FOCUS ON WHAT REALLY MATTERS IN LOW-ALTITUDE GOVERNANCE: A MANAGEMENT-CENTRIC MULTI-MODAL BENCHMARK WITH IMPLICITLY COORDINATED VISION-LANGUAGE REASONING FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-altitude vision systems are becoming a critical infrastructure for smart city governance. However, existing object-centric perception paradigms and loosely coupled vision–language pipelines are still difficult to support management-oriented anomaly understanding required in real-world urban governance. To bridge this gap, we introduce **GovLA-10K**, the **first** management-oriented multi-modal benchmark for low-altitude intelligence, along with **GovLA-Reasoner**, a unified vision-language reasoning framework tailored for governance-aware aerial perception. Unlike existing studies that aim to exhaustively annotate all visible objects, GovLA-10K is deliberately designed around functionally salient targets that directly correspond to practical management needs, and further provides actionable management suggestions grounded in these observations. To effectively coordinate the fine-grained visual grounding with high-level contextual language reasoning, GovLA-Reasoner introduces an efficient Spatially-aware Grounding Adapter (SGA) that implicitly coordinates discriminative representation sharing between the visual detector and the large language model (LLM). Different from existing adapters that primarily focus on global embedding alignment, our SGA is specifically designed to compress and aggregate multi-stream grounding-aware representations, thereby preserving fine-grained spatial cues while enabling their effective integration into the language reasoning process. Extensive experiments indicate that our GovLA-Reasoner effectively improves performance while avoiding the need of fine-tuning for any task-specific individual components. We believe our work offers a new perspective and foundation for future studies on management-aware low-altitude vision–language systems.

1 INTRODUCTION

1.1 BACKGROUND

Low-altitude vision systems, enabled by unmanned aerial vehicles (UAVs) and deep learning techniques, are increasingly deployed in recent years Jiang et al. (2023); Xiang et al. (2026); Shao et al. (2022); Zou et al. (2026). Compared with conventional ground-based perception systems, low-altitude platforms offer a unique vantage point with broader spatial coverage and real-time awareness, making them particularly valuable for large-scale urban management and emergency response Wang et al. (2025b); Yuhong et al. (2025); Wang (2025); Jiang et al. (2025); Hu et al. (2026).

Existing research in low-altitude predominantly targets general-purpose visual perception, such as image classification Al Mazroa et al. (2025); Zhang et al. (2025); Dalai et al. (2026), object detection Alshehri et al. (2025); Zhou & Wei (2025); Nguyen et al. (2026), object tracking Yao et al. (2025); Liu et al. (2025); Ye et al. (2026), and semantic segmentation Yang et al. (2025b); Huang et al. (2025); Li & Huang (2026). These studies aim to improve the completeness and accuracy of visual recognition, and have established the strong technical foundation for understanding low-altitude

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.

scenes Ren et al. (2016); Cai & Vasconcelos (2018); Ge et al. (2021); Zhu et al. (2020); Zhang et al. (2022). To support this paradigm, most publicly available low-altitude datasets adopt a generic, detection-oriented annotation strategy that labels all visible interest objects in a scene Hsieh et al. (2017); Bozcan & Kayacan (2020); Du et al. (2019); Yu et al. (2020); Shermeyer et al. (2021); Sun et al. (2022). Such benchmarks are highly suitable for full-scene coverage, enabling effective and comprehensive visual understanding.

Particularly, with the rapid emergence of large language models (LLMs) and large vision-language models (VLMs), the paradigm of low-altitude scene understanding is undergoing a significant transformation. Unlike conventional task-specific models that rely on limited-scale annotated datasets, these foundation models are pretrained on massive and diverse multimodal corpora, endowing them with strong generalization and knowledge transfer capabilities. As the result, many LLM/VLM-based systems Dai et al. (2023); Zhu et al. (2023); Team et al. (2023; 2024); Bai et al. (2025b); An et al. (2025); Zhu et al. (2025); Wang et al. (2025a); Bai et al. (2025a); Team (2026) demonstrate impressive zero-shot performance in low-altitude scenarios, which can accurately recognize common aerial targets and provide semantically coherent interpretations without requiring task-specific optimization.

1.2 Challenge

However, the requirements of real-world low-altitude management fundamentally differ from exhaustive visual perception. In practical urban governance, the target is not to recognize all objects indiscriminately, but to selectively identify abnormal, risky, or regulation-violating situations, for example, illegally parked vehicles, construction safety hazards, or sanitation-related risks. This shift from comprehensive recognition to selective anomaly confirmation demands higher-level semantic interpretation and situational awareness, which go beyond the scope of traditional object-centric perception. Therefore, the key challenge in real-world low-altitude management does not lie in further pushing the lower bound of identification accuracy, which has already been well addressed by existing visual detection models. Instead, the bottleneck arises from the lack of a management-oriented benchmark dataset that explicitly guide models which objects require attention and why they matter in urban governance.

Beyond this gap, another critical limitation in current low-altitude systems lies in the insufficient coordination between visual detectors and VLMs. Specifically, since VLMs are unlikely to accurately recognize designated targets in every low-altitude urban scenario, deploying scenario-optimized visual detectors to compensate for this limitation is a practical solution. However, existing pipelines generally adopt a loosely coupled paradigm, in which the predicted bounding boxes and images are converted into structured, fine-grained prompts for VLMs, with the aim of achieving more precise object grounding and scene understanding. However, this explicit coordination paradigm inevitably introduces non-negligible side effects. The grounding information extracted by vision detectors must be reinterpreted, structured, and transmitted across heterogeneous stages, during which intermediate representations are prone to information loss and error accumulation. As the result, even minor inaccuracies in visual localization or contextual abstraction at the grounding stage can significantly impact subsequent cross-modal interactions, ultimately leading to substantial degradation in downstream language-based understanding and decision-making performance. Therefore, introducing a more integrated and principled coordination mechanism between visual perception and language reasoning still needs more research.

1.3 Solution

To address these limitations, we propose a coordinated benchmark-framework approach that jointly redefines data annotation and model architecture to satisfy the practical demands of real-world low-altitude management.

On benchmark side, we introduce **GovLA-10K**, the first multimodal dataset tailored for management-oriented low-altitude scenarios. In contrast to existing low-altitude datasets that primarily emphasize exhaustive object-level annotations for generic visual perception, GovLA-10K is deliberately constructed from the perspective of urban governance, focusing on functionally salient targets associated with abnormal, risky, and regulation-violating situations that truly require managerial attention. Beyond selective visual annotations, GovLA-10K further provides aligned language-

based scene descriptions and management suggestions grounded in visual evidence. By shifting the supervision objective from full-scene recognition to selective, semantics-driven event understanding, our GovLA-10K establishes a solid data foundation for low-altitude vision systems.

Beyond benchmark construction, we further propose **GovLA-Reasoner**, a unified vision–language reasoning framework that establishes a more integrated and principled coordination mechanism between visual perception and language reasoning, overcoming the limitations of loosely coupled detector–VLM pipelines. Specifically, instead of explicitly converting localized predictions into structured prompts, GovLA-Reasoner introduces an efficient Spatially-aware Grounding Adapter (SGA) that effectively aggregates discriminative feature representations from the visual detector and further directly feeds them into LLM for high-level contextual reasoning. This implicit feature-space coordination enables end-to-end interaction between vision and language, avoiding fragile intermediate representations and reducing information loss. Peculiarly, unlike existing commonly used adapters such as Q-Former Li et al. (2023) or LLaVA-style Projector Liu et al. (2023) that primarily focus on global embedding alignment, these designs are not well suited for handling the multi-stream localization features produced by detectors. In such representations, complementary spatial cues are distributed across heterogeneous feature sources. Our SGA is specifically designed for localization-aware feature integration. It compresses and coordinates multi-stream grounding features while preserving spatially discriminative cues, enabling structured detection representations to be seamlessly transformed into language-compatible embeddings.

Notably, our GovLA-Reasoner is efficient and deployment-friendly. It delivers effective performance gains without fine-tuning either the visual detector or the language model. Besides, compared with existing adapters, the adopted SGA achieves a better trade-off between performance and computational cost. These two advantages make it well suited for resource-constrained low-altitude intelligence applications. In summary, our work fills a critical gap in low-altitude vision research, with the following key contributions:

- **Benchmark Contribution:** We introduce GovLA-10K, the first management-oriented multimodal benchmark for low-altitude vision systems, shifting the research focus from exhaustive object recognition to selective and actionable understanding of abnormal and risky events.
- **Framework Contribution:** We propose GovLA-Reasoner, a unified vision–language reasoning framework that harmonizes visual perception and language reasoning via implicit feature-space coordination, effectively mitigating error accumulation and information loss inherent in existing loosely coupled pipelines.
- **Practical Significance:** By jointly leveraging GovLA-10K and GovLA-Reasoner, we present an efficient and deployment-friendly solution for low-altitude intelligence. Our framework achieves effective performance gains while better controlling the computational resource cost, establishing a new paradigm for governance-oriented aerial vision systems.

2 BENCHMARK: GOVLA-10K

As mentioned in Section 1, one major limitation of existing low-altitude system research is the lack of management-oriented benchmark that can explicitly guide models to identify which objects deserve attention and why they are important in the context of urban governance. To address this gap, we construct GovLA-10K dataset. The detailed information of the dataset is presented as follows.

2.1 DATA COLLECTION AND ANNOTATION

Data Collection. The construction of GovLA-10K is guided by the practical demands of real-world low-altitude management, where UAVs act as the primary sensing platform. We therefore focus on collecting low-altitude UAV images that cover diverse urban living environments and governance-relevant scenarios, including complex urban layouts, construction activities, traffic conditions, and sanitation-related phenomena.

To ensure both diversity and realism, GovLA-10K is built upon the data collected from two complementary sources: (1) large-scale crawling of publicly available low-altitude aerial images from online platforms and UAV communities, and (2) in-house UAV flights along predefined routes over

representative urban areas such as residential communities, commercial districts, construction zones, and traffic-intensive regions. In total, we initially gather about 16,000 raw UAV images. All images are uniformly cropped into 512×512 patches and further filtered to remove low-quality, redundant, or weakly relevant samples. After screening, 10,572 high-quality images are retained.

Based on extensive investigation of urban management practices and the visual distribution of collected data, we identify nine functionally salient categories that directly correspond to common low-altitude governance concerns, including: illegally parked vehicle, construction debris, construction fencing, brick pile, aggregate pile, construction worker, ground litter, overflowing trash bin, and scaffolding. These categories deliberately emphasize abnormal, risky, or regulation-related targets, rather than exhaustive object inventories.

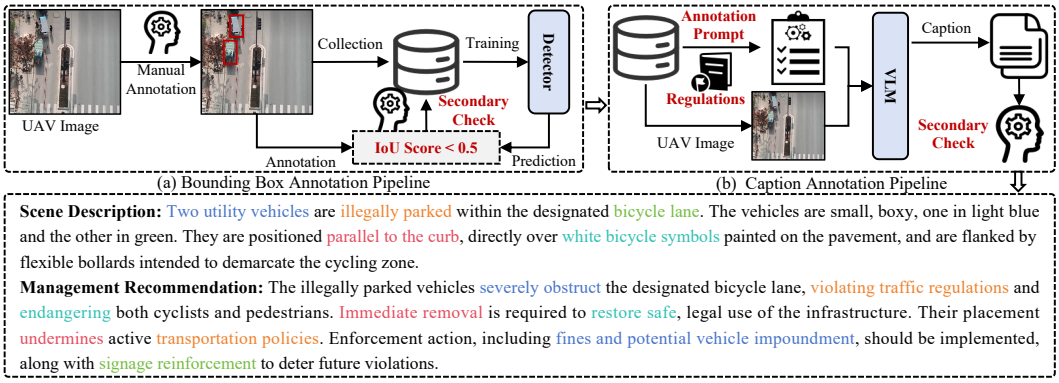


Figure 1: Annotation pipeline of GovLA-10K. Our GovLA-10K adopts a two-stage semi-automatic annotation pipeline. (a) Bounding box annotation stage: interest regions are manually annotated and cross-verified using a powerful detection model to ensure annotation accuracy. (b) Caption annotation stage: verified box annotations and original images are converted into structured, fine-grained prompts and fed into the VLM to generate contextual captions. Peculiarly, all generated captions are also further reviewed by human annotators to maintain high annotation quality. ***For the specially designed prompt template and representative samples from all nine scenarios in GovLA-10K, please refer to Appendix A and Appendix B.***

Data Annotation. To achieve both efficiency and precision, we design a two-stage semi-automatic annotation pipeline that integrates powerful visual detectors with VLMs. As illustrated in Figure 1, in the first stage, several domain experts manually annotate the horizontal bounding boxes (BBoxes) and category labels for the selected targets. After the initial annotation, we train a strong grounding-oriented visual detector based on MM-GroundingDINO Zhao et al. (2024). This detector is then applied to the entire dataset, and its predictions are compared with the original human annotations by computing Intersection-over-Union (IoU) scores. All instances with IoU lower than 0.5 are automatically collected and sent back for secondary manual inspection and correction. This cross-verification mechanism enables rapid large-scale checking while effectively preventing systematic annotation bias and localization errors.

In the second stage, we aim to provide fine-grained semantic descriptions and management-oriented interpretations for each annotated scene. Since the BBox annotations are manually verified and thus do not suffer from the significant localization noise, the explicit prompting paradigm is feasible at this stage. Specifically, the verified BBox annotations and original images are converted into structured, fine-grained prompts, which are jointly fed into the state-of-the-art open-source VLM (Qwen3VL-235B-A22B Bai et al. (2025a)) to generate contextual captions. In addition, the prompt template is carefully constrained to suppress stylistic dominance and maintain consistent, visually grounded descriptions. Each caption includes not only an objective scene description grounded in visual evidence, but also management recommendations aligned with urban governance logic. Moreover, we further incorporate relevant regulations into the prompts to improve the reliability and governance consistency of the generated recommendations and mitigate potential hallucinations. All generated captions are further reviewed by expert annotators and cross-checked against visualized BBoxes to ensure semantic correctness, completeness, and practical relevance.

2.2 Data Statistic

BBox Statistic. Figure 2 (a) and (c) summarize the instance statistics of the nine functional categories in GovLA-10K. Construction Debris (34.0%), Ground Litter (18.8%), and Construction Fencing (18.5%) dominate the dataset, reflecting their prevalence and governance importance in urban low-altitude scenarios, whereas Overflowing Trash Bin (0.3%), Brick Pile (2.0%), and Scaffolding (3.9%) are relatively rare, consistent with real-world distributions. The remaining categories occupy moderate proportions, covering common urban management concerns. We split the dataset into training and testing sets at an approximate 7:3 ratio (7,500 / 3,072 images) using stratified sampling to ensure representative coverage of all categories, especially rare ones.

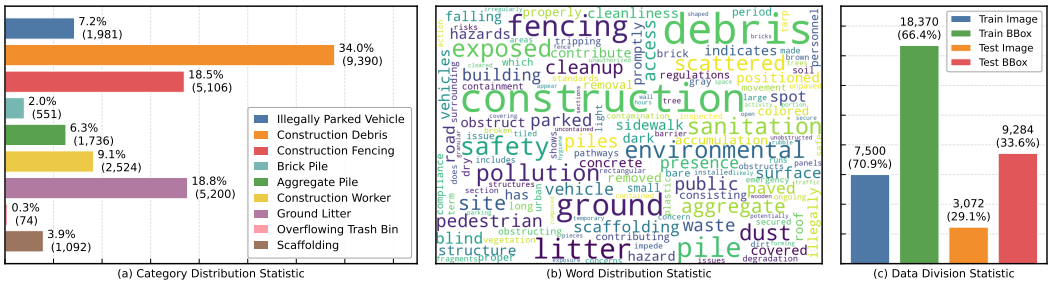


Figure 2: Basic statistics of GovLA-10K. (a) Category-wise instance counts and proportions. (b) High-frequency word distribution in data captions. (c) Numbers of images and bounding boxes in training and test sets.

Caption Statistic. Figure 2 (b) depicts the distribution of high-frequency keywords in generated captions. The results indicate that captions are closely aligned with the annotated target categories and effectively incorporate management-relevant concepts critical for low-altitude urban governance, such as safety, public, and sanitation. Figure 3 further illustrates the distribution of image counts across different caption lengths, which are predominantly concentrated around 60 words. This range allows comprehensive coverage of key targets and scene details while avoiding excessive verbosity, ensuring the overall quality and utility of the captions. These statistics once again provide clear evidence of the effectiveness of our caption annotation pipeline.

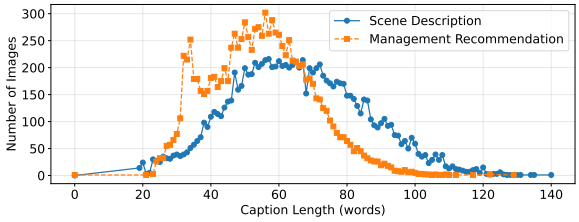


Figure 3: Distribution of image numbers across different caption lengths.

3 Framework: GovLA-Reasoner

Figure 4 illustrates the detail differences between the existing loosely coupled reasoning pipeline and our proposed GovLA-Reasoner. In existing paradigms, the detection model is first optimized for specific scenarios to ensure the precise target grounding. Then, the detection outputs are combined with the original images to construct fine-grained, structured prompts, which are subsequently fed into VLMs to produce detailed descriptions. Although effective, this explicit pipeline inevitably leads to information loss and error accumulation. To address this limitation, our GovLA-Reasoner introduces an efficient Spatially-aware Grounding Adapter (SGA), which can achieve the more coherent end-to-end vision-language coordination. Below is the detailed design of GovLA-Reasoner.

3.1 Stage I: Visual Grounding

As mentioned in Section 1, although VLMs exhibit strong generalization capabilities, they are unlikely to precisely recognize potential targets in every real-world low-altitude scenario. This limitation directly affects the reliability of subsequent language understanding and scene analysis. There-

fore, the primary goal in Stage I is to optimize the vision detector for specific low-altitude scenarios, providing a stable visual foundation for caption generation.

Considering that existing vision detectors have already achieved strong performance, we do not focus on redesigning detector architectures, but on selecting the most suitable model for the current low-altitude management scenario. In this process, our dataset, GovLA-10K, differs significantly from conventional low-altitude datasets: it is designed around practical urban governance needs, focusing only on abnormal or risky targets instead of exhaustive annotations. This ensures that the grounding model is well-adapted to the task. Compared with conventional pure-vision detectors that only take images as input, grounding models offer two key advantages. First, they can take textual prompts as input, naturally enabling selective and open-set recognition, which aligns with both the design of GovLA-10K and the practical requirements of urban management. Second, the generated discriminative features are naturally enhanced through vision-language cross-modal interactions, providing a solid foundation for implicit feature alignment and collaborative reasoning with LLMs.

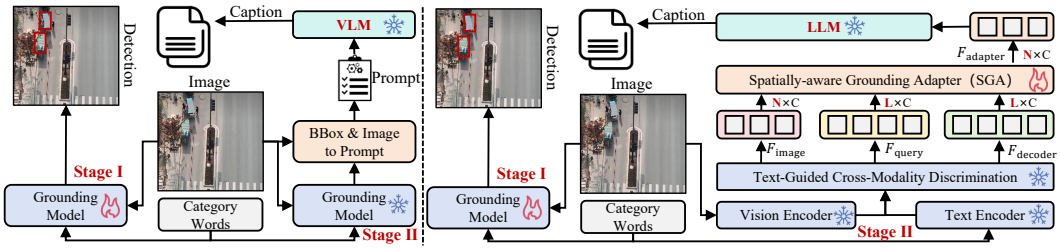


Figure 4: Low-altitude framework overview of the existing pipeline (left) and our proposed GovLA-Reasoner (right). To address the information loss and error accumulation problems caused by the explicit prompt construction, GovLA-Reasoner adopts an implicit feature-space coordination paradigm and further introduces SGA, which adaptively compresses and aggregates the multi-stream discriminative features of grounding model to directly provide the robust feature representation for LLM. Note: snowflakes indicate frozen parameters, whereas flames indicate active (trainable) parameters.

3.2 Stage II: Caption Reasoning

Building upon the precise visual grounding obtained in Stage I, Stage II aims to efficiently receive and exploit grounding-aware visual representations to generate accurate and reliable captions. Instead of following the loosely coupled pipeline that converts detection outputs into structured prompts for VLMs, GovLA-Reasoner introduces an efficient adapter to directly transform grounding features into LLM-compatible representations.

Specifically, existing explicit pipelines typically suffer from three sources of error accumulation: (1) localization errors from the detection model, (2) information loss and ambiguity introduced when converting bounding boxes into the handcrafted structured prompts, and (3) secondary encoding errors when VLM reinterprets these prompts. Since final captions are produced by the LLM conditioned on visual evidence, we eliminate the intermediate prompt construction and re-encoding stages, and instead directly integrate the discriminative grounding features into LLM. This design preserves efficiency while substantially reducing unnecessary transformation-induced degradation, and further enables the end-to-end optimization.

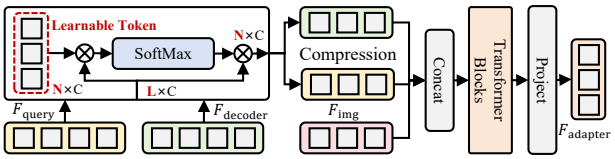


Figure 5: Detailed feature process pipeline of our designed SGA.

Peculiarly, as discussed in Section 1, existing feature adapters commonly focus on global embedding alignment, which is insufficient for grounding-based perception where detectors produce multi-

source localization representations. We therefore propose the Spatially-aware Grounding Adapter (SGA) to adaptively compress and aggregate these signals.

Adapter Inputs. As shown in Figure 5, our SGA receives three groups of grounding features as input: the original image features F_{img}, the text-guided cross-modal query features F_{query}, and the decoded discriminative grounding features F_{decoder}. These representations naturally capture complementary information, including global scene context, language-conditioned alignment, and object-centric discrimination, which together provide rich and structured semantic cues for generating captions.

Token Compression. Since $F_{\text{query}}, F_{\text{decoder}} \in \mathbb{R}^{L\times C}$ are produced by text-guided cross-modal querying and decoding, they represent the responses over all potential interest regions. As a result, these features inevitably contain substantial redundancy and noise (e.g., background activations or false-positive regions). We therefore first perform token-level compression to distill discriminative-relevant semantics. Specifically, we initialize a learnable token set $Q \in \mathbb{R}^{N\times C}(N \ll L)$ with the same dimension as F_{img}. The compressed representations are obtained via global attention modeling:

$$\tilde{F}_{\text{query}} = \text{SoftMax}(Q_{\text{query}}F_{\text{query}}^{\top})F_{\text{query}} \tag{1}$$

$$\tilde{F}_{\text{decoder}} = \text{SoftMax}(Q_{\text{decoder}}F_{\text{decoder}}^{\top})F_{\text{decoder}} \tag{2}$$

Through end-to-end optimization, these learnable tokens can adaptively aggregate informative semantics while suppressing irrelevant activations, and meanwhile help effectively reduce the computational cost.

Token Integration. After compression, we integrate these three heterogeneous discriminative features by concatenating them along the channel dimension:

$$F_{\text{cat}} = \text{Concat}(F_{\text{img}}, \tilde{F}_{\text{query}}, \tilde{F}_{\text{decoder}}) \tag{3}$$

where global appearance cues, language-conditioned responses, and discriminative grounding information are explicitly aggregated into a unified token sequence. This concatenation preserves abundant semantic information and is then fed into the stacked Transformer blocks to perform deep semantic fusion and contextual refinement:

$$F_{\text{fuse}} = \text{Transformer}(F_{\text{cat}}) \tag{4}$$

This design allows cross-source interactions and self-attention-based reasoning among different visual and modal cues, yielding compact and context-aware representations. Finally, a lightweight linear projection is applied to project the fused feature into the embedding space of the LLM:

$$F_{\text{adapter}} = \text{Linear}(F_{\text{fuse}}) \tag{5}$$

so that the adapted tokens can be seamlessly injected into the LLM as visual conditioning tokens, directly supporting high-level caption reasoning without introducing additional prompt engineering or re-encoding stages.

Adapter Optimization. Considering the core purpose of SGA is to transform the grounding-aware visual features into LLM-compatible representations, it is sufficient to optimize SGA solely using the cross-entropy loss between generated captions and ground-truth captions. Formally, the training objective can be written as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{T}\sum_{t=1}^{T}\log P(y_t \mid y_{<t}, F_{\text{adapter}}) \tag{6}$$

where y_t denotes the t-th token in the ground-truth caption, $y_{<t}$ represents the preceding tokens, and F_{adapter} denotes the grounding-aware visual features after adapter transformation. Importantly, all other components, including the grounding model and the LLM itself, are kept frozen during training. This targeted optimization strategy effectively constrains SGA while avoiding perturbation of the pre-trained representations. Moreover, by limiting gradient updates to the lightweight adapter, we can achieve efficient training with minimal computational overhead, ensuring both practicality and scalability for large-scale deployment.

4 EXPERIMENT

Considering the lack of publicly available datasets for low-altitude urban management, we conduct experiments primarily on the proposed GovLA-10K dataset. ***Please see Appendix C for detailed experimental settings.***

4.1 MAIN RESULTS

Table 1: Quantitative comparison between existing mainstream VLMs and our GovLA-Reasoner on GovLA-10K dataset. SD Score, MR Score, and Overall Score denote the semantic consistency of scene description, management recommendation, and overall caption quality, respectively.

Method	Traditional Metrics				Expert Evaluation		
	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SD Score	MR Score	Overall Score
LLaVA-OneVision-1.5-4B An et al. (2025)	7.61	19.10	25.36	4.84	6.52	7.00	6.70
LLaVA-OneVision-1.5-8B An et al. (2025)	6.78	17.25	24.82	2.69	6.62	7.77	7.07
InternVL3-8B Zhu et al. (2025)	5.14	17.68	22.39	2.72	6.45	7.07	6.68
InternVL3.5-4B Wang et al. (2025a)	5.28	17.33	22.61	2.71	6.10	7.12	6.48
InternVL3.5-8B Wang et al. (2025a)	5.64	18.28	22.44	3.01	6.25	7.82	6.93
Qwen2.5-VL-3B Bai et al. (2025b)	8.04	19.21	25.01	5.26	6.30	6.77	6.43
Qwen2.5-VL-7B Bai et al. (2025b)	8.65	19.54	25.63	5.07	7.03	7.07	6.98
Qwen3-VL-4B Bai et al. (2025a)	11.36	23.25	28.92	10.22	7.72	6.45	7.00
Qwen3-VL-8B Bai et al. (2025a)	10.97	21.73	29.44	10.21	7.37	6.90	7.07
GovLA-Reasoner (ours)	**20.31**	**26.63**	**37.97**	**20.31**	**7.87**	**8.18**	**7.83**

Table 2: Comparison of captioning performance and computational cost among different adapters. MSI denotes multi-stream integration capability.

Adapter	MSI	Performance				Computational Cost		
		BLEU-4	METEOR	ROUGE-L	CIDEr-D	Params (M)	FLOPs (G)	Peak Memory (MB)
LLaVA-Projector Liu et al. (2023)	✗	15.54	24.95	34.02	10.56	4.462	9.127	244.020
Q-Former Li et al. (2023)	✗	19.94	26.23	37.72	19.62	21.257	43.520	382.637
SGA (ours)	✓	**20.31**	**26.63**	**37.97**	**20.31**	**17.582**	**36.002**	**317.893**

We first benchmark several representative detectors on GovLA-10K to select a reliable grounding backbone (***detailed results are shown in Appendix D***). MMGroundingDINO achieves the best overall detection performance and is therefore adopted in all subsequent caption reasoning experiments.

Table 1 reports the captioning results using both traditional automatic metrics and expert evaluation. Increasing VLM size does not consistently improve performance: several 8B variants achieve comparable or even inferior results to their smaller counterparts, suggesting that model scaling alone is insufficient to effectively exploit fine-grained grounding information. In contrast, GovLA-Reasoner consistently improves both automatic and expert-evaluated performance, producing responses that are more aligned with reference captions and more accurate for management-oriented reasoning. These gains mainly stem from SGA, which directly integrates multi-stream grounding features into the LLM and jointly exploits spatial, language-conditioned, and object-centric information. Notably, GovLA-Reasoner with only a 4B LLM still outperforms all compared VLMs, including multiple 8B models, demonstrating strong effectiveness and parameter efficiency. ***Qualitative results are provided in Appendix E.***

Table 2 compares both the captioning performance and computational cost of different adapters. SGA achieves the best results across all performance metrics while supporting multi-stream integration, demonstrating its stronger capability in exploiting grounding information. Despite this advantage, its computational cost remains lower than Q-Former, indicating a favorable efficiency-performance trade-off. This efficiency mainly benefits from the token compression mechanism in SGA, which distills discriminative semantics before feature integration. Overall, SGA provides a favorable balance between representation capability and computational efficiency.

4.2 Ablation Study

We mainly conduct ablation studies around the proposed SGA to examine the contribution of its key designs. Considering resource constraints and the consistency of experimental settings, we use traditional automatic captioning metrics for evaluation, which are sufficient to reflect the relative performance changes among different ablation variants.

Depth of SGA. As shown in Table 3, because we deploy stacked Transformer blocks to perform the feature refinement, the block number directly determines how thoroughly heterogeneous grounding features can be aligned and fused. Obviously, using a single layer yields limited gains due to insufficient cross-source interaction, while excessively deep adapters bring marginal or even negative effects. The best overall performance is achieved with two layers, suggesting that moderate semantic fusion is sufficient.

Table 3: Ablation study on the depth of SGA.

Depth	BLEU-4	METEOR	ROUGE-L	CIDEr-D
1	19.54	26.29	37.56	18.68
3	20.28	**26.71**	37.68	18.10
4	19.94	26.25	37.90	18.20
2	**20.31**	26.63	**37.97**	**20.31**

Table 4: Ablation study on the contribution of different discriminative features.

F_{img}	F_{query}	$F_{decoder}$	BLEU-4	METEOR	ROUGE-L	CIDEr-D
✗	✓	✓	17.97	25.00	36.17	14.54
✓	✗	✓	19.62	26.31	37.51	17.81
✓	✓	✗	19.81	26.46	37.45	18.36
✓	✓	✓	**20.31**	**26.63**	**37.97**	**20.31**

Contribution of Discriminative Features in SGA. Because our adapter takes three groups of grounding-aware features as input, we remove each component individually to evaluate its contribution. As shown in Table 4, excluding any feature consistently degrades performance, confirming the rationality of our multi-source design. Notably, removing F_{img} causes the most significant performance drop. This is because the original image feature preserves the most comprehensive and fundamental visual information, providing holistic scene context that cannot be fully recovered from the derived discriminative representations alone.

5 Conclusion

This work aims to address the critical gaps in management-oriented low-altitude intelligence. On the one hand, we introduced GovLA-10K, the first multimodal benchmark designed around functionally salient, governance-relevant targets, shifting low-altitude research from generic perception to selective and semantics-driven reasoning. On the other hand, we propose GovLA-Reasoner, a unified vision–language reasoning framework that implicitly coordinates visual grounding and language generation via the tailored grounding adapter. By directly projecting grounding-aware visual representations into the LLM feature space, GovLA-Reasoner avoids the limitations of the explicit prompt construction pipeline. Extensive experiments demonstrate that GovLA-Reasoner consistently outperforms strong mainstream VLM baselines across all captioning metrics, highlighting both its effectiveness and deployment efficiency. We believe this work establishes a new foundation for management-aware low-altitude vision–language systems and opens promising directions for future research in governance-oriented aerial intelligence.

AI Use Statement

We used generative AI tools for dataset caption generation and English grammar polishing. Specifically, Qwen3VL-235B-A22B was used to generate captions for GovLA-10K, including scene descriptions and management recommendations, based on images and structured prompts incorporating verified bounding-box annotations and relevant regulations. All generated captions were reviewed by expert annotators and cross-checked against the corresponding visual evidence and bounding-box annotations. We also used AI tools to correct grammatical errors in the manuscript. The authors take full responsibility for the final content of the paper and dataset, including all AI-assisted content.

REFERENCES

Alanoud Al Mazroa, Nuha Alruwais, Muhammad Kashif Saeed, Kamal M Othman, Randa Allafi, and Ahmed S Salama. Multi class aerial image classification in uav networks employing snake optimization algorithm with deep learning. *Scientific Reports*, 15(1):23872, 2025.

Mohammed Alshehri, Tingting Xue, Ghulam Mujtaba, Yahya AlQahtani, Nouf Abdullah Almujally, Ahmad Jalal, and Hui Liu. Integrated neural network framework for multi-object detection and recognition using uav imagery. *Frontiers in neurorobotics*, 19:1643011, 2025.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training. In *arXiv*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 8504–8510. IEEE, 2020.

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*, pp. 6154–6162, 2018.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Sagar Dalai, Marco Moreno, Eoin O'Connell, Thomas Newe, and Gerard Dooly. Tinycapsvit: Ultra-lightweight hyperspectral and multispectral image classification for uav edge deployment using capsule vision transformers. *Remote Sensing*, 18(16):2661, 2026.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pp. 0–0, 2019.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pp. 4145–4153, 2017.

Jinpeng Hu, Wei Wang, Yuxiao Liu, and Jing Zhang. Large model in low-altitude economy: Applications and challenges. *Big Data and Cognitive Computing*, 10(1):33, 2026.

Bangju Huang, Junhui Li, Wuyang Luan, Jintao Tan, Chenglong Li, and Longyang Huang. Expanding open-vocabulary understanding for uav aerial imagery: A vision–language framework to semantic segmentation. *Drones*, 9(2):155, 2025.

Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Guodong Guo, Qixiang Ye, Jianbin Jiao, Jian Zhao, and Zhenjun Han. Anti-uav: A large-scale benchmark for vision-based uav tracking. *IEEE Transactions on Multimedia*, 25:486–500, 2023. doi: 10.1109/TMM.2021.3128047.

Xiaoyan Jiang, Khairul Hamimah Abas, and Abdul Rashid Husain. A review: Obstacle perception based on panoramic vision in low-altitude rotorcraft uavs. *Applications of Modelling and Simulation*, 9:213–226, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Sai Li and Shuo Huang. Heterogeneous attention fusion network with multi-branch decoding for uav semantic segmentation. *Journal of King Saud University Computer and Information Sciences*, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shanshan Liu, Xinglin Shen, Shanzhu Xiao, Hanwen Li, and Huamin Tao. A multi-scale feature-fusion multi-object tracking algorithm for scale-variant vehicle tracking in uav videos. *Remote Sensing*, 17(6):1014, 2025.

Thi Lan Nguyen, Cao Truong Tran, and Dinh Tan Nguyen. Etfnet: An efficient transformer-based rgb–ir fusion network for uav object detection. *Information Fusion*, pp. 104658, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Zhenfeng Shao, Gui Cheng, Jiayi Ma, Zhongyuan Wang, Jiaming Wang, and Deren Li. Real-time and accurate uav pedestrian detection for social distancing monitoring in covid-19 pandemic. *IEEE Transactions on Multimedia*, 24:2069–2083, 2022. doi: 10.1109/TMM.2021.3075566.

Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. Rareplanes: Synthetic data takes flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 207–217, 2021.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yizhe Wang, Jie Li, Xiaoguang Yang, and Qing Peng. Uav–ground vehicle collaborative delivery in emergency response: A review of key technologies and future trends. *Applied Sciences*, 15(17): 9803, 2025b.

Ziming Wang. Technological integration in urban emergency management: Challenges, innovations, and pathways to resilient systems. *Science, Engineering and Technology Proceedings*, 1:30–39, 2025.

Tianqi Xiang, Xinghui Xia, Junsong Yuan, and Zhigang Tu. Fre-stformer: A frequency-based spatio-temporal transformer for uav human action recognition. *IEEE Transactions on Multimedia*, pp. 1–13, 2026. doi: 10.1109/TMM.2026.3678499.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Xinghong Yang, Heqing Li, Wei Zhu, and Yi Zuo. Rshrnet: Improved hrnet-based semantic segmentation for uav rice seedling images in mechanical transplanting quality assessment. *Computers and Electronics in Agriculture*, 234:110273, 2025b.

Mufeng Yao, Jinlong Peng, Qingdong He, Bo Peng, Hao Chen, Mingmin Chi, Chao Liu, and Jon Atli Benediktsson. Mm-tracker: Motion mamba for uav-platform multiple object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9409–9417, 2025.

Jingtao Ye, Kexin Zhang, Xunchi Ma, Yuechan Li, Guangming Zhu, Peiyi Shen, Linhua Jiang, Xiangdong Zhang, and Liang Zhang. Breaking smooth-motion assumptions: A uav benchmark for multi-object tracking in complex and adverse conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13594–13603, 2026.

Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1257–1265, 2020.

Huang Yuhong, Ding Haiyu, Chen Weiyan, Kong Luting, Deng Wei, Li Xin, Liu Yang, Wang Guizhen, and Liu Liang. Towards a low-altitude aerial intelligent network: Vision, challenges, and key technologies. *China Communications*, 22(9):1–21, 2025.

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. *arXiv preprint arXiv:2203.03605*, 2022.

Zhen Zhang, Linhuan Jiang, Bo-Hui Tang, Jianchen Liu, Qingwang Wang, Yabin Hu, Liang Huang, and Zhitao Fu. Attention residual hybrid network for unmanned aerial vehicles hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024.

Yu Zhou and Yan Wei. Uav-detr: an enhanced rt-detr architecture for efficient small object detection in uav imagery. *Sensors*, 25(15):4582, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Zizhuang Zou, Mao Ye, Luping Ji, Lihua Zhou, Song Tang, Yan Gan, and Shuai Li. Long-short match for lost control in uav multi-object tracking. *IEEE Transactions on Multimedia*, 28:786–800, 2026. doi: 10.1109/TMM.2025.3632642.

A PROMPT TEMPLATE OF GOVLA-10K

You are an intelligent UAV-based (Unmanned Aerial Vehicle) urban management analysis system. You must strictly follow a two-stage processing workflow:

[INPUT DATA]
1. Summary of detected objects (count statistics):
{detection_summary}
2. Full detection list with bounding boxes:
{detection_all}
3. Some urban public security management regulations
{management regulations}
4. UAV aerial image

[PROCESSING RULES]
■ Stage 1: Scene Description (strictly objective)
- Describe only what is visibly present within the detection boxes; no assumptions or speculation are allowed.
- Group objects of the same category (e.g., "3 illegally parked vehicles" instead of describing each one individually).
- The description must include: object categories, quantities, spatial relationships, and prominent visual features (e.g., color or shape only when clearly visible).
- Forbidden words: may, might, seems, should, suspected, possible, probably.

■ Stage 2: Urban Management Analysis (professional inference)
- Trigger this stage only when **violation categories** are detected:
Illegally Parked Vehicle / Construction Debris / Construction Fencing / Brick Pile / Aggregate Pile / Construction Worker/ Ground Litter / Overflowing Trash Bin / Scaffolding
- Violation assessment principles:
  - Safety issues: blocked fire lanes, sidewalk occupation, illegal operation, risk of falling objects, etc.
  - Environmental issues: garbage exposed for more than 24 hours, dust pollution, sanitation blind spots, etc.

[OUTPUT REQUIREMENTS]
1. First line output must start with the prefix "Scene Description:" followed by Stage 1 content.
2. The second line must start with "Management Recommendation:" and include Stage 2 content.
3. Management recommendations must be grounded in reality and must not be exaggerated.
4. Output must be entirely in English; avoid numbering, bullet points, and technical terms (e.g., bbox).

Figure A.1: Detailed prompt template of GovLA-10K. The contents in {} are dynamically generated and updated by a Python-based parsing pipeline according to each input UAV image, automatically incorporating detection statistics, object-level information, and management regulations to support scalable and consistent caption construction.

As shown in Figure A.1, to support the construction of GovLA-10K, we design a structured and regulation-aware prompt template that emulates an intelligent UAV-based urban management analysis system. This prompt ensures the consistency, reliability, and practical relevance of the generated annotations.

Specifically, the strict two-stage workflow enforces a clear separation between objective perception and professional analysis. Stage 1 constrains the model to produce purely factual, visually grounded scene descriptions based on detected objects and their attributes, significantly reducing hallucinations and standardizing annotation style at scale. Stage 2 is conditionally activated only when predefined violation categories appear, aligning the generation process with real urban management workflows and enabling governance-oriented semantic outputs that connect visual evidence with management reasoning. Furthermore, the integration of urban public security management regulations grounds the generated management recommendations in realistic governance logic, while the explicit output format constraints enhance controllability, automatic quality checking, and large-scale generation efficiency.

Overall, this prompt template acts as a lightweight annotation protocol that embeds objectivity, regulatory awareness, and management-oriented reasoning into GovLA-10K, supporting diverse low-altitude urban perception and governance tasks.



Figure A.2: Some representative scenarios of GovLA-10K. From top to bottom, the examples illustrate: illegally parked vehicles, construction debris, construction fencing, brick piles, and aggregate piles.

B Complete Examples of GovLA-10K

Figures A.2-A.3 present qualitative examples covering all nine categories in the GovLA-10K dataset. As illustrated, GovLA-10K consistently focuses on management-oriented targets in low-altitude urban scenes, rather than indiscriminately annotating all visible objects. This design choice highlights the practical relevance of the dataset for real-world urban governance, where only specific objects and areas are truly critical for management and decision-making. Moreover, the fine-grained captions provided in GovLA-10K go beyond objective visual descriptions of the scenes. They further incorporate management-related interpretations and corresponding recommendations, enabling a closer alignment between visual perception and urban governance requirements. Such governance-aware annotations make GovLA-10K not only a perception benchmark, but also a foundation for high-level reasoning and management-oriented understanding. These qualitative results also intuitively demonstrate the effectiveness of our designed two-stage semi-automatic annotation pipeline,



Figure A.3: Some representative scenarios of GovLA-10K. From top to bottom, the examples illustrate: construction workers, ground litter, overflowing trash bins, and scaffolding.

which is able to reliably identify management-relevant targets and generate structured, governance-aware descriptions at scale.

We believe that GovLA-10K establishes a solid basis for future research on low-altitude urban intelligence, especially for developing models that integrate visual understanding with decision support and governance-oriented reasoning.

C Detail Settings

All models are implemented in PyTorch Paszke et al. (2019) and trained on eight NVIDIA H20 GPUs, with Qwen3-4B Yang et al. (2025a) serving as the LLM in GovLA-Reasoner. We adopt AdamW with a batch size of 16 and an initial learning rate of 1×10^{-4}, using a multi-step learning-rate scheduler. The detection and captioning models are trained for 50 and 25 epochs, respectively. Six representative visual grounding methods, nine mainstream VLMs, and two visual adapters are evaluated under consistent settings. Detection performance is measured by mAP, mAP@50, and mAP@75. For captioning evaluation, we employ both automatic metrics and human assessment. Specifically, BLEU-4 Papineni et al. (2002), ROUGE-L Lin (2004), METEOR Banerjee & Lavie (2005), and CIDEr-D Vedantam et al. (2015) are used as conventional automatic metrics. In addition, we randomly sample 100 test examples and invite 3 domain experts to manually evaluate the generated captions on a 1–10 scale, providing a complementary assessment of overall quality.

D DETECTION RESULTS

Table A.1: Detection performance evaluation of different representative methods on GovLA-10K dataset.

Method	mAP	mAP@50	mAP@75
Faster R-CNN Ren et al. (2016)	35.1	71.1	29.4
Cascade R-CNN Cai & Vasconcelos (2018)	37.2	71.2	34.7
YOLO-X Ge et al. (2021)	31.6	65.3	25.2
Deformable DETR Zhu et al. (2020)	36.6	69.3	34.3
DINO Zhang et al. (2022)	38.6	70.6	37.5
MMGroundingDINO Zhao et al. (2024)	**42.0**	**76.6**	**42.3**

Table A.1 reports the quantitative results of several representative detection frameworks on the proposed GovLA-10K dataset. Instead of focusing on architectural modifications, we deliberately select a set of widely adopted and competitive detectors, including two-stage, one-stage, and transformer-based methods, to provide a comprehensive and objective evaluation. shown in the table, most existing detectors achieve reasonably strong performance, indicating that current detection methods are already sufficient for handling low-altitude urban scenarios. This observation is consistent with our initial motivation that the bottleneck of low-altitude understanding does not mainly lie in pure object detection. Notably, MMGroundingDINO achieves the best results across all metrics. We attribute this advantage to its ability to incorporate textual prompts as inputs, which enables more selective and goal-driven localization. This property is highly aligned with the intrinsic characteristics of GovLA-10K, where interest objects are often defined by management-oriented semantics rather than purely visual categories.

E VISUAL RESULTS

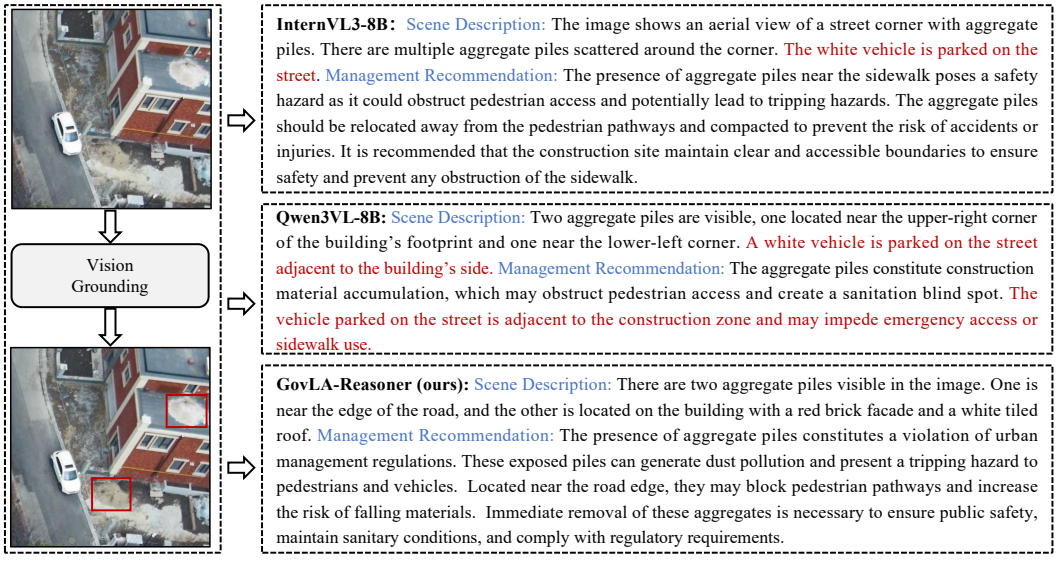


Figure A.4: Caption reasoning result comparison between existing mainstream VLMs and our GovLA-Reasoner. Unexpected contents are highlighted in red for clarity.

Figure A.4 presents the qualitative comparison between several mainstream VLMs and our proposed GovLA-Reasoner on one low-altitude urban scene. Overall, existing VLMs are able to generate semantically rich and fine-grained captions, demonstrating their strong generalization ability and powerful open-world visual understanding. As highlighted by the red text in the figure, the core management concern in this scenario is the identification and analysis of aggregate piles. However,

existing VLMs commonly tend to produce over-generalized or misaligned reasoning results, such as shifting attention to irrelevant objects (e.g., vehicles), or drawing conclusions unrelated to the actual risk source. In contrast, our GovLA-Reasoner always produces more precise, goal-consistent, and practically meaningful descriptions. Therefore, this qualitative comparison once again intuitively verifies the effectiveness of the proposed framework and its advantages in real-world low-altitude scene understanding.