ACM SIGKDD 2026 (KDD ’26)

中国电信 CHINA TELECOM

Edge–Cloud Collaborative Multimodal Dehazing for UAV Vision

Communication-efficient intelligent sensing under complex weather conditions

Sensing + Communication + Computing, co-designed

Junwei Zhao · Qianchun Luo · Jie Wu*

China Telecom Cloud Computing Research Institute

Supported by the China Telecom Key Research Project on Intelligent Ubiquitous Cloud Technologies (No. 65996).

Low-Altitude Perception: Demand and Constraints

Infrastructure inspection | Emergency response | Low-altitude economy | Autonomous aerial sensing

WHAT CONSTRAINS IT

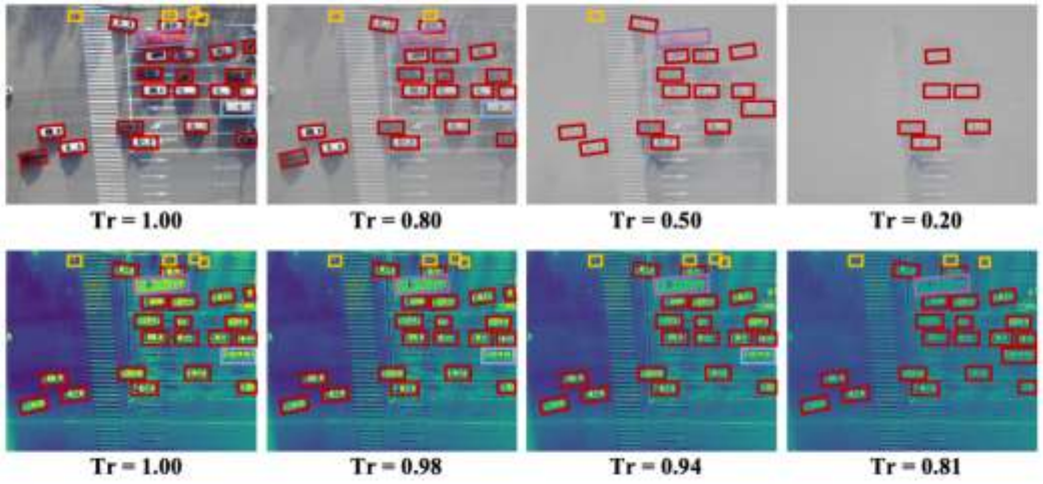
1 **Device**
Onboard compute and power budgets are tight — full-capacity models do not fit.

2 **Link**
The air-to-ground uplink is narrow and fluctuates; the per-frame bit budget is not stable.

3 **Environment**
Dense haze scatters the visible band and erases small ground targets.

4 **Cloud**
Capacity is abundant — but only reachable through constraint 2.

VISIBLE VS. INFRARED AS HAZE INCREASES



Top: visible band — detections collapse as transmittance falls to 0.20.
Bottom: thermal infrared — the same targets persist. Source: paper Fig. 2(c).

Why Neither Cloud-Only nor Edge-Only Works

Cloud-only

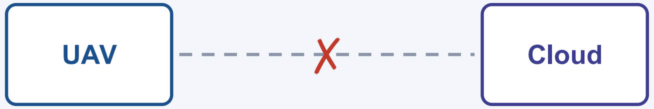
ship everything up, compute remotely

+65.5 kbit/frame

- ✓ Full model capacity
- ✗ Second modality floods the uplink
- ✗ Transmission-bound latency

Edge-only

keep everything onboard

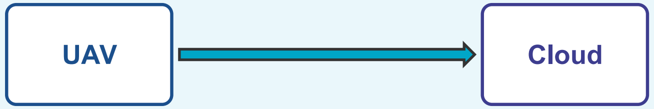


no uplink — but no capacity

- ✓ Lowest latency
- ✗ 4 of 7 baselines exceed Orin Nano memory
- ✗ Fastest that fits still needs 128 ms/frame

Collaborative

decide at the edge, reconstruct in the cloud



+6.6 kbit/frame

- ✓ Full capacity, on the cloud side
- ✓ Sparse, task-selected uplink
- ✓ Bounded latency under a weak link

How do we jointly optimise computation and communication?

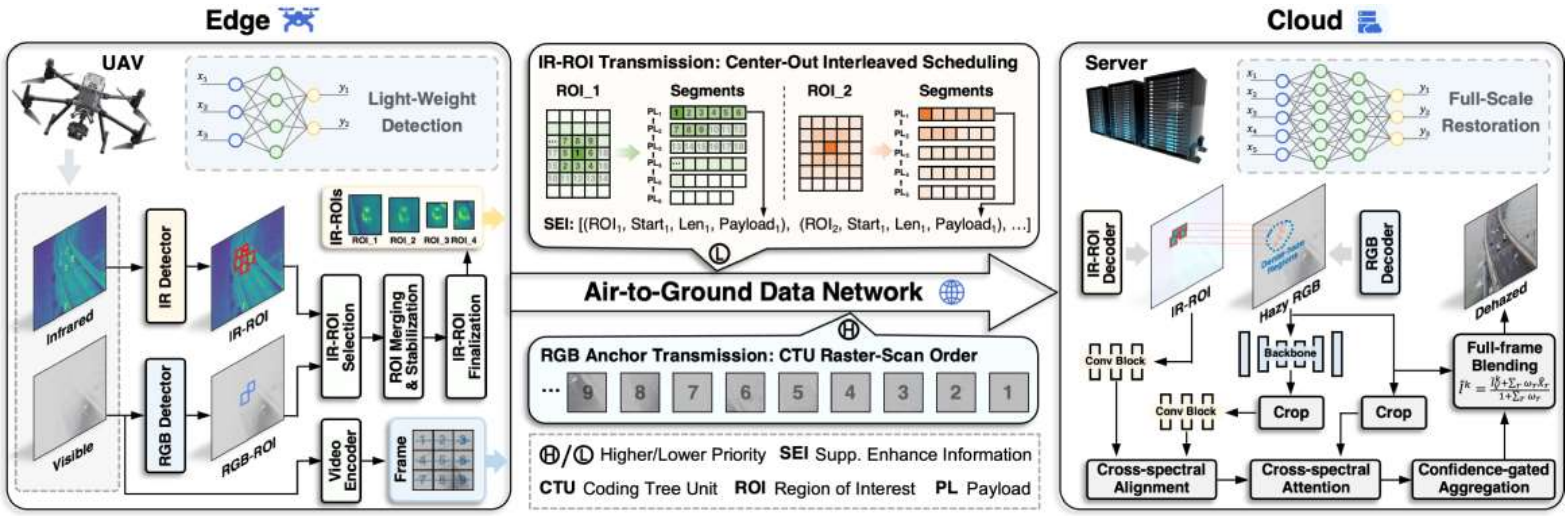
Not “which dehazing model is best”, but: given a fixed number of bits per frame, which bits do you spend?

EVIDENCE · TABLE 1

Full-frame IR alone costs 65.5 kbit/frame — 65.5 ms at 1 Mbps, twice the 33.3 ms budget of a 30 fps stream.

System Overview: The Collaborative Pipeline

Three stages under one contract — the edge decides where, the link carries only that, the cloud reconstructs in parallel.



Edge decides where

VIS is the anchor stream; infrared is requested only where visible evidence has collapsed.

Link carries only that

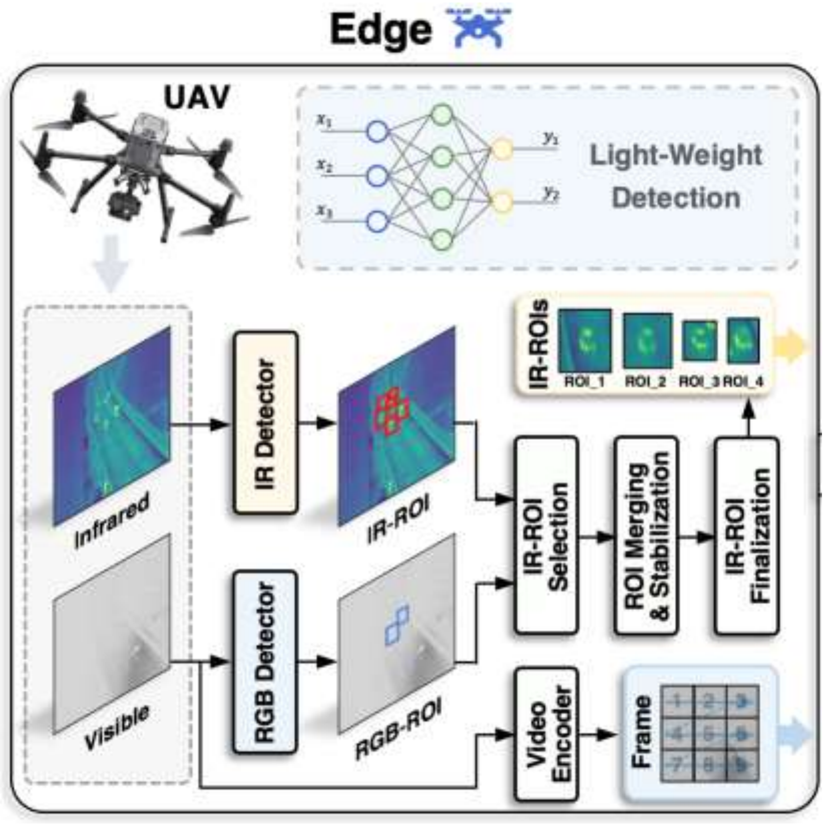
A single H.26x track — IR pixels embedded in-band as SEI, with ROI cores ordered first.

Cloud reconstructs in parallel

Every ROI processed concurrently on GPU, then confidence-gated blending into the frame.

Edge Side: Deciding What Deserves the Uplink

The edge does not compress the image — it performs selection, a far cheaper computation and a far larger bandwidth saving.



Edge stage of the framework (paper Fig. 3, left panel)

THE SELECTION RULE, IN WORDS

Transmit infrared where the IR detector is confident and no visible detection overlaps it — that is, where infrared carries information the visible stream has already lost.

$\Delta = \hat{s}_{IR} \cdot (1 - \mathrm{IoU}_{max})$ · transmit if $\Delta \geq \theta_{diff}$ and area $\geq A_{min}$ · tuned optimum $\theta_{diff} = 0.4$, $A_{min} = 576\ \mathrm{px}^2$ (Table 5)

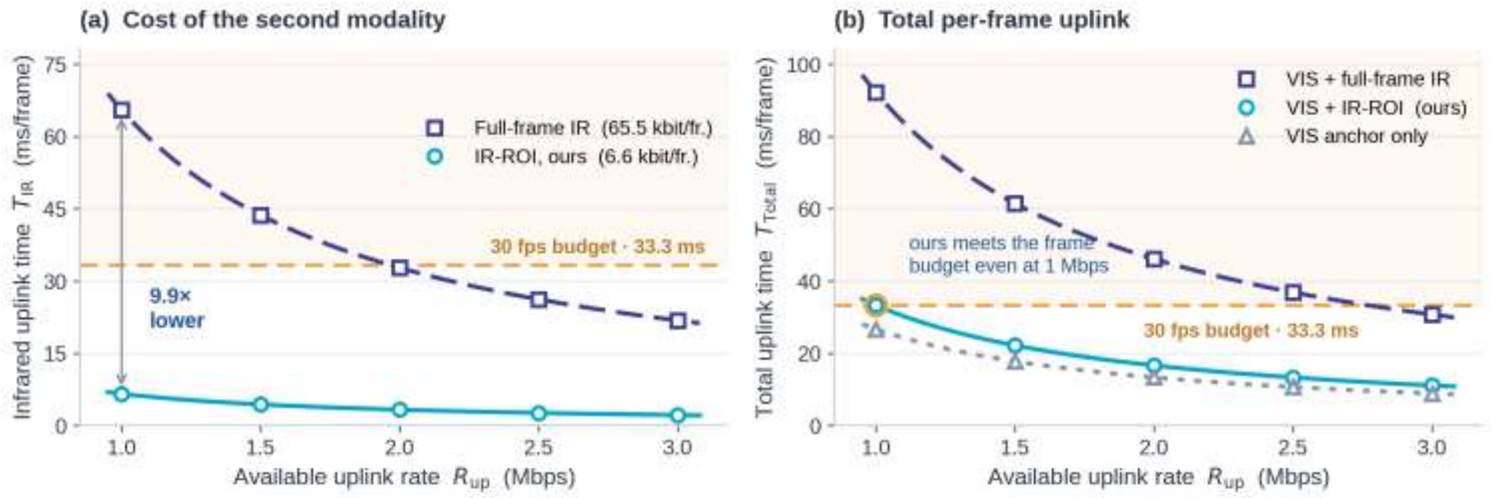
WHAT RUNS AT THE EDGE — AND WHY IT BELONGS THERE

Module	Cost profile	Why the edge
Lightweight VIS + IR detectors	small, fixed-cost inference	must see raw sensor data before encoding
Differential ROI criterion	arithmetic on box lists	needs both modalities co-located
Merge · stabilise · track IDs	negligible	keeps the payload temporally coherent
Encoder + in-band SEI packing	hardware codec	the bit decision belongs where bits originate

6.6 kbit per frame — the only infrared that leaves the aircraft. The payload scales with the area of degraded regions, not with frame size.

Communication-Efficient Transmission

Same restoration quality at roughly one tenth of the infrared uplink — the ratio, not the raw gain, is the contribution.



Analytical uplink model: 512×512 inputs, 30 fps, RGB anchor fixed at 0.8 Mbps. Not a live-network measurement.

6.6 kbit

infrared payload per frame, against 65.5 kbit for full-frame IR

− 89.9 %

infrared uplink volume

− 63.9 %

total per-frame uplink

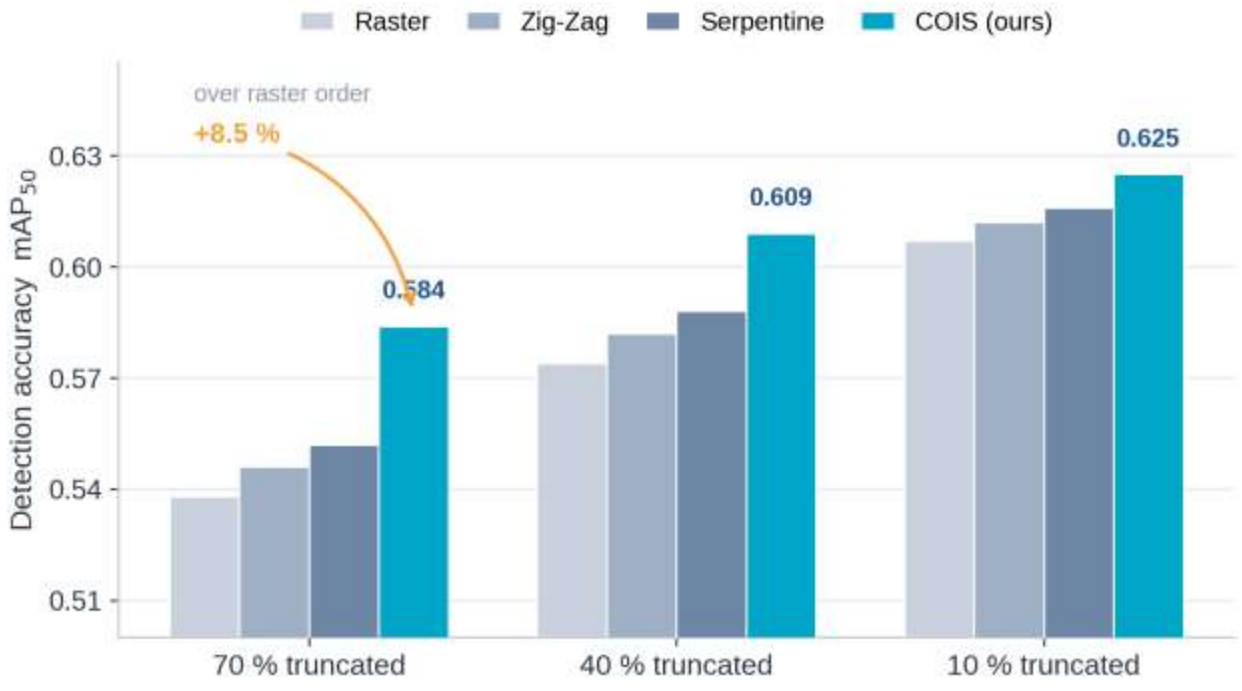
THE COMMUNICATION–COMPUTATION TRADE-OFF IS MEASURED, NOT ASSERTED (VTUAV, Table 4)

Strategy	PSNR	mAP50	Cloud time	IR uplink
VIS anchor only	28.63	0.527	24 ms	0
+ full-frame IR	32.07	0.635	52 ms	65.5 kbit
+ IR-ROI (ours)	**31.92**	**0.632**	**31 ms**	**6.6 kbit**

Give up 0.15 dB — 0.5 % — and buy 40 % of the cloud compute and 90 % of the infrared bits.

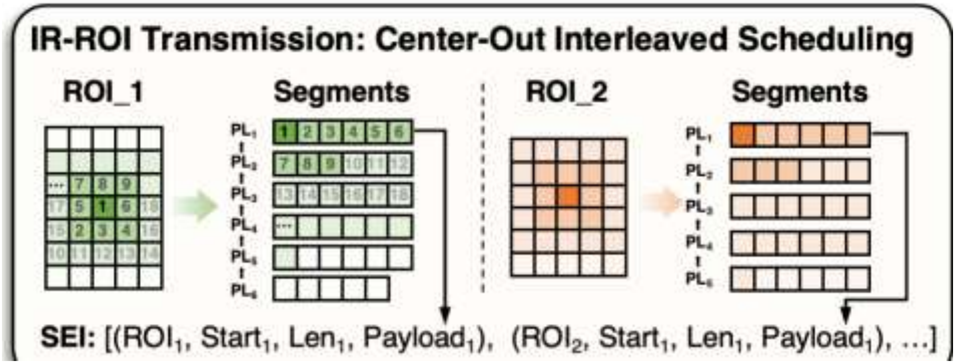
Robustness When the Uplink Fluctuates

Bandwidth loss is modelled as prefix truncation — core-first scheduling keeps a truncated stream semantically useful.



VTUAV · SEI prefix truncation at three levels · four serialisation orders (Table 6)

COIS gains are largest where the link is worst: **+8.5 %** mAP50 at 70 % truncation, tapering to **+3.0 %** at 10 %.



Center-out spiral

Each ROI is serialised from its core outward, so the earliest bits carry the most informative pixels.

Interleaved across ROIs

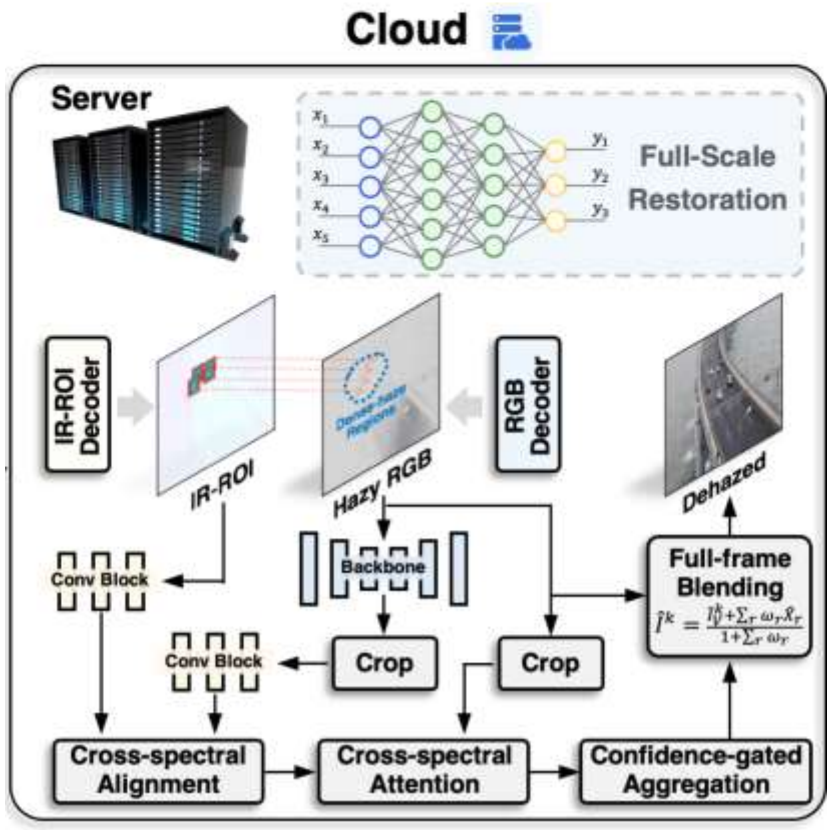
Segments from all active ROIs are interleaved core-first, so no single region monopolises a short prefix.

Any prefix is useful

COIS leads at every truncation level and against every alternative order.

Cloud Side: Capacity Where It Actually Pays Off

The cloud's advantage is parallelism as much as capacity — that is what keeps the frame deadline.



Cloud stage of the framework (paper Fig. 3, right panel)

1. **Scalable parallel execution**
   All ROIs of a frame run concurrently on GPU — cloud time falls from 52 ms (full-frame IR) to 31 ms.

2. **Capacity the device cannot host**
   Cross-spectral alignment and attention that would not fit in onboard memory.

3. **Three-stage processing**
   Align (coarse-to-fine displacement + confidence) → attend (inject IR residual where alignment is trusted) → aggregate (confidence-gated blending).

PIPELINED, SO THROUGHPUT IS max(·) NOT sum(·)

Edge **26 ms**	Cloud **31 ms**	→ max(26, 31) = 31 ms ≤ 33.3 ms

Graceful degradation floor. If no infrared arrives — link outage, no ROI triggered, truncated payload — the pixel falls back to the visible anchor. Enhancement is additive, never load-bearing.

What Actually Enables the Cooperation

Three mechanisms — each solving an edge–cloud systems problem.

1

Where to spend bits

PROBLEM

The uplink budget is smaller than the second modality.

MECHANISM

VIS–IR differential ROI selection — transmit infrared only where visible evidence has collapsed.

EVIDENCE

65.5 → 6.6 kbit/frame at −0.15 dB

Tables 4, 10

2

Surviving a fluctuating link

PROBLEM

Bits available per frame are not known in advance.

MECHANISM

ViStream-SEI with COIS — in-band single-track transport, ROI cores scheduled first.

EVIDENCE

+8.5 % mAP50 at 70 % truncation

Table 6

3

Fusing across a lossy boundary

PROBLEM

VIS and IR arrive misaligned in time and space.

MECHANISM

Confidence-gated cross-spectral fusion — align, gate, and fall back to the anchor when unsure.

EVIDENCE

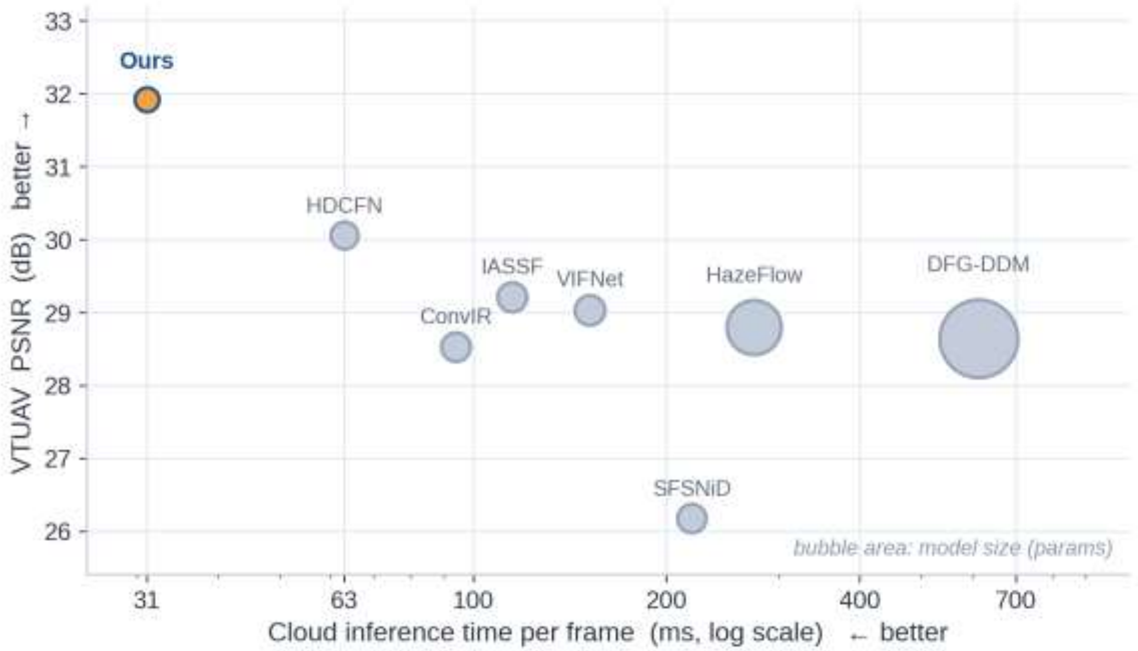
+1.55 dB over HDCFN at 15 px shift

Table 9

Experimental Validation Across Four System Axes

Validated as a system — quality, communication, computation and a real aircraft — not only as an image-quality benchmark.

① ③ QUALITY–COST FRONTIER · VTUAV, 7 baselines



+1.86 dB PSNR over the best baseline · **−49 %** parameters · **4.9×** faster on the edge device

② COMMUNICATION EFFICIENCY

−89.9 % infrared uplink · **+8.5 %** mAP50 at 70 % truncation

① DOWNSTREAM PERCEPTION · YOLOv8m on VTUAV

Metric	HDCFN	Ours	Gain
mAP50	0.581	0.632	+8.8 %
mAR50	0.656	0.702	+7.0 %

The recall gain is the meaningful one — fewer missed small, low-contrast targets.

④ REAL AIRCRAFT, REAL WEATHER

hazy input → ours

Zenmuse H20T UAV · ~15,000 real VIS–IR frames · best NIQE 15.70 / PIQE 34.89

Deployment Concept: Cloud–Network–Edge Inspection

Architectural positioning — not evaluated in the paper. The paper models uplink at 1–3 Mbps citing an LTE–UAV feasibility study; it runs no 5G field trial, no end-to-end glass-to-glass latency measurement, and no multi-UAV scheduling evaluation.

UAV sensing	Edge gateway	Mobile network	Cloud AI platform	Decision support
VIS + IR capture	ROI selection, in-band packing 26 ms/frame, Orin-class device	uplink-constrained bearer payload shaped to 1–3 Mbps	ROI-parallel enhancement 31 ms/frame on GPU	defect and hazard detection, alerting, archival

Standards-compatible transport

The infrared payload rides as user_data_unregistered SEI inside a single H.26x track. No second stream, no bespoke protocol, no change to the bearer — it traverses ordinary video infrastructure.

Degradation is graceful, not cliff-edged

Bandwidth loss costs enhancement quality, never stream integrity. The visible anchor always survives, so the analytics pipeline never goes dark.

WHAT THIS WORK CONTRIBUTES AT EACH LAYER

Layer	Operator asset	Contribution of this work
Aircraft / edge gateway	UAV + onboard compute (Orin-class)	26 ms/frame selection stage — fits the device budget
Access network	mobile uplink bearer	payload shaped to a 1–3 Mbps budget; single standard video track
Cloud	GPU inference platform	ROI-parallel enhancement at 31 ms/frame

Contributions: Computing, Communication, Intelligence

Three separable contributions — and it is worth being precise about which is which.

COMPUTING

system architecture contribution

Collaborative workload allocation

An asymmetric split: cheap, decisive selection at the edge; expensive, parallel reconstruction in the cloud — executed as a pipeline.

- 26 ms edge / 31 ms cloud
- throughput = max(·), not sum(·)
- 3.47 M parameters — lowest in the comparison

COMMUNICATION

communication efficiency contribution

ViStream-SEI with COIS

Single-track, in-band ROI transport whose any prefix is useful — and which needs no change to the bearer.

- −89.9 % infrared uplink at −0.15 dB
- +8.5 % mAP50 under 70 % truncation
- standards-compatible H.26x SEI

INTELLIGENCE

algorithm contribution

Differential selection + gated fusion

Infrared used exactly where the visible band has failed, gated by cross-spectral alignment confidence.

- +1.86 dB PSNR · +8.8 % mAP50
- +7.0 % mAR50
- robust to ±2-frame and 15-px misalignment

The system contribution is the ratio, not any single number: **0.5 % of quality for 90 % of the bits and 40 % of the cloud compute.**

Thanks for watching!