

Distributed Benchmarking of Vision-Language Models under Semantic Adversarial Attacks in UAV Emergency Scenarios

Amina Ben Taher*, Samiha Ayed[†], Lamia Chaari Fourati[‡], Sofiane Aissani[§]

* LIST3N, University of Technology of Troyes, France

National School of Electronics and Telecommunications of Sfax, Tunisia

Digital Research Center of Sfax (CRNS), SM@RTS Laboratory, Sfax University, Tunisia

ORCID: 0009-0004-7401-4410

aminabentaher15@gmail.com

[†] Lab-STICC, IMT Atlantique, Brest, France

ORCID: 0000-0002-4678-6956

samiha.ayed@imt-atlantique.fr

[‡] Digital Research Center of Sfax (CRNS), SM@RTS Laboratory, Sfax University, Tunisia

ORCID: 0000-0003-0401-5050

lamia.chaari@isims.usf.tn

[§] LIST3N, University of Technology of Troyes, France

ORCID: 0000-0003-4926-0455

sofiane.aissani@utt.fr

Abstract—Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) are increasingly used in UAV-based perception systems operating in safety-critical environments. However, semantic adversarial manipulations can insert realistic emergency cues, such as fire or smoke, raising concerns about their robustness.

This work introduces a distributed benchmark for evaluating multimodal perception systems under semantic adversarial attacks. The framework considers three-node deployments where adversarial images are independently assigned to nodes with probability ρ , while textual context propagates sequentially across nodes. Homogeneous and heterogeneous configurations based on BLIP-2 and InstructBLIP are evaluated using Majority, Confidence-Weighted, and Reliability-Weighted Voting. Robustness is assessed through clean-conditioned attack success rate (ASR), consensus failure, and node-level accuracy.

Experiments over five independent adversarial-assignment runs show that the heterogeneous deployment maintains Majority-Voting ASR below 2.5% across all evaluated conditions, compared with 6.93% for homogeneous BLIP-2 and 23.50% for homogeneous InstructBLIP at $\rho = 0.8$. Confidence-Weighted Voting is consistently more vulnerable, while consensus failure decreases with increasing adversarial exposure, showing that higher agreement does not necessarily imply greater robustness.

The proposed benchmark provides a reproducible protocol for stress-testing distributed multimodal perception systems and highlights the importance of architectural diversity and aggregation strategy for robust UAV-based emergency monitoring.

Index Terms—Vision-Language Models, Multimodal Large Language Models, Semantic Adversarial Attacks, Multimodal AI Security, Distributed Inference, UAV Security

I. INTRODUCTION

Vision-Language Models (VLMs) [1] and Multimodal Large Language Models (MLLMs) [2] are increasingly deployed

in UAV-based emergency monitoring systems, where they support scene understanding and decision-making in safety-critical environments [3]. In distributed deployments, multiple AI nodes collaboratively process visual information and combine their predictions to improve system reliability.

Recent advances in generative AI have enabled semantic adversarial attacks capable of inserting realistic fire, smoke, or other emergency cues while preserving visual plausibility [4]. Such manipulations may trigger false alarms and compromise operational decisions in emergency-response scenarios.

Although multimodal robustness has been extensively studied, existing evaluations primarily focus on individual models. The behavior of distributed multimodal systems under partial-node compromise remains largely unexplored. Understanding how heterogeneous architectures and aggregation strategies affect system-level robustness is therefore important for the secure deployment of VLMs and MLLMs.

To address this challenge, we introduce a distributed benchmark for evaluating semantic adversarial attacks against collaborative VLM and MLLM inference systems in UAV emergency monitoring scenarios. The proposed framework considers adversarial node-assignment probabilities ranging from 20% to 80%, evaluates homogeneous and heterogeneous deployments based on BLIP-2 and InstructBLIP architectures, and compares Majority Voting, Confidence-Weighted Voting, and Reliability-Weighted Voting. System robustness is assessed using clean-conditioned attack success rate (ASR), consensus failure, and node-level accuracy.

By shifting the focus from individual model robustness to collective system behavior, the benchmark provides a repro-

ducible evaluation protocol and experimental framework for stress-testing distributed multimodal perception systems before deployment in safety-critical environments.

II. BACKGROUND AND RELATED WORK

VLMs and MLLMs are increasingly used in multimodal perception systems, including safety-critical applications such as autonomous systems, emergency monitoring, and UAV-based disaster response [5]. Prior studies have demonstrated their vulnerability to visual and textual adversarial perturbations [9], including attacks that transfer across architectures and black-box settings [10], [11]. Robustness has also been investigated under feature-level perturbations [13], distribution shifts [14], and multimodal attacks against safety-aligned models [15]–[17].

Several benchmarks have been proposed to systematically evaluate adversarial robustness [6]–[8]. However, these approaches primarily evaluate models independently rather than their behavior within collaborative multimodal inference systems.

Robustness to compromised participants has also been studied in distributed and Byzantine-resilient systems [20]–[22]. Unlike conventional Byzantine settings, our perception nodes remain benign, while their visual observations may be semantically manipulated before inference. This distinction motivates evaluating robustness at the distributed perception level.

Existing adversarial evaluations therefore remain predominantly model-centric, with limited attention to partial semantic compromise in distributed multimodal inference. To address this gap, we benchmark homogeneous and heterogeneous VLM deployments under partial adversarial exposure and evaluate multiple aggregation strategies using clean-conditioned attack success rate, consensus failure, and propagation effect. This system-level evaluation enables the analysis of architectural diversity and aggregation behavior in UAV emergency monitoring.

III. DISTRIBUTED BENCHMARK FRAMEWORK

We propose a distributed benchmark to evaluate semantic adversarial manipulations in collaborative VLM/MLLM inference. The framework consists of N perception nodes whose outputs are aggregated into a global system decision. Given a clean UAV image x , an adversary generates a semantically manipulated counterpart $\tilde{x}$ containing realistic fire or smoke cues. For each scene, $\tilde{x}$ is independently assigned to each node with probability ρ , while non-selected nodes receive x .

A. System Overview

As illustrated in Fig. 1, clean and adversarial UAV images are distributed across perception nodes under partial-compromise scenarios. The benchmark supports homogeneous and heterogeneous configurations based on BLIP-2 and InstructBLIP architectures. Inference follows a sequential $A \rightarrow B \rightarrow C$ pipeline: image assignment is independent across nodes, while textual context is propagated between consecutive

nodes. Node predictions are then combined using system-level aggregation strategies and evaluated through attack success rate, consensus failure, and propagation effect.

The evaluation addresses two research questions:

RQ1: How does semantic adversarial manipulation affect distributed multimodal systems under partial-node compromise?

RQ2: How does architectural diversity influence distributed robustness and consensus?

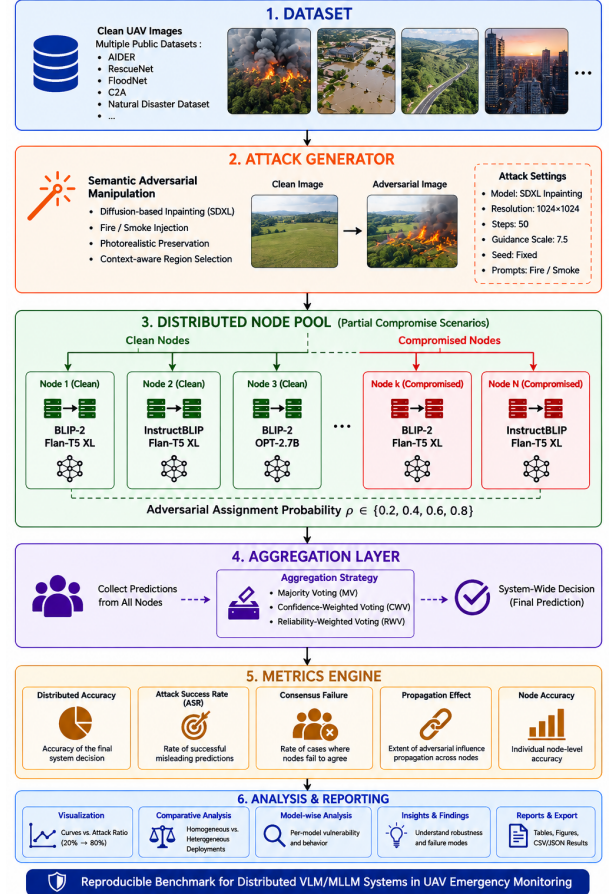


Fig. 1. Overview of the proposed distributed benchmark framework.

B. Aggregation Strategies

We evaluate three aggregation strategies. Majority Voting (MV) selects the class predicted by most nodes. Confidence-Weighted Voting (CWV) weights predictions by generation confidence and selects the class with the largest cumulative score. Reliability-Weighted Voting (RWV) uses fixed node weights estimated from predictive accuracy on a separate clean calibration set. These strategies enable comparison between conventional, confidence-aware, and reliability-aware decision fusion.

IV. EXPERIMENTAL EVALUATION

A. Threat Model

We consider a gray-box adversary capable of injecting semantically manipulated UAV images containing realistic fire

and smoke cues into a subset of distributed perception nodes. The attacker has no access to model parameters, gradients, or internal representations, but possesses knowledge of the system architecture and deployment configuration.

The attacker transforms a clean UAV image into an adversarial sample by inserting visually plausible emergency indicators while preserving the overall scene realism. Unlike conventional pixel-level perturbations, the attack operates at the semantic level by modifying the scene content itself. As a result, manipulated images remain photorealistic while conveying misleading emergency information.

For each scene, each distributed node receives the adversarial image independently with probability ρ , whereas non-selected nodes process the original observation. Thus, ρ controls the expected adversarial exposure across node-scene assignments rather than specifying an exact compromised node fraction for each scene.

The objective of the attack is to trigger false emergency detections and disrupt collaborative decision-making. By exposing different nodes to inconsistent observations of the same scene, the attack may induce prediction drift, consensus instability, and propagation of incorrect decisions throughout the distributed system.

The threat model considered in this work is illustrated in Fig. 2.

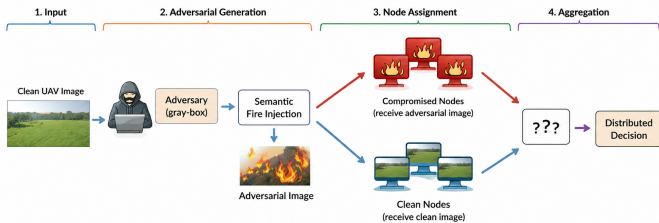


Fig. 2. Partial-node semantic adversarial attack model in distributed UAV perception systems.

B. Dataset Construction

To evaluate distributed multimodal systems under semantic adversarial attacks, we construct a benchmark composed of clean UAV scenes collected from multiple public emergency-monitoring datasets. Only images without visible crisis indicators (e.g., fire, smoke, flooding, or structural damage) are retained to ensure that emergency events are introduced exclusively through adversarial manipulation.

The benchmark aggregates images from five complementary datasets covering diverse environments, including urban areas, road networks, vegetation regions, and critical infrastructures. All images are converted to RGB format and resized to a common resolution before being merged into a unified corpus.

Table I summarizes the datasets used for benchmark construction.

Representative examples of clean UAV scenes are shown in Fig. 3. Each selected image is paired with an adversarially manipulated counterpart containing synthetic fire and smoke, enabling controlled evaluation under semantic attacks.

TABLE I
CLEAN UAV IMAGES RETAINED FOR BENCHMARK CONSTRUCTION.

Dataset	Images
AIDER	5922
RescueNet	4494
FloodNet	2343
C2A	10215
Natural Disaster Dataset	2976
Total	25,950

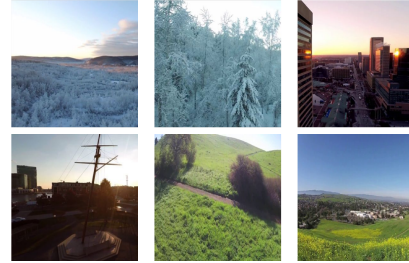


Fig. 3. Examples of clean UAV scenes used in the benchmark.

C. Experimental Setup

The semantic attack generation follows our previous work on UAV crisis monitoring [19]. Adversarial samples were generated using Stable Diffusion Inpainting at 768×768 resolution, with a base seed of 2026, 80–100 inference steps, and a guidance scale of 16.0–20.0. Context-aware prompts were adapted to four scene categories (forest, urban, road, and infrastructure) and three severity levels (low, medium, and high). The manipulation consists of inserting realistic fire or smoke cues while preserving the surrounding scene. Once generated, the adversarial images were paired with their clean counterparts and reused throughout all experiments. The main generation parameters are summarized in Table II.

TABLE II
SEMANTIC ATTACK GENERATION CONFIGURATION.

Parameter	Configuration
Generator	Stable Diffusion Inpainting
Resolution	768×768
Base seed	2026
Inference steps	80–100
Guidance scale	16.0–20.0
Scene categories	Forest, urban, road, infrastructure
Severity levels	Low, medium, high
Manipulation	Fire/smoke insertion

The distributed system consists of three heterogeneous multimodal nodes: BLIP-2 Flan-T5 XL, InstructBLIP Flan-T5 XL, and BLIP-2 OPT-2.7B (Table III). Node A uses the following binary classification prompt:

“Is there visible fire, flames, or smoke in this UAV image? Answer with one word only: fire or normal.”

For Nodes B and C, the preceding node’s textual response is appended as context to the same binary question, forming a sequential $A \rightarrow B \rightarrow C$ inference chain.

We evaluate 2,000 scenes at four adversarial assignment probabilities, $\rho \in \{0.2, 0.4, 0.6, 0.8\}$. For each scene, each

TABLE III
EVALUATED MULTIMODAL ARCHITECTURES.

Model	Family	Language Backbone
BLIP-2	VLM	Flan-T5 XL
InstructBLIP	MLLM	Flan-T5 XL
BLIP-2	VLM	OPT-2.7B

of the three perception nodes independently receives the adversarial image with probability ρ ; otherwise, it receives the corresponding clean image. Therefore, ρ represents the probability of adversarial assignment over node-scene pairs rather than the exact fraction of compromised nodes within an individual three-node scene. The realized compromised-node fraction for a given scene can thus take values in $\{0, 1/3, 2/3, 1\}$.

To account for the stochasticity of adversarial assignment, each experiment is repeated using five independent assignment seeds (2026–2030). For each seed, a new compromise plan is generated and applied consistently across all deployment configurations to ensure a fair comparison. Reported results are averaged across the five runs, and error bars indicate ± 1 standard deviation across runs.

Figure 4 shows the overall evaluation process. Clean or adversarial inputs are assigned to the three nodes according to ρ , and their predictions are then combined by the system-level aggregation mechanism.

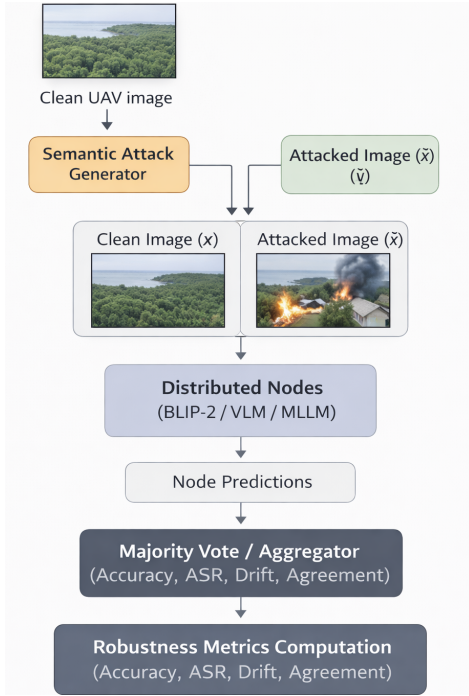


Fig. 4. Distributed adversarial evaluation pipeline.

D. Evaluation Metrics

Distributed Accuracy measures the proportion of scenes for which the final majority-vote decision matches the ground-truth label.

TABLE IV
SYSTEM-LEVEL ROBUSTNESS OF THE HETEROGENEOUS DEPLOYMENT

ρ	Attack Success Rate (ASR, %)			Distributed Behavior (%)	
	Majority	Confidence-Weighted	Reliability-Weighted	Consensus Failure	Propagation Effect
0.2	1.35 $\pm$ 0.25	2.57 $\pm$ 0.25	1.35 $\pm$ 0.25	23.28 $\pm$ 0.63	80.81 $\pm$ 0.46
0.4	2.04 $\pm$ 0.12	4.05 $\pm$ 0.21	2.04 $\pm$ 0.12	22.20 $\pm$ 0.41	82.47 $\pm$ 0.46
0.6	2.46 $\pm$ 0.16	4.50 $\pm$ 0.19	2.46 $\pm$ 0.16	20.46 $\pm$ 0.85	84.68 $\pm$ 0.83
0.8	2.17 $\pm$ 0.31	3.74 $\pm$ 0.22	2.17 $\pm$ 0.31	18.72 $\pm$ 0.86	86.15 $\pm$ 0.59

Attack Success Rate (ASR) measures the fraction of samples that are correctly classified in the clean setting but become misclassified after adversarial manipulation. For an aggregation strategy g ,

$$\text{ASR}_g = \frac{|\{x : g(x) = y \wedge g(\tilde{x}) \neq y\}|}{|\{x : g(x) = y\}|}. \quad (1)$$

Here, x denotes the clean input, $\tilde{x}$ its partially compromised counterpart, and y the ground-truth label. A clean baseline was evaluated on the same 2,000 scenes used for each adversarial assignment probability.

Consensus Failure (CF) measures the proportion of scenes for which the distributed nodes fail to reach unanimous agreement. For M scenes and N nodes,

$$\text{CF} = \frac{1}{M} \sum_{j=1}^M \mathbb{I}(|\{y_{1j}, \dots, y_{Nj}\}| > 1), \quad (2)$$

where y_{ij} is the prediction of node i for scene j .

Propagation Effect (PE) measures the extent to which downstream nodes reproduce the prediction of the first node in the sequential distributed inference pipeline. For each scene,

$$\text{PE} = \frac{1}{N-1} \sum_{i=2}^N \mathbb{I}(y_i = y_1), \quad (3)$$

where y_1 is the prediction of Node A and y_i denotes the prediction of each downstream node ($i \in \{2, \dots, N\}$). A value of $\text{PE} = 1$ indicates that all downstream nodes agree with Node A, while $\text{PE} = 0$ indicates complete disagreement.

E. Experimental Results

We first evaluate the heterogeneous distributed deployment composed of three multimodal perception nodes: BLIP-2 Flan-T5 XL, InstructBLIP Flan-T5 XL, and BLIP-2 OPT-2.7B, under adversarial assignment probabilities $\rho \in \{0.2, 0.4, 0.6, 0.8\}$.

Results are averaged over five independent adversarial-assignment runs using seeds 2026–2030. Table IV summarizes the clean-conditioned ASR and distributed behavior.

Majority Voting and Reliability-Weighted Voting maintain low ASR, ranging from 1.35% to 2.46%. Confidence-Weighted Voting is consistently more vulnerable, with ASR ranging from 2.57% to 4.50% and reaching its maximum at $\rho = 0.6$. Reliability-Weighted Voting produces the same system-level decisions as Majority Voting across all five runs and therefore yields identical ASR values.

These results indicate that the aggregation strategy directly affects system robustness. In particular, confidence weighting

gives greater influence to highly confident predictions and can therefore amplify incorrect adversarial responses. The attack effectiveness is also non-monotonic with respect to ρ , since the highest mean ASR for Majority Voting occurs at $\rho = 0.6$ rather than $\rho = 0.8$.

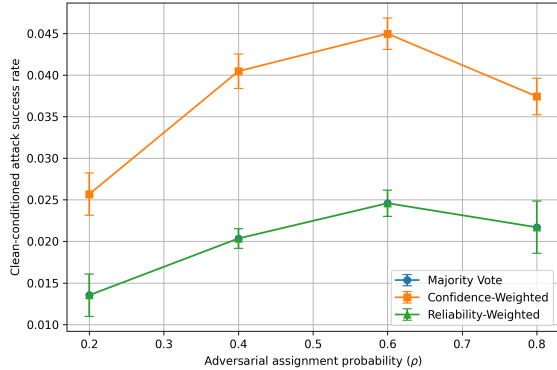


Fig. 5. Clean-conditioned attack success rate across aggregation strategies.

Figure 5 reports the corresponding clean-conditioned ASR. Confidence-Weighted Voting consistently exhibits higher ASR than Majority and Reliability-Weighted Voting, reaching 4.50% at $\rho = 0.6$. Majority and Reliability-Weighted Voting overlap because they produce identical system-level decisions in all evaluated runs.

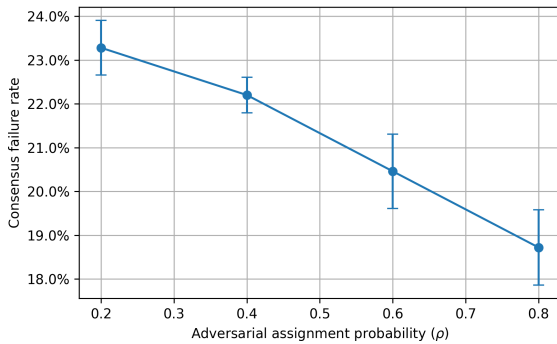


Fig. 6. Consensus failure rate under increasing adversarial assignment probability.

As shown in Fig. 6, consensus failure decreases from $23.28 \pm 0.63\%$ at $\rho = 0.2$ to $18.72 \pm 0.86\%$ at $\rho = 0.8$. In parallel, the propagation effect increases from $80.81 \pm 0.46\%$ to $86.15 \pm 0.59\%$.

This behavior shows that increasing adversarial exposure does not necessarily produce greater disagreement. Instead, the distributed nodes can increasingly converge toward similar decisions under adversarial influence. Consequently, a low disagreement rate should not automatically be interpreted as greater robustness, since it may also reflect adversarially induced false consensus.

To assess the role of architectural diversity, we compare the heterogeneous deployment with homogeneous BLIP-2 and homogeneous InstructBLIP configurations. All configurations use the same 2,000 scenes, the same pre-generated

TABLE V
HOMOGENEOUS VS. HETEROGENEOUS ROBUSTNESS OVER FIVE RUNS

Configuration	ρ	ASR (%)	Consensus Failure (%)	Propagation Effect (%)
Homogeneous BLIP-2	0.2	1.52 ± 0.17	1.28 ± 0.21	98.87 ± 0.17
	0.4	3.65 ± 0.21	0.81 ± 0.17	99.27 ± 0.16
	0.6	5.05 ± 0.25	0.60 ± 0.14	99.45 ± 0.14
	0.8	6.93 ± 0.20	0.38 ± 0.18	99.63 ± 0.16
Homogeneous InstructBLIP	0.2	1.62 ± 0.41	14.31 ± 0.47	91.56 ± 0.41
	0.4	6.23 ± 0.59	24.66 ± 0.62	84.08 ± 0.54
	0.6	13.54 ± 0.49	31.68 ± 0.50	76.95 ± 0.49
	0.8	23.50 ± 0.66	35.15 ± 0.36	70.53 ± 0.47
Heterogeneous	0.2	1.35 ± 0.25	23.28 ± 0.63	80.81 ± 0.46
	0.4	2.04 ± 0.12	22.20 ± 0.41	82.47 ± 0.46
	0.6	2.46 ± 0.16	20.46 ± 0.85	84.68 ± 0.83
	0.8	2.17 ± 0.31	18.72 ± 0.86	86.15 ± 0.59

adversarial samples, and the same five adversarial-assignment seeds, allowing a direct paired comparison across deployment configurations.

Table V shows that the heterogeneous deployment provides the most stable system-level robustness. Its Majority-Voting ASR remains between 1.35% and 2.46% across all adversarial assignment probabilities.

In contrast, homogeneous BLIP-2 becomes progressively more vulnerable, with ASR increasing from $1.52 \pm 0.17\%$ at $\rho = 0.2$ to $6.93 \pm 0.20\%$ at $\rho = 0.8$. Homogeneous InstructBLIP exhibits the strongest degradation, increasing from $1.62 \pm 0.41\%$ to $23.50 \pm 0.66\%$ over the same range.

At $\rho = 0.8$, the heterogeneous deployment therefore achieves substantially lower ASR than both homogeneous BLIP-2 and homogeneous InstructBLIP. These results indicate that architectural diversity can reduce correlated adversarial failures and limit the increase in system-level ASR observed in homogeneous deployments.

The distributed-behavior metrics further reveal that consensus alone is not a sufficient indicator of robustness. Homogeneous BLIP-2 shows very low consensus failure, decreasing to 0.38% at $\rho = 0.8$, while its propagation effect reaches 99.63%. This indicates that high agreement may correspond to correlated responses rather than robust consensus.

Conversely, the heterogeneous deployment maintains higher cross-node disagreement but substantially lower system-level ASR. Architectural diversity therefore appears to provide complementary decision behavior that limits correlated adversarial failures, even when individual nodes disagree more frequently.

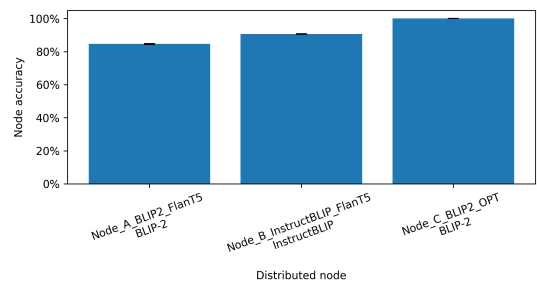


Fig. 7. Per-node accuracy across multimodal architectures

Figure 7 compares the node-level accuracy of the three multimodal architectures under adversarial exposure. Bars represent the mean accuracy over five independent adversarial-

assignment runs, while error bars indicate ± 1 standard deviation across runs. BLIP-2 Flan-T5 (Node A) achieves approximately 85% accuracy, InstructBLIP Flan-T5 (Node B) reaches about 91%, and BLIP-2 OPT (Node C) achieves nearly 100%. The small standard deviations indicate consistent performance across runs and highlight architecture-dependent differences in robustness.

Overall, these results answer RQ2 by demonstrating that architectural diversity strongly influences distributed robustness. At low adversarial exposure, the three deployments exhibit comparable ASR, whereas their behavior diverges strongly as ρ increases. The heterogeneous configuration maintains ASR below 2.5% across all tested conditions, while homogeneous BLIP-2 reaches 6.93% and homogeneous InstructBLIP reaches 23.50% at $\rho = 0.8$.

The results also show that aggregation strategy matters within the heterogeneous deployment: Confidence-Weighted Voting is consistently more vulnerable than Majority and Reliability-Weighted Voting. Together, these findings indicate that both model diversity and aggregation policy influence the resilience of distributed multimodal perception systems.

V. CONCLUSION

This paper introduced a distributed benchmark for evaluating the robustness of VLMs and MLLMs against semantic adversarial attacks in UAV emergency monitoring. The framework evaluates both homogeneous and heterogeneous deployments under partial adversarial exposure and compares multiple aggregation strategies.

Experiments over five independent adversarial-assignment runs show that robustness depends on both deployment architecture and aggregation strategy. The heterogeneous configuration maintains Majority-Voting ASR below 2.5% across all evaluated adversarial assignment probabilities, whereas homogeneous BLIP-2 and homogeneous InstructBLIP reach 6.93% and 23.50% ASR, respectively, at $\rho = 0.8$. Confidence-Weighted Voting is consistently more vulnerable than Majority and Reliability-Weighted Voting. Moreover, the results show that higher cross-node agreement does not necessarily imply greater robustness, highlighting the benefit of architectural diversity in limiting correlated adversarial failures.

The aggregation stage operates only on compact prediction-level outputs, while the main computational cost arises from multimodal inference at the distributed nodes. This design supports parallel node execution and avoids exchanging images or intermediate model representations. Overall, the proposed benchmark provides a reproducible protocol for stress-testing distributed multimodal perception systems before deployment in safety-critical environments. Future work will investigate computational scalability, including latency, memory, energy, and communication overhead, as well as robust aggregation, compromised-node detection, and defense mechanisms for larger distributed VLM/MLLM systems.

REFERENCES

- [1] A. Sharshar, L. U. Khan, W. Ullah and M. Guizani, "Vision-Language Models for Edge Networks: A Comprehensive Survey," in IEEE Internet of Things Journal, vol. 12, no. 16, pp. 32701-32724, 15 Aug. 15, 2025, doi: 10.1109/IIOT.2025.3579032.
- [2] Jin, Y., Li, J., Gu, T. et al. Efficient multimodal large language models: a survey. *Vis. Intell.* 3, 27 (2025). <https://doi.org/10.1007/s44267-025-00099-6>
- [3] S. Kotsiantis, T. Panagiotakopoulos, N. Piperigkos and A. S. Lalos, "Visual Language Models: Foundations and Applications," in IEEE Access, vol. 14, pp. 32431-32454, 2026, doi: 10.1109/ACCESS.2026.3664520.
- [4] REN, Pengfei, WANG, Jingjing, CHEN, Jianrui, et al. Generative AI meets semantic communications: Opportunities, challenges and security risks. *IEEE Network*, 2025.
- [5] Y. Jing et al., "Multi-modal Large Language Models for Perception in Autonomous Driving: Architecture, Taxonomy, and Challenges," in IEEE Internet of Things Journal, doi: 10.1109/IIOT.2026.3668085.
- [6] J. Donato, "Benchmarking LLM Robustness Against Prompt-Based Adversarial Attacks," in *2025 20th European Dependable Computing Conference Companion Proceedings (EDCC-C)*, Lisbon, Portugal, 2025, pp. 60-63, doi: 10.1109/EDCC-C66476.2025.00031.
- [7] O. Kushnerov, R. Shevchuk, S. Yevseiev et al., "Comparative Benchmarking of Deep Learning Architectures for Detecting Adversarial Attacks on Large Language Models," *Information*, vol. 17, no. 2, p. 155, 2026.
- [8] R. Cantini, A. Orsino, M. Ruggiero et al., "Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge," *Machine Learning*, vol. 114, no. 11, p. 249, 2025.
- [9] Zhang, J., Yi, Q., Sang, J.: Towards adversarial attack on vision-language pre-training models. In: *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 1168-1176 (2022)
- [10] Wang, Y., Hu, W., Dong, Y., Zhang, H., Su, H., Hong, R.: Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning. *IEEE Transactions on Multimedia* 27, 6410-6422 (2025)
- [11] Jia, X., Gao, S., Guo, Q., Qin, S., Ma, K., Huang, Y., Liu, Y., Tsang, I.W., Cao, X.: Semantic-aligned adversarial evolution triangle for high-transferability vision-language attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47(10), 8489-8506 (2025)
- [12] Fox, J., Buchanan, W.J., Papadopoulos, P.: Adversarial robustness of vision in open foundation models. *IEEE Access* 13, 216272-216290 (2025)
- [13] Liu, C., Wang, Y., Cao, H., Liu, B., Jiang, D.: Evaluating the adversarial robustness of vision-language models via internal feature perturbations. *IEEE Transactions on Circuits and Systems for Video Technology* (2025)
- [14] Lu, X., Zhang, Y., Zhao, Y., Liu, J.: Rethinking adversarial attacks in reinforcement learning from policy distribution perspective. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2024)
- [15] Ying, Z., Liu, A., Zhang, T., Yu, Z., Liang, S., Liu, X., Tao, D.: Jailbreak vision language models via bi-modal adversarial prompt. *IEEE Transactions on Information Forensics and Security* 20, 7153-7167 (2025)
- [16] Wang, Y., Hu, W., Dong, Y., Liu, J., Zhang, H., Hong, R.: Align is not enough: Multimodal universal jailbreak attack against multimodal large language models. *IEEE Transactions on Circuits and Systems for Video Technology* 35(6), 5475-5489 (2025)
- [17] Shi, X., Chen, S., Zhang, G., Wei, W., Li, Y., Fan, Z., Liu, J.: Jailbreak attack with multimodal virtual scenario hypnosis for vision-language models. *Pattern Recognition* 172, 112391 (2026)
- [18] Zhang, H., Chen, J., Li, Y., Xu, C.: Multimodal evidential learning for open-world weakly-supervised video anomaly detection. *Pattern Recognition* 156, 110866 (2024)
- [19] Ben Taher, A., Ayed, S., Fourati, L.C.: Detecting semantic adversarial attacks on UAVs in crisis monitoring. In: *International Conference on Advanced Information Networking and Applications*, pp. 310-321. Springer Nature Switzerland, Cham (2026)
- [20] Cajaraville-Aboy, D., Fernández-Vilas, A., Díaz-Redondo, R.P., Fernández-Veiga, M.: Byzantine-robust aggregation for securing decentralized federated learning. *IEEE Access*, vol. 13, pp. 190947-190963 (2025).
- [21] Lee, D.Z., Kolluri, A., Saxena, P., Chang, E.-C.: A practical and secure Byzantine robust aggregator. In: *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2877-2891. ACM (2025).
- [22] Mastromauro, L., Andrade, D.S., Ozmen, M.O., Kinsy, M.A.: Survey of attacks and defenses on consensus algorithms for secure data replication in distributed systems. *IEEE Access*, vol. 13, pp. 143631-143667 (2025).