 2.66s per sample over Qwen3-VL, while achieving comparable performance to InternVL3.5-4B. Full computational results are provided in Appendix B.1.

Effect of evidence budget. We vary the inference-time CnC evidence budget from 32 to 160 tokens. Performance peaks at a compact budget and does not monotonically improve with more retained tokens, suggesting that evidence quality is more important than token quantity. We use 96 tokens as the default budget.

Visualization of graph routing. Figure 5 visualizes layer-wise routing responses and feature-space projections. After message passing, responses become more focused and selected tokens form more coherent object-like clusters, showing that graph routing progressively reorganizes native visual tokens into decoder-friendly evidence candidates.

Figure 6: Selected-node cluster analysis. The selected foreground tokens form compact object-level clusters around small target candidates rather than scattering over large homogeneous backgrounds.

5 Conclusion

We presented Q-REX, a query-deferred evidence organization framework for generative referring expression grounding in high-resolution UAV imagery. Instead of relying on early language-conditioned pruning or iterative cropping, Q-REX reorganizes native MLLM visual tokens before coordinate generation through spatial-semantic graph routing and masked background aggregation. Experiments on SkyFind show consistent improvements over task-specific grounding methods and direct MLLM-based generative grounding. Ablations further indicate that compact foreground evidence organization and decoder-side adaptation are complementary. These results highlight visual evidence organization as a promising direction for generative high-resolution small-object grounding.

Limitations

Q-REX has two main limitations. First, CnC operates on native MLLM visual tokens and is bounded by their spatial granularity; extremely small targets may lose fine-grained details before graph routing. Second, the routing is query-agnostic: it preserves compact object-like candidates before decoding, but does not use the referring expression during evidence selection to disambiguate visually similar candidates. Future work will explore higher-resolution general-purpose visual encoders and tighter vision–language interaction for language-aware evidence preservation and selection.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

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

Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. 2022. Visual grounding with transformers. In *Proceedings of the International Conference on Multimedia and Expo*.

Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. 2024. Convllava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738*.

Kanoko Goto, Takumi Hirose, Mahiro Ukai, Shuhei Kurita, and Nakamasa Inoue. 2025. Referring expression comprehension for small objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21231–21242.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. LLaVA-UHD: An LMM perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

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. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Muchen Li and Leonid Sigal. 2021. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. 2023b. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18653–18663.

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. 2024. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Gen Luo, Yiyi Zhou, Jiamu Sun, Xiaoshuai Sun, and Rongrong Ji. 2023. A survivor in the era of large-scale pretraining: An empirical study of one-stage referring expression comprehension. *IEEE Transactions on Multimedia*, 26:3689–3700.

Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. 2025. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9206–9217.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2025. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22857–22867.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. 2025. ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6602–6618, Suzhou, China. Association for Computational Linguistics.

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

Kunyu Wang, Guanbo Wu, Xueyang Fu, Xingbo Wang, Kean Liu, Xin Lu, Chengjie Ge, Wei Zhai, and Zheng-Jun Zha. 2026. SkyFind: A large-scale benchmark unveiling referring expression comprehension for UAV. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. 2023. UNINEXT: Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15320–15330.

Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. 2022. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9499–9508.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *European conference on computer vision*, pages 387–404. Springer.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4683–4693.

Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. 2022. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15502–15512.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 74840–74857. PMLR.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. 2022. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer.

A Additional Ablation Results

This appendix provides additional ablation studies and analyses that are omitted from the main paper due to space limitations. These experiments further examine the behavior of the proposed Correlate-and-Compress (CnC) module from multiple perspectives, including target coverage, spatial-semantic graph construction, graph routing depth, evidence budget, background aggregation, qualitative hard cases, inference efficiency, and implementation details.Unless otherwise specified, all evaluation results are reported from a single run, and we fix the random seed across experiments to facilitate reproducibility.

A.1 Target Coverage Analysis

Since CnC performs query-agnostic evidence organization before language-conditioned decoding, an important question is whether the selected foreground tokens can preserve the referred target without access to the expression. We therefore evaluate the target coverage of selected foreground tokens with respect to the ground-truth bounding box.

Given a ground-truth box b and the selected foreground token set S_{fg}, each selected native visual token is mapped back to its corresponding image cell r_i. The target coverage is computed as:

$$\text{Coverage}(b, S_{\text{fg}}) = \frac{\text{Area}\left(b \cap \bigcup_{i \in S_{\text{fg}}} r_i\right)}{\text{Area}(b)}.$$

We report the average coverage over the validation set, as well as Cov@0.25, Cov@0.5, and Cov@0.75. Cov@η denotes the percentage of samples for which the selected foreground token cells cover at least η of the ground-truth box. We additionally report Center-Hit, which measures whether at least one selected token center lies inside the ground-truth box, and Overlap-Hit, which measures whether at least one selected token cell overlaps with the ground-truth box.

As shown in Table 5, the full CnC selector consistently achieves higher target coverage than all selection baselines. Compared with the MLP selector without graph routing, CnC improves the average target coverage from 65.90 to 88.64 and Cov@0.5 from 62.89 to 94.71. The gain is particularly pronounced under the stricter Cov@0.75 metric, where CnC improves the coverage from 48.41 to 81.80. These results suggest that graph routing does not merely select isolated target-related tokens, but helps preserve a more complete set of object-level visual evidence. In contrast, random and uniform selection obtain substantially lower coverage, indicating that the improvement is not simply due to retaining a fixed number of tokens.

These results support our query-agnostic evidence organization design. The selector is not intended to identify the final referent from the expression. Instead, it serves as a high-recall visual evidence organizer that preserves compact object-like candidates before the language decoder performs expression-conditioned disambiguation.

A.2 Spatial-Semantic Edge Ablation

CnC constructs a spatial-semantic graph over native visual tokens. Spatial edges encode local geometric continuity, while semantic edges connect visually similar tokens across non-local regions. We ablate these two edge types to examine their respective contributions.

Table 6 shows that removing either spatial or semantic edges leads to a substantial performance drop. Without spatial edges, the model loses local structural continuity, and the test Acc@0.5 decreases from 65.35 to 55.00. Without semantic edges, visually similar evidence distributed across non-local neighborhoods cannot be effectively connected, reducing the test Acc@0.5 to 55.60. The full graph achieves the best performance, suggesting that local spatial structure and non-local semantic affinity provide complementary cues for organizing foreground evidence.

A.3 Graph Routing Depth

We further study the effect of graph routing depth. The number of message-passing layers controls the effective receptive field of the selector. A shallow graph may be insufficient to aggregate object-level evidence, whereas an overly deep graph may oversmooth token representations or introduce irrelevant background information.

As shown in Table 7, graph routing depth exhibits a non-monotonic effect. Without message passing ($L = 0$), the selector scores each token independently and achieves 62.78 Acc@0.5. Increasing the depth to $L = 2$ improves the performance to 65.35, indicating that multi-hop spatial-semantic propagation helps refine object-like evidence. Further increasing the depth to $L = 3$ and $L = 4$ brings only marginal gains, while $L = 5$ results in a clear performance drop. This trend suggests that moderate graph propagation expands the effective receptive field of evidence selection, whereas

Method	Avg. Cov.	Cov@0.25	Cov@0.5	Cov@0.75	Center-Hit	Overlap-Hit
Random-K	7.77	11.09	5.29	2.79	7.72	26.14
Uniform-K	7.29	10.07	4.29	2.07	7.25	29.95
Feature-norm	52.78	71.77	34.28	24.88	61.23	80.19
MLP Selector ($L = 0$)	65.90	85.66	62.89	48.41	70.49	92.80
CnC Selector	**88.64**	**99.34**	**94.71**	**81.80**	**92.34**	**99.86**

Table 5: Target coverage analysis on SkyFind-Val. Cov@η denotes the percentage of samples where the selected foreground token cells cover at least η of the ground-truth box.

Variant	Val Acc@0.5	Val mIoU	Test Acc@0.5	Test mIoU
w/o Spatial Edge	43.15	34.09	55.00	49.09
w/o Semantic Edge	45.00	35.61	55.60	48.51
Full Q-REX	**53.15**	**45.61**	**65.35**	**52.85**

Table 6: Ablation of spatial-semantic graph construction. Spatial edges preserve local geometric continuity, while semantic edges capture non-local visual affinity.

Graph Layers	Test Acc@0.5	Test mIoU
$L = 0$	62.78	50.86
$L = 1$	60.92	48.95
$L = 2$	65.35	52.85
$L = 3$	66.00	53.15
$L = 4$	**66.18**	**53.27**
$L = 5$	63.29	49.63

Table 7: Effect of graph routing depth on SkyFind-Test. Moderate graph depth improves grounding accuracy, while excessive depth leads to performance degradation.

Evidence Budget	Acc@0.5	mIoU
32	51.18	44.90
64	52.48	44.67
96	53.15	45.61
128	52.48	44.90
160	52.12	44.67

Table 8: Inference-time evidence budget analysis. Q-REX achieves strong performance with a compact evidence budget, and larger budgets do not necessarily improve grounding accuracy.

Variant	Val Acc@0.5	Val mIoU	Test Acc@0.5	Test mIoU
Foreground-only	49.60	42.87	61.00	46.63
Mean-pooling BG	51.00	41.09	63.00	50.29
Unmasked BG Attention	52.12	42.29	64.50	51.28
Masked BG Attention	**53.15**	**45.61**	**65.35**	**52.85**

Table 9: Ablation of background aggregation. Masked background aggregation preserves complementary scene context under a compact token budget.

excessive routing may over-smooth token features and reintroduce background interference. We therefore use $L = 2$ as the default setting in the main experiments to balance accuracy and efficiency.

A.4 Evidence Budget Analysis

We analyze the inference-time evidence budget by varying the number of CnC evidence tokens. This experiment examines whether increasing the retained visual evidence consistently improves generative grounding performance.

Table 8 shows that increasing the token budget from a very small value improves grounding performance, indicating that a sufficient amount of foreground evidence is necessary for preserving tiny targets. However, the performance does not improve monotonically as the budget further increases. This suggests that redundant visual tokens may reintroduce background interference, while compact and well-organized evidence is more beneficial for coordinate generation. We therefore use 96 tokens as the default CnC evidence budget in the main experiments, which provides a strong high-recall setting while maintaining a compact visual context.

A.5 Background Aggregation Ablation

Foreground evidence alone may be insufficient for UAV referring expression grounding, since many expressions depend on surrounding roads, buildings, or relative layouts. Therefore, CnC compresses unselected tokens into a fixed-size background memory. We compare different background aggregation strategies.

Removing background memory evaluates whether compact foreground evidence alone is sufficient for grounding. Mean-pooling background tokens provides a simple global context baseline, while unmasked background attention allows background queries to attend to all visual tokens, including selected foreground tokens. In contrast, masked background attention explicitly prevents background queries from redundantly attending to selected foreground tokens, encouraging the background memory to summarize complementary scene context. This design is expected to be especially useful for relational expressions and UAV scenes with dense surrounding context.

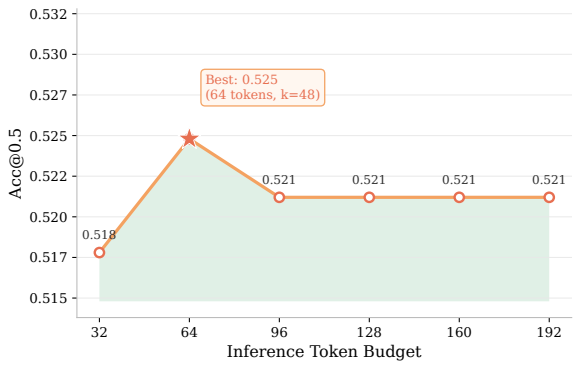

Figure 7: Effect of inference-time evidence budget on SkyFind validation performance.

Method	Latency (s) ↓	Peak Mem. (GB) ↓	Acc@0.5 ↑	mIoU ↑
Qwen3-VL	2.5597	5.2048	53.25	45.87
InternVL3.5-4B	1.8782	11.0718	64.45	52.85
Q-REX	2.6623	6.1531	65.35	52.85

Table 10: Accuracy–efficiency comparison with MLLM-based grounding baselines on SkyFind-Test.

B Hard Sample Analysis

We provide additional qualitative examples on SkyFind in Figure 8. These examples include challenging cases with tiny targets, weak visual cues, dense same-class distractors, and ambiguous surrounding context. Q-REX is able to preserve compact object-like evidence candidates while retaining contextual memory for language-conditioned disambiguation.

Nevertheless, failure cases may occur when the target is visually indistinguishable from nearby distractors or when the referring expression requires very fine-grained relational reasoning beyond the preserved contextual memory. These cases suggest that further improving the interaction between visual evidence organization and language-side reasoning remains an important direction for future work.

B.1 Efficiency Analysis

We report the inference efficiency of Q-REX and representative MLLM baselines under the same evaluation setting. Latency denotes the average end-to-end inference time per sample, including visual encoding, evidence organization if applicable, and autoregressive box generation. Peak memory denotes the maximum GPU memory allocated during inference.

C Implementation Details

C.1 Dataset and Preprocessing

We provide additional details on the data preprocessing and training configuration used in the main experiments. SkyFind-Train contains approximately 157k unique expression-box pairs. Following the official preprocessing protocol, we construct multi-scale training samples to cover the large scale variation in UAV imagery.

To balance visual resolution and GPU memory usage, training images are grouped into four resolution buckets. Images are resized with bicubic interpolation according to their assigned bucket, and bounding boxes are linearly scaled to the resized image coordinates. The bucket configuration is shown in Table 11. During evaluation, all predicted boxes are mapped back to the original image coordinates before computing metrics.

Bucket ID	Resolution	Samples	Proportion
Bucket-1	768×512	66,446	22.0%
Bucket-2	1280×736	90,592	30.0%
Bucket-3	1344×1024	36,208	12.0%
Bucket-4	1920×1088	108,601	36.0%

Table 11: Configuration of multi-scale resolution buckets used during training.

C.2 Supervision Format and Evaluation

During training and inference, bounding boxes are serialized into a standard JSON-style answer format:

$$\texttt{\{"bbox_2d": [}x_1\texttt{, }y_1\texttt{, }x_2\texttt{, }y_2\texttt{], "label": ""\}}. \tag{17}$$

This format follows the generative grounding interface of the base MLLM and does not require a task-specific detection head. During loss computation, all prompt tokens are masked out by setting their labels to -100. The language modeling loss is computed only on the generated answer tokens, including the end-of-sequence token.

For evaluation, the generated coordinate sequence is parsed into a bounding box and mapped back to the original image coordinate system. We report Accuracy@0.5 (Acc@0.5), which measures the percentage of predictions whose IoU with the ground-truth box is greater than 0.5, and mean IoU (mIoU), which measures the average localization quality across all samples.

Figure 8: Additional qualitative examples on SkyFind. These cases involve tiny targets, weak visual cues, dense distractors, or ambiguous surrounding context. Q-REX preserves compact evidence candidates and contextual memory for language-conditioned disambiguation.

C.3 Model and Hyperparameters

We implement Q-REX based on Qwen3-VL-2B-Instruct. The native visual encoder and the backbone parameters of the base MLLM are frozen. Q-REX updates only the Correlate-and-Compress (CnC) module and the LoRA parameters inserted into the language decoder. CnC operates on the pooler-level native visual tokens used by the base MLLM, while other internal visual features are kept unchanged to preserve compatibility with the original generative interface.

Table 12 summarizes the main hyperparameters. The default CnC evidence budget is 96 tokens, including $K_{\text{fg}} = 72$ selected evidence tokens and $K_{\text{bg}} = 24$ contextual memory tokens. The graph router uses two message-passing layers and connects each token to its top-$m = 4$ semantic neighbors in addition to spatial neighbors.

LoRA Configuration	
Target Modules	q, k, v, o_proj
Rank (r)	16
Scaling Factor (α)	32
Dropout Rate	0.05
Bias	None
CnC Configuration	
Graph Hidden Dimension	512
Message-Passing Layers	2
Semantic Top-m	4
Positional Embedding Dim.	64
Dropout Rate	0.1
Total Evidence Budget	96
Selected Evidence Tokens	72
Contextual Memory Tokens	24
STE Temperature (τ)	1.0

Table 12: Hyperparameter settings for the LoRA adapters and the Correlate-and-Compress module.

C.4 Optimization and Hardware Cost

Training is performed in two stages. In the first stage, we warm up the LoRA adapters and the graph router for 10,000 steps with a constant learning rate of 1×10^{-4}. In the second stage, we jointly optimize the CnC module and LoRA parameters using a cosine learning rate schedule, decaying from 3×10^{-5} to 1×10^{-6}.

We use the AdamW optimizer with a weight decay of 1×10^{-2} and momentum parameters $\beta = (0.9, 0.95)$. Gradients are clipped to a maximum norm of 1.0. To reduce memory usage, training uses bfloat16 mixed precision and gradient checkpointing. The micro-batch size is set to 1 per GPU, and gradient accumulation is used to obtain larger effective batch sizes for different resolution buckets.

All experiments are conducted on two NVIDIA

A800 GPUs with 80GB memory using Distributed Data Parallel (DDP). The first-stage warm-up takes approximately 60 single-GPU hours, and the second-stage joint training takes approximately 140 single-GPU hours. The full training cycle therefore takes around 200 GPU hours. The final checkpoint contains only the CnC weights and LoRA adapter weights, making the trained parameters lightweight relative to the frozen base MLLM.

D Artifact Use and Responsible Scope

This work uses existing scientific artifacts, including the SkyFind benchmark, the Qwen3-VL-2B-Instruct pretrained model, LoRA, and the compared grounding or multimodal baselines. We cite the original papers or official model reports for all datasets, models, and methods used in our experiments. The dataset statistics, preprocessing protocol, train/validation/test usage, evaluation metrics, and implementation details are reported in Section 4.1 and Appendix C; we do not introduce an additional data collection procedure beyond the existing benchmark.

We use SkyFind solely for non-commercial academic research, following its official terms of use. The dataset is released under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which restricts commercial use and redistribution of modified or derived versions without authorization. Accordingly, we do not redistribute SkyFind images, annotations, or repackaged dataset files. Qwen3-VL-2B-Instruct is used as the pretrained MLLM backbone under its Apache-2.0 license. If our code is released, it will contain only our own implementation, configuration files, and scripts; users will be required to obtain third-party datasets and pretrained weights from their official sources and comply with the corresponding licenses.

The artifacts introduced by this work are intended for research on high-resolution referring expression grounding. They should not be interpreted as a deployed UAV perception or surveillance system, nor should they be used for safety-critical decision-making. Our experiments are conducted on object localization from benchmark images and referring expressions, without collecting user data, linking samples to real-world identities, or inferring protected attributes. For the qualitative examples shown in the paper and appendix, we manually avoid displaying samples that visibly reveal personally identifiable information or offensive content. Since UAV imagery may naturally contain public outdoor scenes, complete removal of all background-sensitive visual context is not always possible without altering the benchmark; nevertheless, our method and analysis are limited to localization accuracy and visual evidence organization rather than identity, demographic, or behavioral inference.

E Use of AI Assistants

AI-based writing assistants were used only for language polishing, grammar correction, wording refinement, and LaTeX editing during manuscript preparation. All research ideas, method design, experiments, analyses, and final claims were developed and verified by the authors. No AI assistant was used to generate data, fabricate results, or make scientific conclusions. The authors take full responsibility for the manuscript.

F Ethical Considerations

This work studies referring expression grounding in high-resolution UAV imagery using an existing public research benchmark. We do not collect new images, user data, or personally identifiable information. The method is designed for offline academic evaluation of object localization and does not identify individuals, infer protected attributes, track people, or analyze personal behavior.

UAV imagery may contain public outdoor scenes and potentially sensitive background context. We therefore avoid displaying qualitative examples that visibly reveal personally identifiable information or offensive content. We also acknowledge the potential misuse of UAV grounding systems in surveillance or safety-critical applications. Q-REX should be regarded as a research prototype rather than a deployed UAV perception system, and any real-world use should require careful legal, ethical, and domain-specific review.

All datasets, pretrained models, and code artifacts should be used in accordance with their original licenses and terms of use.