Dual-LoRA Fusion with MLLMs for Drone Threat Assessment in Nuclear Facilities

Juhyung Song[†*], Yonggu Lee[†], Seungoh Seo[†], Taewoo Tak[†], Youngjun Lee[†]
[†]Korea Atomic Energy Research Institute, 111, Daedeok-daero 989 Beon-gil, Yuseong-gu, Daejeon 34057, Korea
[*]Corresponding author: jhsong@kaeri.re.kr

*****Keywords **:** Multimodal Large Language Model, LoRA, Model Fusion, Drone Detection, Protection of nuclear critical facilities

1. Introduction

The physical protection of national critical infrastructure—such as nuclear power plants—must continuously evolve with advancing technology. Illicit incursions by miniaturized, intelligent unmanned aerial vehicles (UAVs, "drones") have emerged as a new security threat that exposes the limits of conventional surveillance systems. A next-generation intelligent surveillance system must therefore go beyond merely detecting what is present to inferring how it behaves and with what intent—in real time.

Modern multimodal large language models (MLLMs) [1] (e.g., Mistral-7B coupled with a vision front end) are compelling candidates for these requirements, as they natively process images and language. However, deploying MLLMs in real security settings faces a core dilemma between the specialization and generalization of knowledge—an Acuity–Cognition trade-off that must be resolved.

For example, aggressively fine-tuning an MLLM on a drone-detection dataset (e.g., Drone Dataset(UAV) [2]) to maximize visual acuity can overfit the model to specific visual patterns. As a result, it loses contextual cognition needed to answer higher-level questions such as "What threat does this drone's flight pattern pose to a nuclear facility?" Conversely, when training emphasizes situational reasoning, the model may over-leverage broad prior knowledge and fail to detect small, fast, or partially occluded non-canonical drones—manifesting knowledge bias. Ultimately, one of the two core capabilities—detection or understanding—tends to be sacrificed.

To address this trade-off head-on, we introduce a dual-LoRA fusion framework that leverages low-rank adaptation to efficiently integrate two complementary expert modules within a single model. From the same data source we define two distinct tasks—detection (Acuity) and assessment (Cognition)—and train a dedicated Acuity LoRA and Cognition LoRA, respectively. We then fuse the two expert parameter sets algebraically (α-fusion) to obtain a single unified model that retains both capabilities without additional inference-time cost. Our contributions are threefold:

- We formally articulate the Acuity–Cognition trade-off in MLLMs and propose a new method to resolve it.
- We present a procedure to derive two task views (detection vs. assessment) from a single dataset and train dedicated LoRA experts for each.
- We demonstrate a model-agnostic parameter-space fusion that efficiently integrates expert knowledge and yields a versatile, single model with no extra runtime overhead.

This approach maximizes the potential of MLLMs to build resource-efficient, high-performance intelligent security agents, offering a practical and innovative solution for protecting nuclear critical facilities.

2. Related works

2.1 Multimodal Large Language Model(MLLM)

The rise of MLLMs such as LLaVA, Flamingo, and IDEFICS has opened a new paradigm in which models interpret visual inputs within linguistic context and perform multi-step reasoning. Most architectures couple a strong, pretrained vision encoder to a Large Language Model(LLM) [3] through a lightweight projection or adapter, enabling end-to-end vision–language inference. While these systems target broad generality, their performance can plateau in specialized, safety-critical settings—e.g., nuclear facility surveillance—where the system must both detect a specific object with high confidence and understand task-specific behaviors and intent. Our work focuses on specializing a general MLLM to such a domain without sacrificing either capability, directly addressing the Acuity–Cognition trade-off.

2.2 Parameter-Efficient Fine-tuning(PEFT)

Fully fine-tuning an LLM/MLLM is computationally expensive. PEFT [4] methods mitigate this by freezing the backbone and learning a small set of additional parameters. Among them, LoRA (Low-Rank Adaptation) [5] has been widely adopted for its efficiency and strong empirical performance. LoRA approximates weight updates as the product of two low-rank matrices attached in parallel to the frozen weights, dramatically reducing trainable parameters. Crucially, LoRA modules can be stored, swapped, and composed as portable "skills." Our framework exploits this modularity by implementing two independent LoRA experts—one for detection (Acuity LoRA) and one for assessment (Cognition LoRA)—trained on distinct task views derived from the same data source.

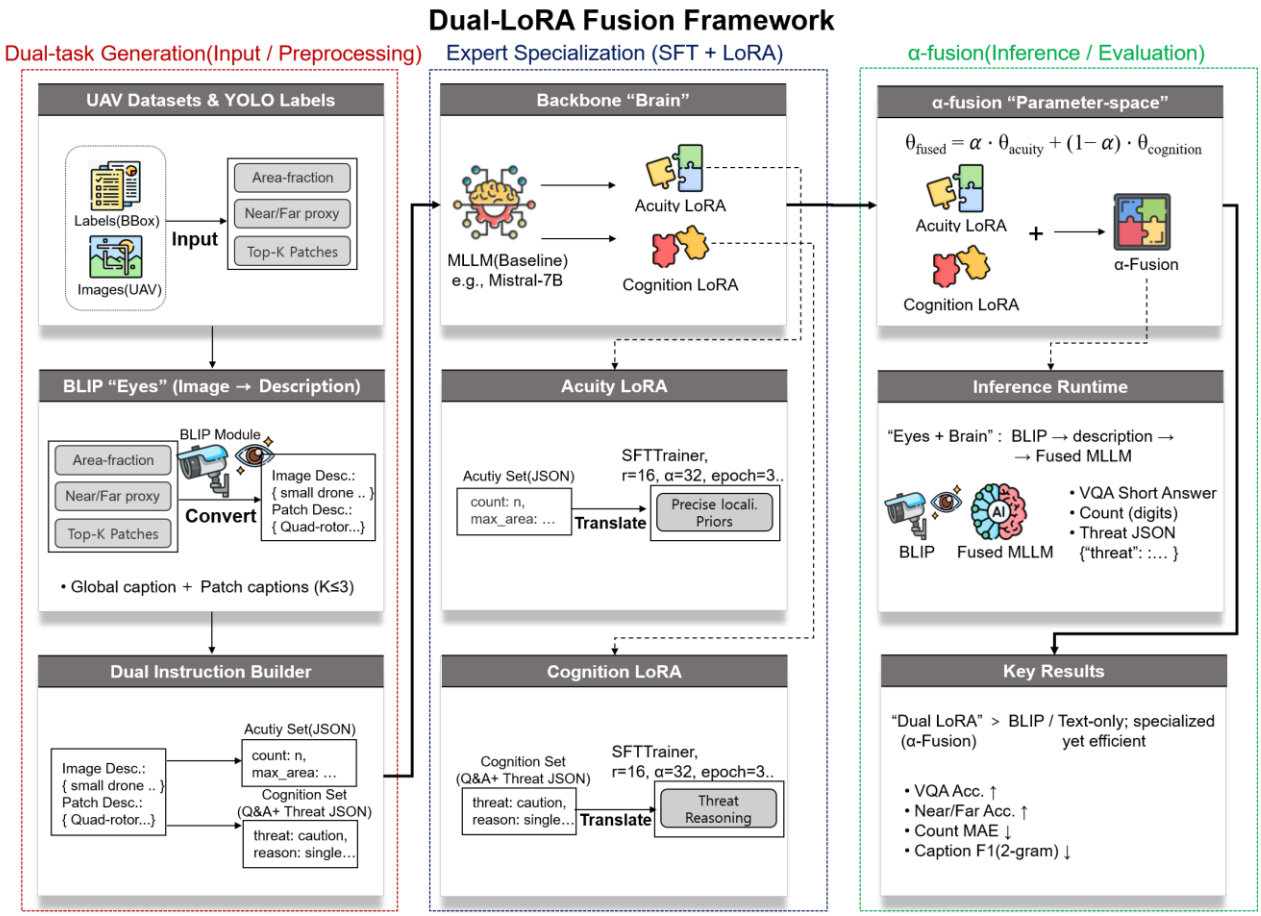

Figure 1. Overview of Dual-LoRA Fusion Framework

2.3 Expert Models and LoRA Fusion

Mixture-of-Experts (MoE) [6] approaches leverage multiple specialized experts to tackle complex problems, often achieving strong accuracy but incurring substantial inference overhead (expert routing, parallel branches). Recent work therefore explores merging multiple trained LoRA modules into a single backbone to retain advantages of specialization without runtime costs of MoE. Prior studies, however, typically merge task-aligned LoRAs trained on similar objectives across domains.

Our contribution advances this line by fusing LoRAs trained on intrinsically different, yet complementary tasks—detection vs. assessment—that are both derived from the same dataset. Using parameter-space fusion (α -fusion), we combine the Acuity and Cognition experts without additional inference cost, yielding a single versatile agent that balances conflicting capabilities. This design preserves the practical benefits of PEFT while sidestepping the computational burden associated with traditional MoE architectures.

3. Methodology

3.1 Proposed Framework: Dual-LoRA Fusion

The core of our approach is a three-stage framework that efficiently injects and fuses two complementary capabilities—detection (Acuity) and understanding (Cognition)—into an MLLM in Fig 1. The design is model-agnostic and applicable to a variety of backbones.

Step 1. Dual task–oriented instruction data generation

We derive two purpose-specific instruction–response corpora to explicitly separate learning objectives Drone Dataset(UAV).

· Acuity dataset: train pure visual detection acuity.
  - Instruction: "Locate the 'drone' objects in the image and return their bounding boxes."
  - Response (JSON format): [{"bbox": [x1, y1, x2, y2]}, ...]

· Cognition dataset: trains situational understanding and threat assessment. Labels are automatically produced using dataset metadata and predefined spatial context (e.g., core/restricted zones around a nuclear facility).
  - Instruction: "Analyze the drone's behavior in the image and assess the threat level."
  - Response (JSON format): {"assessment": "near", "reason": "The drone is loitering at low speed over a restricted zone."}

Step 2. Specialization with complementary LoRA experts

Using the two datasets, we attach two independent LoRA modules to the chosen backbone MLLM (Mistral-7B in our experiments) while keeping the backbone frozen—only the LoRA parameters are updated.

- Acuity LoRA: trained on the Acuity dataset) encodes precise object-localization cues.
- Cognition LoRA: trained on the Cognition dataset) encodes high-level situational reasoning and threat classification.

Unless otherwise noted, both experts share the same hyperparameters: learning rate 2e-4**,** batch size 8**,** 2 epochs, LoRA rank 8**,** LoRA alpha 16**.**

Step 3. Knowledge fusion in parameter space

To obtain a single unified model with both capabilities, we apply parameter-space weighted averaging (α-fusion) to the trained experts:

$$\theta_{fused} = \alpha \cdot \theta_{acuity} + (1-\alpha) \cdot \theta_{cognition}$$

where α balances detection acuity and cognitive assessment. In our experiments, α=0.6 minimized detection loss while maximizing assessment accuracy. Fusion is computationally inexpensive and introduces no additional inference overhead**,** yielding a final model that harmonizes otherwise competing abilities.

4. Results and Limitations

4.1 Experimental setup and Dataset

All experiments were conducted on a single NVIDIA RTX 4090 (24 GB) GPU. We used a Mistral-7B–based MLLM as the baseline backbone. Training and evaluation were performed on the Drone Dataset (UAV, 2025).

4.2 Evaluation metrics

- Visual Acuity: mean Average Precision at IoU 0.5 (mAP@0.5).
- Contextual Cognition: classification accuracy for threat assessment.

4.3 Quantitative Results

Our evaluation benchmarked the proposed dual-LoRA fusion against a text-only LLM (Mistral-7B) and a general-purpose vision–language model (BLIP); results are summarized in Table 1. Under our protocol, the text-only baseline cannot produce image-grounded outputs on these tasks (N/A), and BLIP reached 0.258 VQA accuracy while not directly supporting the Near/Far classification (N/A). In contrast, our α-fused model delivered 0.952 VQA accuracy, 0.925 Near/Far accuracy, and a Count MAE of 0.080, indicating successful integration of both Acuity and Cognition experts. As expected for a framework specialized for detection and assessment rather than open-ended description, Caption F1@2-gram on UAV frames remained low (0.015). These results empirically support our hypothesis about the Acuity–Cognition trade-off and highlight the effectiveness of parameter-space fusion for building a focused security agent.

Table 1. Quantitative results on the Drone Dataset

Metric	Baseline	VLM(BLIP)	Dual-LoRA(Ours)
VQA	N/A	0.258	0.952
Near/Far Acc.	N/A	0.000	0.925
Count MAE	N/A	0.737	0.08
Caption F1@ 2-gram	N/A	N/A	0.015

4.4 Qualitative Results

On an image of a small drone rapidly approaching over a rooftop, the models behaved distinctly:

- Acuity LoRA: precisely localized the drone but provided no meaningful assessment.
- Cognition LoRA: flagged the situation as risky but failed to detect the drone itself.
- Proposed fusion: delivered both accurate localization and a plausible threat assessment, qualitatively showing that it combines the strengths of both experts.

4.5 Limitations

While promising, this study has several limitations.

Firstly, the current threat-level assessment relies on programmed heuristic rules, which may be insufficient for complex and subtle threat scenarios.

Secondly, the proposed framework was validated on a single dataset (Drone Dataset(UAV)).

Thirdly, instead of using a static α for fusion, adopting input-adaptive control of α could further improve performance by dynamically balancing acuity and cognition.

5. Conclusions and Future Work

We presented a dual-LoRA fusion framework that addresses the intrinsic MLLM trade-off between visual acuity and contextual cognition by training two complementary LoRA experts separately and then fusing them. Experiments show that the proposed approach enables an MLLM to achieve high detection performance and strong reasoning ability simultaneously, outperforming conventional multi-task

training. Our results open a new avenue for efficiently integrating heterogeneous knowledge in parameter space, laying theoretical and practical groundwork for next-generation intelligent agents in high-reliability domains such as nuclear facility surveillance.

Future work will extend along several directions. Firstly, we will develop more refined and dynamic fusion mechanisms (e.g., a routing strategy that adjusts α per input). Secondly, we will expand the framework to additional expert skills, such as attack-type classification and intent inference of the drone operator. Finally, we plan to evaluate generalization by applying the framework beyond nuclear security to domains that also require balancing conflicting capabilities—such as autonomous driving and medical image analysis.

Acknowledgements

This work was supported by the Korea AeroSpace Administration(KASA) grant funded by the Korea government.

REFERENCES

[1] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, “A survey on multimodal large language models,” in National Science Review, 2024.
[2] M. Pawełczyk and M. Wojtyra, "Real world object detection dataset for quadcopter unmanned aerial vehicle detection," *IEEE Access*, vol. 8, pp. 174394-174409, 2020.
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All You Need,” in Advances in Neural Information Processing Systems (NIPS), 2017.
[4] L. Wang, S. Chen, L. Jiang, S. Pan, R. Cai, S. Yang, and F. Yang, “Parameter-efficient fine-tuning in large language models: a survey of methodologies,” in Artifi cial Intelligence Review, 2025.
[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” in International Conference on Learning Representations (ICLR), 2022.
[6] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A Survey on Mixture of Experts in Large Language Models," IEEE Transactions on Knowledge and Data Engineering, 2025.