

AeroLat: Channel-Aware Latent Space Semantic Communication for Decentralized UAV Swarms

Rajdeep Ghosh[✉], *Graduate Student Member, IEEE*, Goparaju Venkata Seshachala Sree Vatsava[✉],
Sudip Misra[✉], *Fellow, IEEE, ACM*

Abstract—Communication in latent space offers an intriguing alternative to symbolic messages for decentralized autonomous Unmanned Aerial Vehicle (UAV) swarms operating over bandwidth-constrained, time-varying wireless links. However, when homogeneous frozen models are prompted with discretized perceptual inputs, their broadcast states collapse toward the shared prompt template. In view of this, we propose *AeroLat*, a channel-aware latent semantic communication framework that uses evidence injection. The resulting latent states are then passed through an explicit communication model that encompasses bandwidth-limited serialization, additive noise and information staleness, which facilitates a joint assessment of communication fidelity and swarm-level coordination. Across multi-seed simulations, AeroLat provably remains resilient to codec choice, faults and increasing swarm size. It consistently reproduces the latent-swarm anomaly, while no-whitening controls recover the collapse. In particular, AeroLat is capable of reducing false similarity by 97.5%.

Index Terms—Semantic communication, latent-space communication, multi-agent systems, LLMs, decentralized UAV swarms, representational collapse.

I. INTRODUCTION

UAV swarms for mission-oriented applications such as post-disaster surveillance and solid waste disposal violations, carry zero-shot perception and lightweight onboard Large Language Model (LLM) reasoners. The open coordination question is how such agents should share their beliefs, thoughts, or perceptions. The pre-dominant answer to it is based on natural language or structured symbolic messaging. However, this imposes what prior work calls a *symbolic bottleneck* — auto-regressive decoding latency at the sender, quantization of a rich internal belief distribution into one discrete report, and re-encoding ambiguity at the receiver. In coherence with this, recent advancements in this direction of research moved inter-agent communication into the models’ continuous latent space. For UAV swarms, the appeal of latent communication extends beyond avoiding the overhead of language generation. When a language model reads text, it builds a vector, a long list of numbers, that represents its internal state: what it currently *has in mind* before it picks a word. Writing a sentence throws most of that away, because the model holds a whole distribution over possible meanings but has to emit one specific string of words. Latent-space communication skips this intermediate symbolic

representation by directly exchanging internal model states between agents that share a compatible representation space.

A. Motivation

Let us consider a mission in which a small fleet surveys an unstructured river basin or urban fringe for waste accumulation and debris. No agent sees the whole scene; each holds a partial, *genuinely ambiguous* view. The fleet must converge on whether an anomaly exists while flying, over a shared wireless medium, on companion computers that also run perception. Three properties of this setting drive every subsequent design decision. First, the *symbolic bottleneck* presents a concrete system-level challenge, due to its measurable overhead in both communication and decision-making. If agents must serialize belief into language before transmitting, three losses compound. In measurements, it takes 0.88 s per message on a data-center GPU and 3.1 s on an edge-grade GPU, which is much higher than a single forward pass. On the other hand, the latent channel introduces a substantial communication overhead that remains largely unaccounted for.

B. Contributions

We propose AeroLat, a latent semantic-communication stack for decentralized UAV swarms. The proposed system treats the broadcast hidden state as a signal to be engineered rather than as a tensor to be passed. Each agent firstly articulates its full zero-shot visual posterior into a frozen onboard LLM, secondly takes the last-layer hidden states as its message, and lastly removes the shared prompt-template subspace with a linear operator that every agent derives independently and identically, training-free, coordination-free, and with one projection per message. The whitened state then crosses an explicitly modeled aerial link with erasures, jitter, bandwidth-limited serialization, noise and quantization, and the task information it carries is measured rather than asserted. Consequently, consensus tracks the scene rather than the prompt, both in emulation and across twenty-five cold-boot flights of the full stack. In brief, the contributions of the proposed work are as follows:

- We propose *evidence injection* and *template whitening*, a shared linear operator that requires neither training nor inter-agent coordination, but restores discriminative similarity to 0.025— a unique contribution to the state-of-the-art in decentralized UAV swarm communication.
- We benchmark the proposed model against existing dispatching methodologies for swarms, implemented on an

R. Ghosh, G. V. S. S. Vatsava, and S. Misra are with the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India (e-mail: rajghosh2507.24@kgpian.iitkgp.ac.in; goparajusreevatsava@gmail.com; sudipm@iitkgp.ac.in).

identical perception stack, across twenty seeded runs using Welch and Mann–Whitney tests with Holm correction.

- We validate the complete stack in a PX4/AirSim campaign on an edge-class GPU, spanning codec comparison, fault injection, impaired channels, and fleet scaling to show the efficacy of the proposed work.

II. RELATED WORK

Researchers have been developing new methodologies for a semantic communication stack for decentralized UAV swarms. Table I summarizes the existing landscape. In this section, we briefly present some of the existing literature categorized as: 1) Latent Inter-Agent Communication and 2) Semantic Communication and UAV Networking.

A. Latent Inter-Agent Communication

Interlat [1] established end-to-end communication via last hidden states, with learned compression into short latent prefixes. LatentMAS [2] demonstrated training-free latent collaboration via last-layer embeddings and shared KV working memory, reporting up to 14.6% accuracy gains and 4-4.3× faster inference over text-based multi-agent systems. CIPHER [3] bypasses token sampling with expectation embeddings. ThoughtComm [4] establishes identifiability guarantees through sparsity-regularized autoencoders, while KV-Comm [5] enables latent communication through the transmission of calibrated KV caches. Dense inter-model messaging without a shared latent geometry is explored in [13]. Cross-model latent alignment is enabled by relative representations [14] and semantic-alignment translation [15]; continuous-latent reasoning within a single model traces back to COCONUT [16] and CODI [17], and has been extended to vision-language models in [18]. Homogeneous LLM committees with redundant, near-identical states were formalized as a pathology in [6], along with diversity metrics from the effective-rank family.

B. Semantic Communication and UAV Networking

DeepSC and its successors [7] transmit task-relevant semantics with learned joint source-channel coding; [8] frames semantics as a 6G opportunity; [9] couples semantic coding with knowledge graphs. Cooperative aerial fleets have been studied for mission class-timely inspection of targets by drone squads under energy and trip constraints, validated in real-field experiments as well as simulation [10]; that line optimizes *where drones fly*, whereas we address *what they say to each other* once airborne, so the two are complementary layers of the same system. Cross-layer allocation driven by application-layer state [11] is precisely the design pattern our entropy-gated codec policy would instantiate for semantic traffic. The work [19] addresses load management in communication channels. The authors of work [20] explored the semantics of communication under index modulation via a semantic-aware stream splitting scheme. In the work [21], the authors explored a Hybrid architecture for UAVs powered either solely

by a laser or solely by a battery to improve reach and durability. Finally, cooperative DNN inference through device placement and model partitioning across heterogeneous edge nodes [12] addresses the bottleneck our embodied scaling campaign independently identifies. The work in [22] presents a new channel model that simultaneously accounts for the effects of mobility and shadowing.

Synthesis: After a detailed analysis of the existing methodologies, there exists an unresolved gap at the intersection of strands of communicating informative, non-collapsed LLM latent states over constrained mobile channels while sustaining task-relevant semantics in an embodied decentralized swarm during operation. Current approaches to latent multi-agent systems demonstrate that hidden states, expectation embeddings, or KV representations can replace explicit text and reduce inference overhead, but largely assume lossless, co-located software channels and do not examine mobility, bandwidth, packet loss, or information staleness. Studies of networking for UAVs model constraints on mobility, contention, energy, and edge computing, but treat payload semantics as opaque.

III. SYSTEM MODEL

A. Platform and Timing

Each UAV $i \in \{1, \dots, N\}$, where N is the swarm size, runs a PX4 [23] flight controller consuming 10 Hz position setpoints over MAVLink/UDP, an perception–reasoning stack, CLIP ViT-B/32 [24] for zero-shot anchoring, and Qwen2.5-0.5B [25] for latent generation on a rolling inference interval $T_{\text{inf}} = 3$ s; and a TCP mesh socket for the semantic side-channel. The two loops share no locks, and an inference deadline miss degrades only message freshness, never flight stability. This decoupling is the architectural invariant of the proposed work as depicted in Fig. 1.

B. Message Generation

At each inference tick, drone i captures frame $I_i(t)$, computes the CLIP posterior $p_i(t) \in \Delta^{C-1}$ over $C=13$ scene/anomaly classes and the image embedding $v_i(t) \in \mathbb{R}^{512}$, builds an evidence-verbalized prompt $x_i(t)$, and extracts the last-layer hidden states of the final $k=8$ tokens:

$$h_i(t) = \text{vec}(H_{i,-k}^{(L)}) \in \mathbb{R}^{7168}, \quad d_{\text{model}} = 896. \quad (1)$$

The broadcast state is the whitened residual $\psi_i(t)$. Payload sizes measured on the wire (16-B header included) are 28,688 B (fp32), 14,352 B (fp16), 7,184 B (int8), and 3,600 B (int4) per state.

C. Channel Model

Each directed link ($i \rightarrow j$) applies, in order:

- 1) *Source coding*: quantization $Q_b(\cdot)$ at $b \in \{32, 16, 8, 4\}$ bits per dimension (symmetric per-vector scaling for integer modes), with optional top- s magnitude sparsification charged an explicit 32-bit index cost per retained coefficient.

$$R_{\text{link}} = R_{\text{tot}} / (N(N-1)) \quad (2)$$

TABLE I

RELATED WORKS ACROSS THE TWO CAPABILITY CLUSTERS THIS PROBLEM REQUIRES. ✓ = ADDRESSED; ✗ = NOT ADDRESSED; *p* = PARTIALLY ADDRESSED. THE UPPER BLOCK REPRESENTS LATENT-SPACE MULTI-AGENT COMMUNICATION; THE MIDDLE BLOCK REPRESENTS SEMANTIC COMMUNICATION; THE LOWER BLOCK REPRESENTS UAV NETWORKING AND MOBILE COMPUTING.

Work	Domain	latent-space semantics			communication & platform realism			
		LLM latent payload	Anti-collapse	Measured info. bounds	Channel model	Mobility / node speed	Physical platform	Seeded stats
Interlat [1]	latent MAS	✓	✗	✗	✗	✗	✗	<i>p</i>
LatentMAS [2]	latent MAS	✓	✗	<i>p</i>	✗	✗	✗	✓
CIPHER [3]	latent MAS	✓	✗	✗	✗	✗	✗	<i>p</i>
ThoughtComm [4]	latent MAS	✓	<i>p</i>	<i>p</i>	✗	✗	✗	✓
KVComm [5]	latent MAS	✓	✗	✗	✗	✗	✗	✓
Collapse study [6]	LLM committees	✓	<i>p</i>	✗	✗	✗	✗	✓
DeepSC [7]	semantic comm.	✗	✗	<i>p</i>	✓	✗	✗	✓
Semantics for 6G [8]	semantic comm.	✗	✗	<i>p</i>	✓	✗	✗	<i>p</i>
KG semantic comm. [9]	semantic comm.	✗	✗	✗	✓	✗	✗	✓
Drone squads [10]	UAV fleets	✗	✗	✗	<i>p</i>	✓	✓	✓
CLEVER [11]	HetNet video	✗	✗	✗	✓	✓	✗	✓
Cooperative DNN inf. [12]	edge inference	✗	✗	✗	<i>p</i>	✗	<i>p</i>	✓
AeroLat (proposed work)	UAV latent comm.	✓	✓	✓	✓	✓	✓	✓

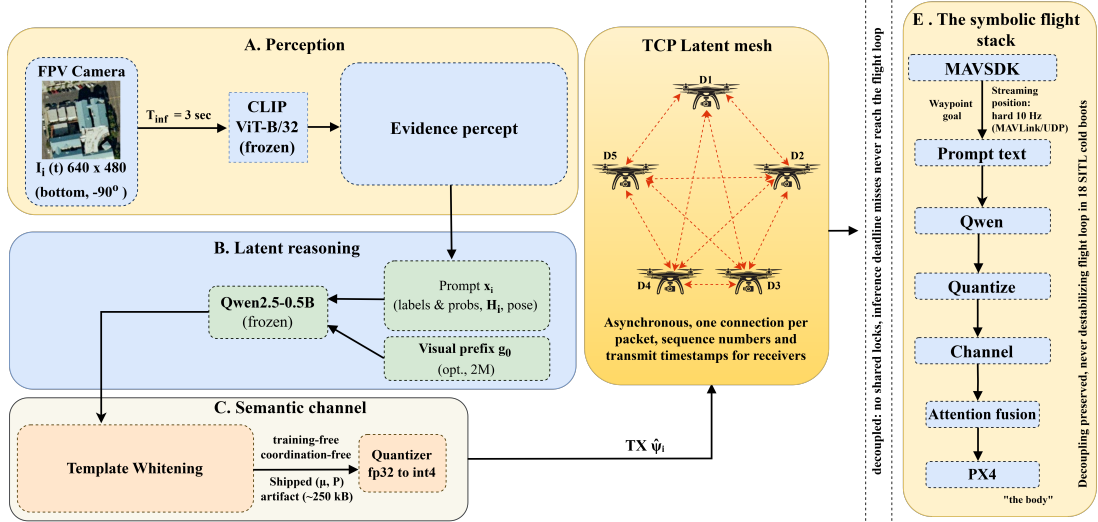


Fig. 1. AeroLat system architecture for one agent D_i . (A) zero-shot CLIP perception emits the full posterior, image embedding, visual entropy, and pose rather than an arg max label; (B) that evidence is verbalized into the prompt of the frozen LLM; (C) template whitening removes the shared common-mode subspace. The centre part depicts the asynchronous TCP latent mesh, one connection per packet with sequence numbers and transmit timestamps so receivers can measure message age. (D) staleness-aware attention fusion; (E) the symbolic flight stack.

- 2) *Noise*: for analog transmission, Additive White Gaussian Noise (AWGN) at per-symbol SNR γ , i.e.,

$$\hat{\psi} = \psi + w, w \sim \mathcal{N}(0, s^2 I) \quad (3)$$

with $s^2 = \mathbb{E}[\psi^2]/10^{\gamma/10}$

- 3) *Staleness*: the receiver discards states older than an age-of-information threshold $A_{\max}$, and fusion discounts survivors by e^{-A/T_a} .

The offered load of the full mesh is $\Lambda(N) = N(N-1)8n_B/T_{\text{inf}}$; with fp32 payloads this evaluates to 0.5, 1.5, 6.9, and 29.1 Mb/s at $N=3, 5, 10, 20$. Quadratic growth is stated plainly here because it motivates both quantization and learned compression.

D. Complexity and Overhead

For each agent and at every inference tick, the computational cost consists of one CLIP forward pass and one LLM forward pass. The whitening projection draws a complexity of $O(mD)$ with $m=4$ and $D=7168$. The incurred cost is negligible relative to the transformer forward passes and is measured at <1 ms. Apart from these, quantization requires $O(D)$ operations and fusion incurs $O(ND)$ complexity, since attention is performed over at most $N-1$ peer states. The per-agent computation is therefore $O(ND)$ beyond the two fixed forward passes, and is independent of mission length because no history is retained. Communication cost is one broadcast of n_B bytes per tick per agent, giving the aggregate offered load $\Lambda(N) = N(N-1)8n_B/T_{\text{inf}}$ as stated above.

TABLE II
LIST OF SYMBOLS

Symbol	Description
$\mathcal{D}_i$	agent (drone) i
T_{inf}	perception/inference interval (3 s)
$I_i(t), v_i(t)$	frame and CLIP image embedding of $\mathcal{D}_i$
$p_i(t), H_i(t)$	CLIP posterior over $C=13$ classes; its entropy
$x_i(t)$	evidence-verbalized prompt
$h_i(t)$	raw latent from last $k=8$ hidden states in $\mathbb{R}^{7168}$
μ	whitening mean
P	top- m template subspace
$\psi_i(t)$	whitened broadcast state
$\hat{\psi}_j$	peer state as received (post-channel)
λ	fusion coefficient
a_{jm}	attention weight
A_m	message age
T_a	staleness discount constant
$Q_b(\cdot)$	quantizer at b bits/dim
n_B	payload bytes
p_ℓ	erasure probability
γ	per-symbol SNR
R_{tot}	shared medium budget
R_{link}	per-link rate
$\Lambda(N)$	aggregate offered mesh load
D_{sem}	semantic distortion $1 - \cos(\psi, \hat{\psi})$
$C(t)$	whitened consensus
v	node speed
f_d	Doppler shift
T_c	coherence time
τ	latency
ρ	information density

E. Metrics

The communication quality, task performance, and swarm-state diversity is evaluated using the following metrics. For M evaluated transmissions (or mission episodes), the semantic-relay success rate is

$$\text{SR} = \frac{1}{M} \sum_{m=1}^M \mathbf{1}\{\mathcal{S}_m\}, \quad (4)$$

where $\mathcal{S}_m$ denotes satisfaction of the task-specific success criterion in episode m . Coordination latency is calculated as,

$$\tau = t_{\text{enc}} + t_{\text{ch}} + t_{\text{int}}, \quad (5)$$

where the three terms denote encoding, channel, and receiver-side interpretation time, respectively. We access the efficacy of transmitted bits in preserving the task-relevant information by evaluation the information density as,

$$\rho = \frac{\hat{I}(\hat{\psi}; Y)}{b}, \quad (6)$$

where $\hat{I}(\hat{\psi}; Y)$ is the estimated mutual information between the received semantic state and task variable Y , and b is the number of transmitted bits. We additionally measure the breadth of the receiver's decoded hypothesis. Let $q_{(1)} \geq q_{(2)} \geq \dots \geq q_{(K)}$ denote the decoded class probabilities in descending order. In order we can define this as,

$$P_{50} = \min \left\{ k : \sum_{j=1}^k q_{(j)} \geq \frac{1}{2} \right\}. \quad (7)$$

Thus, smaller P_{50} indicates a more concentrated decoded belief, whereas larger P_{50} indicates greater hypothesis breadth.

Consensus $C(t)$ is the mean off-diagonal entry of the $N \times N$ Gram matrix $S_{ij} = \cos(\psi_i(t), \psi_j(t))$ of whitened states. Finally, we monitor latent-state collapse using effective rank. Let $\Psi(t) = [\psi_1(t), \dots, \psi_N(t)]^\top \in \mathbb{R}^{N \times D}$ be the stacked swarm latent set, $\bar{\Psi}(t)$ its column-centered form, and $s_1 \geq \dots \geq s_R > 0$ the nonzero singular values of $\bar{\Psi}(t)$. With $\pi_r = s_r / \sum_q s_q$ so that $\pi_r \geq 0$ and $\sum_r \pi_r = 1$, the effective rank is

$$\text{erank} = \exp \left(- \sum_r \pi_r \ln \pi_r \right) \quad (8)$$

The process costs one degree of freedom, giving $\text{erank} \in [1, N-1]$. The measure approaches 1 when the swarm states collapse to a common direction and approaches $N-1$ when they span comparably weighted, independent directions.

IV. THE COLLAPSE PATHOLOGY AND THE AEROLAT FIX

A. Template Whitening

Let $\{h_b(t)\}_{b=1}^B$ be a calibration buffer of raw latents over diverse scenes. With mean μ and the top- m right singular vectors $P \in \mathbb{R}^{m \times D}$ of the centered buffer, every agent broadcasts P spans the template subspace, and the residual carries the scene-specific variation. Because all agents run identical frozen weights and a deterministic calibration protocol, (μ, P) agree across the fleet *without coordination*. All similarity, consensus, fusion, and channel-distortion computations operate on ψ . With $m=4$, the operator costs one $D \times m$ projection per message.

B. Fusion Under Impairments, and a Second Collapse Mechanism

Receiver j fuses its own ψ_j with peer states via scaled dot-product attention *including self*, discounted by message age A_m :

$$\begin{aligned} a_{jm} &\propto \exp(\psi_j^\top \psi_m / \sqrt{D}) e^{-A_m/T_a}, \\ \psi_j^{\text{new}} &= \text{norm} \left((1 - \lambda) \psi_j + \lambda \sum_{m \neq j} \tilde{a}_{jm} \psi_m \right), \end{aligned} \quad (9)$$

The fusion coefficient λ controls the contribution of peer information and is selected empirically. One structural requirement here is that fusion must use the *fresh* percept latent as ψ_j , not the previous fused state. Iterating (9) on fused states is a DeGroot averaging process that converges to a common vector for any $\lambda > 0$ regardless of scene diversity. This leads to a dynamical collapse that is distinct from prompt-induced representational collapse. To avoid repeated averaging across rounds, AeroLat re-initializes the local percept every T_{inf} .

V. SEMANTIC RATE-DISTORTION AND INFORMATION ACCOUNTING

Before assessing the channel's performance, we first characterize the impact of transmission impairments on the geometry and task-relevant information carried by the latent state. The following section consequently analyses AeroLat using complementary distortion and information measures against the considered channel operating points.

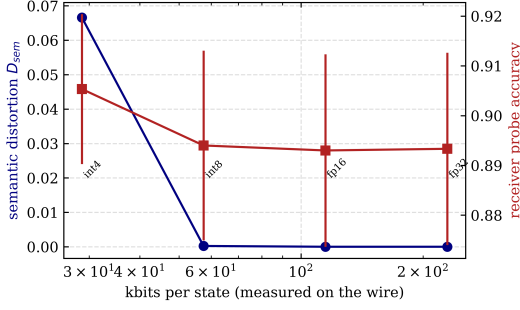


Fig. 2. Semantic rate-distortion over the real latent corpus (ten seeds). Dense quantization from fp32 (229.5 kb/state) to int8 (57.5 kb) is semantically lossless ($D_{\text{sem}} \leq 2 \times 10^{-4}$); int4 (28.8 kb) costs $D_{\text{sem}} = 0.067$ in geometry while receiver probe accuracy remains 0.905: within this codec family, rate can be cut $8\times$ before task distortion begins.

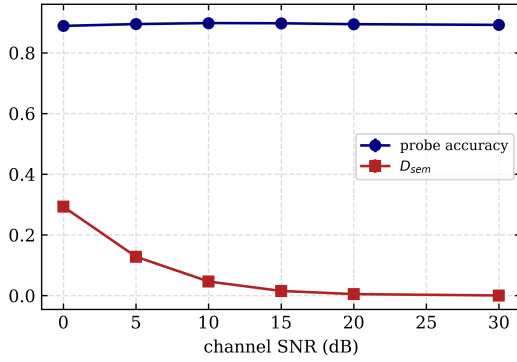


Fig. 3. AWGN robustness. From 30 dB down to 0 dB SNR, the geometric distortion grows ($D_{\text{sem}} 0.000 \rightarrow 0.293$) while receiver probe accuracy is flat ($0.892 \rightarrow 0.889$): the whitened latent code degrades gracefully, with no digital cliff.

A. Empirical Rate-Distortion

Sweeping quantization and sparsity over the corpus yields Fig. 2. Top- s sparsification is reported honestly as *not* competitive at these dimensions once the 32-bit index overhead is included. Generally, for below $\sim 4\times$ compression, a sparse int8 vector costs more wire bytes than a dense one. AWGN robustness appears in Fig. 3. The flat accuracy down to 0 dB is the graceful-degradation profile that a digital symbolic packet, catastrophic under residual bit errors, cannot exhibit.

B. Measured Information Density

We quantify the task-relevant information preserved in the received latent state $\hat{\psi}$ using two lower bounds on $I(\hat{\psi}; Y)$, with Y the scene/anomaly class: a *Fano* bound from probe error e ,

$$I \geq H(Y) - H_b(e) - e \log_2(|\mathcal{Y}| - 1), \quad (10)$$

and a neural-critic *InfoNCE* bound $I \geq \log N_b - \mathcal{L}_{\text{NCE}}$ trained per operating point. Table III reports both together with $\rho = \hat{I}/b$. Quantization is nearly free in task information (Table III), so wire density improves $8\times$ at no measured semantic cost.

VI. LEARNED COMPRESSION: THE FOUR-LOSS PROTOCOL

The four-loss objective below follows the taxonomy introduced by Interlat [1]; our contribution here is empirical—prior

TABLE III
MEASURED INFORMATION PER STATE AND INFORMATION DENSITY ρ (NEURAL INFO NCE AND FANO BOUNDS; TASK CEILING $H(Y) = 2.52$ BITS; THE MARGINAL OVERSHOOT AT 0 DB REFLECTS FINITE-BATCH ESTIMATOR VARIANCE).

Operating point	bits/state	$\hat{I}_{\text{NCE}}$	Fano	ρ (b/kb)
fp32, noiseless	229,504	≥ 2.52	1.91	0.011
fp16	114,816	≥ 2.52	1.91	0.022
int8	57,472	≥ 2.52	1.90	0.044
int4	28,800	≥ 2.52	1.88	0.088
fp32 @ 10 dB	229,504	≥ 2.52	1.83	0.011
int8 @ 10 dB	57,472	≥ 2.52	1.83	0.044
fp32 @ 0 dB	229,504	≥ 2.53	1.78	0.011

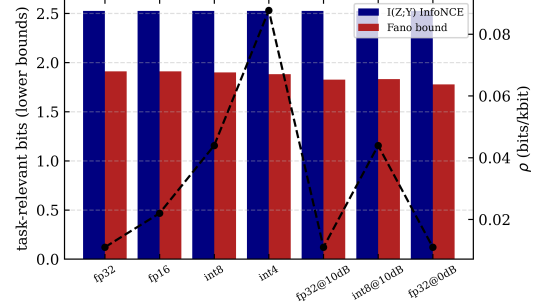


Fig. 4. Task-relevant information (InfoNCE and Fano lower bounds, bars) and density ρ (line) per channel operating point. The state carries the entire task entropy essentially and retains it under int4 quantization and 0 dB AWGN simultaneously.

work states these losses but does not train or dissect them, and the leave-one-loss-out ablations that follow assign each term a measured role. The compressor M_c maps the full latent trajectory H_L to $H_K \in \mathbb{R}^{8 \times 896}$ and is trained with the frozen LLM as measurement device:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + a \mathcal{L}_{\text{sep}} + b \mathcal{L}_{\text{pref}} + c \mathcal{L}_{\text{geom}}, \quad (11)$$

where $\mathcal{L}_{\text{task}}$ is the cross-entropy of the frozen receiver producing the correct structured survey report conditioned on H_K as a soft prefix. $\mathcal{L}_{\text{sep}} = -\text{JS}(p_\theta(\cdot | C, H), p_\theta(\cdot | C, \tilde{H}))$ with batch-mismatched $\tilde{H}$ (the “must-listen” pressure). $\mathcal{L}_{\text{pref}}$ is a confidence-weighted teacher-student KL between full-trajectory and compressed conditioning. Lastly, $\mathcal{L}_{\text{geom}} = 1 - \cos(\bar{z}^{(A)}, \bar{z}^{(D)})$ aligns step-averaged actor features. As shown in Fig. ?? and Table IV, training converges in less than 38 epochs, and the Jensen—Shannon separation between matched and mismatched prefixes increases by approximately 0.1 nat, which confirms dependence on the transmitted latent. Ablations show that $\mathcal{L}_{\text{pref}}$ prevents shortcut learning, $\mathcal{L}_{\text{geom}}$ preserves latent orientation, and $\mathcal{L}_{\text{sep}}$ enforces prefix dependence with little effect on task loss. The optional compressor reduces the int8 payload from 57.5 to 14.3 kb. As shown in Fig. ?? Ablations show that $\mathcal{L}_{\text{pref}}$ prevents shortcut learning, $\mathcal{L}_{\text{geom}}$ preserves latent orientation, and $\mathcal{L}_{\text{sep}}$ enforces prefix dependence with little effect on task loss. The optional compressor reduces the int8 payload from 57.5 to 14.3 kb. The visual-prefix projector further improves the accuracy of validation probes to 0.973.

TABLE IV

COMPRESSOR TRAINING AND LEAVE-ONE-LOSS-OUT ABLATIONS: INITIAL $\rightarrow$ FINAL LOSS VALUES (FULL RUN: 30 EPOCHS; ABLATIONS: 15; FOR PARITY THE FULL RUN READS $\mathcal{L}_{\text{task}}=0.405$ AT EPOCH 14).

Run	$\mathcal{L}_{\text{task}}$	$\mathcal{L}_{\text{sep}}$	$\mathcal{L}_{\text{pref}}$	$\mathcal{L}_{\text{geom}}$
full	2.37 $\rightarrow$ 0.39	-.01 $\rightarrow$ -.11	3.36 $\rightarrow$ 1.52	.38 $\rightarrow$.07
no-sep	$\rightarrow$ 0.39	(off)	$\rightarrow$ 1.55	$\rightarrow$.08
no-pref	$\rightarrow$ 0.008	$\rightarrow$ -.13	(off)	$\rightarrow$.13
no-geom	$\rightarrow$ 0.41	$\rightarrow$ -.10	$\rightarrow$ 1.87	(off)

TABLE V

THE COLLAPSE, REPRODUCED AND REPAIRED ($n=400$ PER CONDITION; CHANCE = 0.167). ROW 4 IS THE EARLIER DRAFT’S FLAGSHIP “1.00 SIMILARITY,” EXPLAINED.

Metric	Argmax prompt (prior system)	Evidence prompt (AeroLat)
Probe accuracy, raw h	0.593 ± 0.071	0.923 ± 0.020
Probe accuracy, whitened ψ	0.500 ± 0.065	0.878 ± 0.037
Fano MI (raw), bits	0.61	1.96
Off-diag. cosine, raw	0.9978	0.9989
Off-diag. cosine, whitened	0.184	0.025
Effective rank, raw $\rightarrow$ whitened	7.9 $\rightarrow$ 4.1	164 $\rightarrow$ 233

VII. EXPERIMENTAL EVALUATION

This section assesses AeroLat across representation quality, channel robustness, baseline performance, and swarm scalability in both controlled and full-stack settings. For the simulation phase, we utilized a high-fidelity Software-In-The-Loop (SITL) architecture that interfaces with a detailed AirSim (Unreal Engine) neighborhood environment [26], providing a continuous, unstructured 3D testing ground.

A. Protocol

We construct the evaluation corpus from 1,000 aerial images across six scene classes in the UC Merced Land Use Dataset [27], a benchmark of high-resolution aerial imagery for land-use classification. The fitted whiteners are exported as the deployment artifact. We perform semantic-layer experiments (probe evaluation, channel sweeps, ablations, and scaling studies) over 10–20 random seeds and report mean $\pm$ standard deviation with 95% bootstrap confidence intervals. Baseline comparisons are done with Welch’s t -test, Mann–Whitney U -test with Holm correction, and Cohen’s d . Full-stack validation is performed using SITL missions on an NVIDIA T1000 (4GB), and semantic-layer experiments and training are conducted on an RTX 6000 Ada, on which Airsim runs.

B. Collapse and Representation Diagnostics

The collapse and its mitigation are summarized in Table V. Under argmax prompting, the mean cosine similarity between raw latents across scenes is 0.9978. Template whitening is what restores discriminative similarity to 0.025 across mixed scenes while increasing latent diversity (erank = 233), indicating that semantic content and representation geometry constitute distinct failure modes. We additionally test the representation on a six-class aerial corpus with $H(Y) = 2.52$ bits and 1,000 samples, with three diagnostics:

- *D1 (probe recoverability)*. Five-fold linear-probing produces 0.915 ± 0.032 accuracy on raw states and $0.916 \pm$

0.030 accuracy after whitening, in comparison to 0.167 chance accuracy predicting the scene class from the broadcast vector.

- *D2 (representational sensitivity)*. With the template fixed and only vision varied, the ratio of between-scene to within-scene latent distances is 1.34.
- *D3 (conflict fusion/superposition)*. Fusing whitened latents from two *conflicting* scenes at $\lambda=0.4$, the fused state’s probe top-2 contains *both* source hypotheses in 92.75% of 400 trials (snapping to a single hypothesis: 7.25%).

C. Baseline Comparison

Table ?? and Fig. 5 report the head-to-head, under a *semantic relay* protocol, where the sender perceives a scene, encodes a message with the codec under test, the message crosses the emulated channel, and the receiver must identify the sender’s scene from the message alone. The JSON baseline remains more efficient in latency and payload size, requiring only a 120-byte message. The latent representation also leads to a sharper decoded belief ($P_{50} = 1.58$) compared to the symbolic baselines (≈ 2.25), while D3 confirms the preservation of conflicting hypotheses in 92.75% of trials. To summarize, AeroLat improves SR by 0.24 over the best symbolic baseline, at the expense of a significantly larger transmitted payload.

D. Ablations

- λ sweep and the two fusion regimes Under evidence-refreshed dynamics (Fig. ??), λ traces a clean trade-off: same-scene consensus rises $0.50 \rightarrow 0.84 \rightarrow 1.00$ as λ goes $0 \rightarrow 0.4 \rightarrow 1.0$, while fused-state probe accuracy holds at 1.000 up to $\lambda=0.3$, is 0.978 at $\lambda=0.4$, and degrades beyond (0.933 at 0.5; 0.489 at 1.0). Under pure *gossip* dynamics (iterating fusion on already-fused states), the swarm converges to cross-scene consensus 1.0 for *any* $\lambda \geq 0.1$ with fused accuracy collapsing to 0.489.
- Component ablations:
 - *Whitening*: After experiments, it was found that probe accuracy remained unchanged ($0.915 \rightarrow 0.916$) while the mean off-diagonal cosine dropped from $0.999 \rightarrow 0.188$ on the full corpus geometry, not content, matching Table V.
 - *CLIP anchor*: With visual evidence removed from the prompt, probe accuracy falls to $0.18 \approx$ chance versus 0.78 with it. The results identified a collapse mechanism and demonstrate that scene-specific visual content originates from the visual anchor. The effect of the number of extracted tokens is shown in Fig. ??, and the ablations in relation to the compressor loss are summarized in Table IV.

E. Mobility: Delay and Delivery Under Node Motion

The channel, as depicted in Sec. III-C, is static, but the aerial nodes are not. Henceforth, extend the link model with a standard block-fading treatment of mobility and sweep it offline. The results found show that int4 quantization is

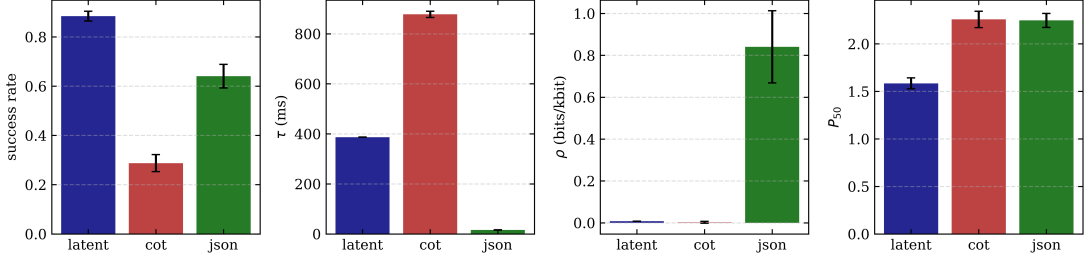


Fig. 5. Baseline comparison across twenty seeds. Left to right: relay success rate, coordination latency τ , information density ρ , and hypothesis breadth P_{50} .

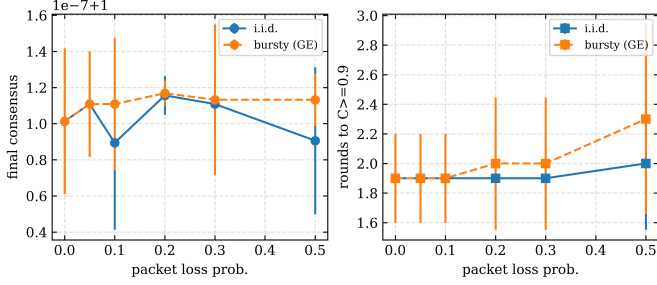


Fig. 6. Consensus under packet loss (i.i.d. and Gilbert–Elliott burst; $N=5$ fusion mesh, ten seeds). Final same-scene consensus is 1.00 at every loss rate up to 50%; loss costs convergence *time* only (1.9 $\rightarrow$ 2.3 rounds).

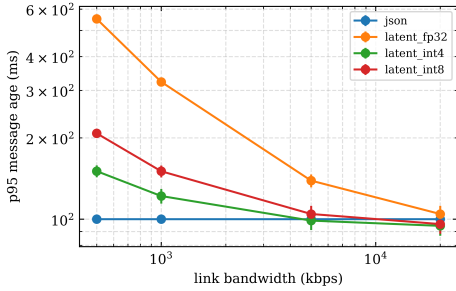


Fig. 7. Age of information under bandwidth caps (50 ms link latency): p95 message age versus link rate per codec. At 500 kb/s, quantization moves the latent message from 552 ms (fp32) to 208 ms (int8) and 151 ms (int4), against JSON’s 100 ms floor; at ≥ 5 Mb/s all codecs converge.

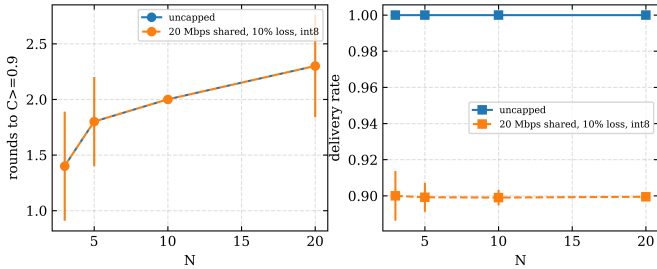


Fig. 8. Mesh scaling (ten seeds): rounds to $C \geq 0.9$ grow sub-linearly (1.4 $\rightarrow$ 2.3 for $N=3 \rightarrow 20$), and under a shared 20 Mb/s medium with int8 payloads and 10% loss the mesh holds 0.90 delivery with unchanged convergence at every N .

effectively *task-lossless*. The mobility results present that the same compression is also crucial for reliable latent communication in time-varying channels. At a 0.9 delivery floor, the fleet supports $N=4$ at fp32 and $N=10$ at int4—the $8\times$ rate



Fig. 9. The frame of what the drones actually see. The heterogeneity is the point: at any instant the swarm holds genuinely different, individually ambiguous views— $\mathcal{D}_1$ reads the spawned cluster as *debris* at only $p=0.25$ with $H=2.28$ nats—and it is exactly this residual uncertainty that the whitened latent channel transports and the argmax-based JSON codec destroys

reduction that costs nothing in task information buys a $2.5\times$ larger swarm. The result provides an operational interpretation of the offered-load trend directly in terms of the scalability and reliability constraints relevant to network design.

F. Full-Stack SITL Campaign

Table VI reports the embodied campaign, and Fig. 9 shows representative onboard percepts from a capture mission. The findings from the Table VI and Fig. 9 are as follows:

- 1) *The task result carries to flight.* AeroLat detected the anomaly in all five missions, compared to the JSON baseline, which detected the anomaly only in one of three runs, and CoT, which failed in both evaluated missions. The difference in performance is due to communication and inference overhead rather than to differences in visual sensing, even though all the stated methodologies use the same perception stack. In particular, CoT took about 3.1 s to generate each message at the edge platform, increasing the mean message age to 22.6 s compared with 51 ms for AeroLat.
- 2) *The collapse is reproduced live.* Thus, the whitened AeroLat representation was much more sensitive to changes in the visual environment.

TABLE VI

EIGHTEEN-MISSION SITL CAMPAIGN. ONE MISSION HANG WAS RECLAIMED BY THE PER-MISSION WATCHDOG; 17/18 COMPLETED NOMINALLY. BOLD MARKS THE VALUES DISCUSSED IN THE TEXT: THE LATENT CODEC’S SUCCESS RATE AND THE TWO FAILURE SIGNATURES (CoT’s MESH FLOODING, THE NO-WHITENING CONTROL’S COLLAPSED CONSENSUS).

Configuration	Missions	SR	Encode ms (med/p95)	Whitened $C(t)$	Delivery	Mesh delay
Symbolic CoT	2	0.00	3,137 / 3,464	0.638 ± 0.083	1.00	22.6 s
JSON dispatch	3	0.33 ± 0.47	71 / 79	0.886 ± 0.041	1.00	39 ms
Latent, dead drone {1, 2, 3}	3	0.67 (2/3)	114–120	0.83 ± 0.06	1.00	13–60 ms
Latent, 20% loss + int8 + 50/25 ms jitter	2	0.50	117 / 128	0.656 ± 0.100	0.80	136 ms
Latent, <i>no whitening</i> (control)	2	0.00	146 / 159	0.999 ± 0.000	1.00	134 ms
Latent	5	1.00	117 / 127	0.678 ± 0.094	1.00	51 ms

TABLE VII

EMBODIED SCALING CAMPAIGN. THE CAMPAIGN IS INCREMENTAL BY DESIGN: EACH SIZE MUST SUCCEED BEFORE THE NEXT IS ATTEMPTED, SO THE FAILURE BOUNDARY IS ITSELF A MEASUREMENT.

N	Missions	SR	Enc. ms	Whitened $C(t)$	erank	Msg age
3	5	1.00	117	0.678 ± 0.094	1.9	51 ms
5	3	1.00	119	0.649 ± 0.070	3.8	78 ms
10	2	1.00	148	0.632 ± 0.068	7.6	2.9 s
15	0/2	watchdog 900 s / boot failure; 11/15 airborne, 132 frames				

G. Embodied Scaling and the Edge-Compute Ceiling

The emulated sweep establishes the communication load that the network can support. A separate flight campaign is used to measure the scalability of the full embodied stack. We attempted $N \in \{5, 10, 15, 20\}$ incrementally (two seeded missions per size, frames, and per-tick latent dumps enabled, automatic abort once a size fails twice), on the same single-edge host that runs the simulator and all N autopilots, and the shared perception-LLM pipeline simultaneously. There are four observations:

- 1) *The task metric is flat where flight completes.* Every completed mission at every size confirmed the anomaly (SR=1.00), and mesh delivery remained 1.00 throughout, across the entire campaign to $N=10$, not one latent packet was dropped.
- 2) *No collapse at scale.* The effective rank of the circulating whitened states grows almost linearly with fleet size from 1.9 to 3.8 and 7.6, corresponding to approximately $0.75N$. This indicates that adding drones adds *dimensions* to the swarm’s shared thought-space rather than redundant copies of a single thought, and that same-scene consensus remains within the discriminative band (0.63–0.68) rather than drifting toward the 0.999 artifact.
- 3) *The cost that grows is compute, not bandwidth.* Median encode latency rises from 117 to 148 ms and, decisively, mean received-message age jumps from 78 ms at $N=5$ to 2.9 s at $N=10$ while delivery stays perfect. The delay is inference queuing on the shared GPU, not the network. The anytime, staleness-aware fusion absorbs this; SR is unharmed, but the trend identifies the true scaling bottleneck on edge silicon. It is the mirror image of the emulated result that the *mesh* holds at $N=20$ as in Fig. 8. The embodiment shows that the *processor* saturates first.

H. Comparison with the existing methodologies

Against Interlat, we share the extraction protocol and loss taxonomy but replace the function-call channel with an emulated lossy, noisy, band-limited network crossed by a live TCP mesh during flight, and we confront the collapse regime that embodied swarms. Against LatentMAS, we share the training-free ethos, but KV-cache working memory presupposes lossless co-located transport, whereas AeroLat’s fixed 7,168-dimensional state is engineered for a budgeted link with quantified distortion. Against the collapse literature [6], we contribute the mechanism, a closed-form decentralized mitigation, and the observation that collapse splits into representational and dynamical channels. Against DeepSC-class semantic communication [7]–[9], we use the evaluation toolkit but transmit frozen LLM native states, achieving graceful degradation without training a codec.

VIII. CONCLUSIONS

AeroLat advances latent inter-agent communication from a software-benchmark mechanism to a network-realistic, statistically validated semantic communication system for UAV swarms. It identifies and mitigates representational collapse, explicitly models the communication impairments experienced by transmitted latent states, and quantifies the task-relevant information preserved across the channel.

As part of future work, we will look at hardware-in-the-loop radio experiments to validate AeroLat under measured fading, interference, and contention. We also plan to extend the framework to heterogeneous UAV fleets via latent alignment among different frozen models.

REFERENCES

- [1] Z. Du, R. Wang, H. Bai, Z. Cao, X. Zhu, Y. Cheng, B. Zheng, W. Chen, and H. Ying, “Enabling agents to communicate entirely in latent space,” in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026, pp. 27 106–27 129.
- [2] J. Zou, R. Qiu, G. Li, X. Yang, K. Tieu, P. Lu, K. Shen, H. Tong, Y. Choi, J. He *et al.*, “Latent collaboration in multi-agent systems,” *arXiv preprint arXiv:2511.20639*, 2025.
- [3] C. Pham, B. Liu, Y. Yang, Z. Chen, T. Liu, J. Yuan, B. Plummer, Z. Wang, and H. Yang, “Let models speak ciphers: Multiagent debate through embeddings,” in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 51 899–51 928.
- [4] Y. Zheng, Z. Zhao, Z. Li, Y. Xie, M. Gao, L. Zhang, and K. Zhang, “Thought communication in multiagent collaboration,” *Advances in Neural Information Processing Systems*, vol. 38, pp. 123 389–123 418, 2026.

- [5] H. Ye, Z. Gao, M. Ma, Q. Wang, Y. Fu, M.-Y. Chung, Y. Lin, Z. Liu, J. Zhang, D. Zhuo *et al.*, “Kvcomm: Online cross-context kv-cache communication for efficient llm-based multi-agent systems,” *Advances in Neural Information Processing Systems*, vol. 38, pp. 17 882–17 928, 2026.
- [6] D. Patel, “Representational collapse in multi-agent llm committees: Measurement and diversity-aware consensus,” *arXiv preprint arXiv:2604.03809*, 2026.
- [7] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, “Deep learning enabled semantic communication systems,” *IEEE transactions on signal processing*, vol. 69, pp. 2663–2675, 2021.
- [8] M. Sana and E. C. Strinati, “Learning semantics: An opportunity for effective 6g communications,” in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2022, pp. 631–636.
- [9] N. Hello, P. Di Lorenzo, and E. C. Strinati, “Semantic communication enhanced by knowledge graph representation learning,” in *2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2024, pp. 876–880.
- [10] N. Bartolini, A. Coletta, G. Maselli *et al.*, “A multi-trip task assignment for early target inspection in squads of aerial drones,” *IEEE Transactions on Mobile Computing*, vol. 20, no. 11, pp. 3099–3116, 2020.
- [11] S. Colonnese, F. Cuomo, L. Chiaraviglio, V. Salvatore, T. Melodia, and I. Rubin, “Clever: A cooperative and cross-layer approach to video streaming in hetnets,” *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1497–1510, 2017.
- [12] P. Dai, B. Han, K. Li, X. Xu, H. Xing, and K. Liu, “Joint optimization of device placement and model partitioning for cooperative dnn inference in heterogeneous edge computing,” *IEEE Transactions on Mobile Computing*, vol. 24, no. 1, pp. 210–226, 2024.
- [13] S. Wu, Y. Wang, and Q. Yao, “Dense communication between language models,” *arXiv e-prints*, pp. arXiv–2505, 2025.
- [14] L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodolà, “Relative representations enable zero-shot latent space communication,” *arXiv preprint arXiv:2209.15430*, 2022.
- [15] V. Maiorca, L. Moschella, A. Norelli, M. Fumero, F. Locatello, and E. Rodolà, “Latent space translation via semantic alignment,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 55 394–55 414, 2023.
- [16] S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. Weston, and Y. Tian, “Training large language models to reason in a continuous latent space, 2024,” URL <https://arxiv.org/abs/2412.06769>, vol. 98, 2022.
- [17] Z. Shen, H. Yan, L. Zhang, Z. Hu, Y. Du, and Y. He, “Codi: Compressing chain-of-thought into continuous space via self-distillation,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 677–693.
- [18] T.-H. Pham and C. Ngo, “Multimodal chain of continuous thought for latent-space reasoning in vision-language models,” *arXiv preprint arXiv:2508.12587*, 2025.
- [19] I. Maity, S. Misra, and C. Mandal, “Core: Prediction-based control plane load reduction in software-defined iot networks,” *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1835–1844, 2020.
- [20] X. Guo, Y. Xu, D. He, H. Hong, Z. Chen, C. Zhang, Y. Wu, and W. Zhang, “Spatial semantic communication: When semantic transmission meets index modulation,” *arXiv preprint arXiv:2607.19934*, 2026.
- [21] A. M. Abdelhady, C. Diaz-Vilor, M. Barzegaran, H. Jafarkhani, and A. M. Eltawil, “Optimization of hybrid laser-battery-powered uav-assisted backscatter communications,” *IEEE Transactions on Communications*, vol. 73, no. 11, pp. 11 690–11 706, 2025.
- [22] P. S. Bithas, V. Nikolaidis, A. G. Kanatas, and G. K. Karagiannidis, “Uav-to-ground communications: Channel modeling and uav selection,” *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5135–5144, 2020.
- [23] L. Meier, D. Honegger, and M. Pollefeys, “Px4: A node-based multithreaded open source robotics framework for deeply embedded platforms,” in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 6235–6240.
- [24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
- [25] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu *et al.*, “Qwen2. 5-coder technical report,” *arXiv preprint arXiv:2409.12186*, 2024.
- [26] S. Shah, D. Dey, C. Lovett, and A. Kapoor, “Airsim: High-fidelity visual and physical simulation for autonomous vehicles,” in *Field and service robotics: Results of the 11th international conference*. Springer, 2017, pp. 621–635.
- [27] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. 18th ACM SIGSPATIAL Int. Conf. Advances in Geographic Information Systems (GIS)*, San Jose, CA, USA, 2010, pp. 270–279.