

Sustainable Intelligence for the Wild: Democratizing Ecological Monitoring via Knowledge-Adaptive Edge Expert Agents

Jiaxing Li¹, Hao Fang¹, Chi Xu¹, Miao Zhang¹, Jiangchuan Liu¹, William I. Atlas²,
Katrina M. Connors³ and Mark A. Spoljaric⁴

¹Simon Fraser University, Vancouver, Canada

²Wild Salmon Center, Portland, USA

³Pacific Salmon Foundation, Vancouver, Canada

⁴Haida Fisheries Program, Skidegate, Canada

{jla641,fanghaof,chix,mza94,jcliu}@sfu.ca, watlas@wildsalmoncenter.org,
kconnors@psf.ca, mark.spoljaric@haidanation.com

Abstract

Rapid biodiversity loss underscore the urgency of effective monitoring, yet manual surveys remain resource-intensive. While on-device AI offers a scalable alternative, its performance in the wild is often challenged by environmental variability. Current methods rely heavily on cloud resource, which requires continuous uploading of field data for model retraining. This approach is unsuitable for remote deployments because it consumes limited power and network connectivity. To address these constraints, this research proposes a shift from model adaptation to knowledge adaptation. We introduce an architecture that separates visual perception from reasoning, combining a visual encoder with a dynamic knowledge base. We use an explicit knowledge base to replace implicitly encoding expert knowledge into model parameters. This method also supports knowledge sustainability by preserving expert insights in a structured form. Through cross-disciplinary collaboration with biologists and Indigenous communities, this work advances ethical AI co-development, fostering responsible and culturally informed ecosystem management.

1 Problem Statement

Global ecosystems are undergoing rapid disruption, driving widespread species declines across terrestrial, marine, and freshwater habitats, with freshwater species experiencing the steepest losses [World Wide Fund for Nature, 2024; Waples *et al.*, 2008; Di Lorenzo and Mantua, 2016; Frölicher and Laufkötter, 2018; Kilduff *et al.*, 2015; Dörner *et al.*, 2018]. Reliable wildlife sensing and analysis is therefore essential for timely, evidence-based conservation decisions [Kshitiz *et al.*, 2023; Gordon *et al.*, 2023; Díaz Musmanni, 2023; Atlas *et al.*, 2021; Schindler and Hilborn, 2015]. Yet in practice, ecological monitoring still scales through manual surveys and expert annotation: modern camera-trap and in-river video systems can generate millions of images per

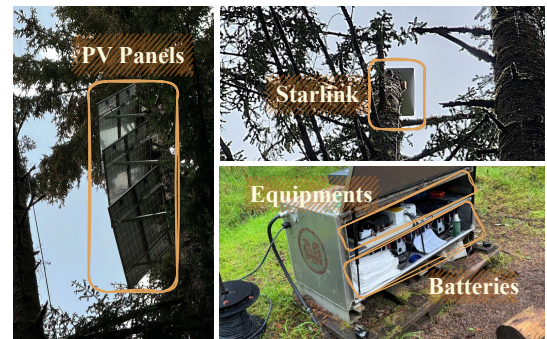


Figure 1: The Off-grid Reality: Our pilot station relies entirely on solar harvesting (PV Panels) and satellite transmission (Starlink), creating strict energy budgets.

season, while expert review requires tens of thousands of hours—creating persistent data backlogs and delayed management actions [Nature Tech Collective, 2025; Norouzzadeh *et al.*, 2018].

Prior work has progressively shifted salmon monitoring from manual review to automated sensing and learning. Video-based weir systems first applied computer vision to expedite in-season fish counting [Atlas *et al.*, 2023], while sonar-based pipelines addressed low-visibility rivers through detection and tracking in sonar videos [Kay *et al.*, 2022; Kay *et al.*, 2024]. SALINA further advanced this direction with real-time sonar analytics and energy-aware deployment in remote ecosystems within Indigenous territories [Xu *et al.*, 2024]. [Xu *et al.*, 2025] explored multimodal foundation AI with expert-in-the-loop workflows to support species identification, counting, and length measurement for sustainable fisheries management in infrastructure-limited Indigenous rivers. FUSED further studied sovereignty-preserving multimodal retrieval across independently governed ecological data domains [Xu *et al.*, 2026]. Yet field deployment remains dominated by the prevailing **model adaptation** paradigm (i.e. continually uploading raw data for annotation and cloud retraining, then downloading and deploying updated model weights.) [Xu *et al.*, 2025; Folkman *et al.*, 2025], which still breaks under off-grid operation and fails to

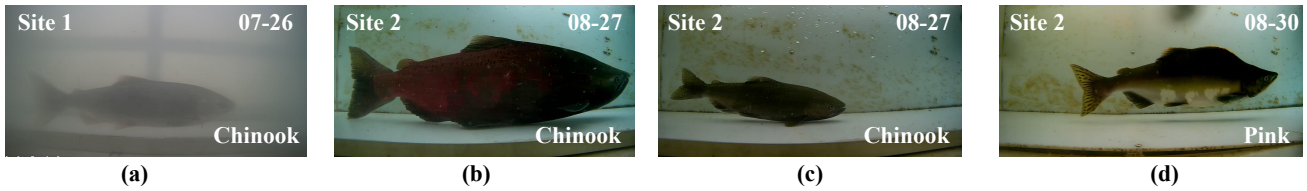


Figure 2: The Adaptation Gap: Visual traits of salmon species drift significantly across space (Estuary vs. Spawning grounds) and time, causing static models to fail.

preserve expert logic. This gap leads to two core barriers: an operational barrier and an knowledge barrier.

1.1 The Operational Barrier: Infrastructure Constrained AI in Off-Grid Rivers

The first barrier is operational. Many priority watersheds are *remote ecosystems that lack basic infrastructure support*, where deployments must sustain sensing, compute, and data delivery under strict resource constraints [Axford *et al.*, 2024]. In our pilot station (Figure 1), the entire pipeline runs off-grid using solar harvesting (PV panels) and battery storage, while wide-area communication depends on a high-wattage satellite terminal (Starlink). This coupling creates a hard trade-off between *ecological coverage* and *system up-time*: continuous connectivity competes directly with the energy budget required for sensing, local storage, and on-device inference [Xu *et al.*, 2024].

Our pilot deployment directly reveals the severity of this constraint. In temperate rainforest conditions with frequent overcast periods, maintaining always-on connectivity quickly depletes battery reserves to safety cutoffs (30%) within 72 hours, forcing a departure from “cloud-first” assumptions toward sparse, low-power communication that restricts connectivity to short daily windows (~2 hours/day).

These realities further expose the limits of the prevailing **model adaptation** paradigm [Xu *et al.*, 2025; Folkman *et al.*, 2025]. In infrastructure-constrained wild, model updates often require repeatedly transferring gigabytes of raw video over satellite links (e.g., Starlink with an average up-link of 14.84 Mbps [Hill, 2025]) for cloud annotation and retraining, and then pushing updated checkpoints back to the field. This closed loop is costly in energy, bandwidth, and turnaround time, and increase operational cost and carbon footprint [Yang *et al.*, 2024; Xu *et al.*, 2024].

1.2 The Knowledge Barrier: From Vanishing Expertise to Executable Digital Heritage

The second barrier is knowledge. Salmon appearance and environmental context shift rapidly across space and time (Figure 2). Even within a single species, visual traits drift throughout the migration cycle: a Chinook salmon near an estuary appear silvery and streamlined, it transforms into darker or bright red spawning forms with distinct skeletal changes as it moves upstream [Axford *et al.*, 2024].

The model adaptation paradigm attempts to absorb field expertise by training on large-scale datasets, implicitly encoding domain knowledge into model parameters. However, spatiotemporal drift can still cause performance degradation when models are deployed across different sites, seasons, and

life stages. Meanwhile, failures are hard to trace, and new knowledge cannot be iterated or incorporated quickly in a principled, transparent way [Xu *et al.*, 2025; Liu *et al.*, 2024].

Furthermore, our stakeholder engagements underscore a compounding crisis: expert scarcity and the challenge of retaining practical field expertise [Wu *et al.*, 2022]. High-quality monitoring relies heavily on senior field biologists and Indigenous stewardship practitioners, whose knowledge is often tacit, highly localized, and recorded only in fragmented notes [Paul *et al.*, 2024; Ma *et al.*, 2024; Liu *et al.*, 2024]. As experts retire or rotate across watersheds, this “river logic” risks being lost—precisely when climate-driven variability increases the need for adaptive decisions.

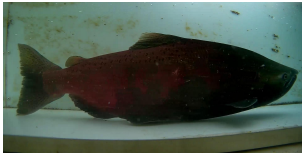
This motivates a shift beyond black-box model adaptation and single-metric “better performance”: through active collaboration with domain experts, we pursue a new application paradigm that jointly targets (i) **sustainable AI** for infrastructure-constrained wild deployments and (ii) **knowledge sustainability** that preserves and operationalizes localized stewardship logic.

2 Alignment with Sustainable Development Goals and the LNOB principle

The 2030 Agenda for Sustainable Development outlines a blueprint for global peace and prosperity [United Nations, 2015]. A core tenet, the Leaving No One Behind (LNOB) principle, underscores the necessity of extending technological advancements to vulnerable communities and fragile ecosystems [United Nations, 2021]. This research shifts from data-intensive cloud retraining to bandwidth-efficient knowledge adaptation. In doing so, our architecture aligns with several Sustainable Development Goals (SDGs):

SDG 13: Climate Action (Green AI). Conventional AI monitoring incurs high carbon costs due to frequent cloud retraining and high-bandwidth transmission. This research supports SDG 13.2 (“Integrate climate change measures”) by adopting a resource-efficient architecture. The system minimizes computational energy by substituting intensive parameter updates with lightweight knowledge graph maintenance. Additionally, it reduces communication costs by transmitting compact semantic text instead of raw video.

SDG 14: Life Below Water & SDG 15: Life on Land. Effective biodiversity conservation depends on timely, granular ecological data, yet manual surveys remain prohibitively expensive and difficult to scale. This research advances SDG 14.2 (“Sustainably manage marine ecosystems”) and SDG 15.5 (“Halt biodiversity loss”) by deploying autonomous agents capable of fine-grained species identification and behavioral analysis in the wild.



Prompt: Identify the species of salmon in this video frame. The possible species are: Chinook, Sockeye, Pink, Coho, and Chum. Describe which visual features (such as body shape, coloration, markings, mouth shape, or fin characteristics) support your classification.

Vision Language Model: This is a **Sockeye salmon**. It has a **deep red body**, especially on the sides, with a **dark green-black back** and a **blackish head**. The **body is robust and elongated**, with a **slightly pointed snout** and **large, dark spots along the sides**. Its **fins are dark**, and the **tail fin has a distinct spotted pattern**.

Figure 3: The semantic gap: A general-purpose VLM (Qwen3-VL-8B) identifies visual cues (green) but hallucinates the species label (red) without domain-specific exclusionary logic.

SDG 17: Partnerships for the Goals. Sustainable conservation depends on integrating computer science with ecology. We establish a unified architecture that connects AI researchers, field biologists, and conservation agencies to facilitate the transfer of technology into ecological practice. This research advances SDG 17.16 (“Enhance the Global Partnership”) and SDG 17.17 (“Encourage multi-stakeholder collaborations”) by mobilizing expertise across disciplines.

Alignment with LNOB. A “digital divide” currently restricts the application of advanced AI in remote conservation areas. This research aligns with LNOB by democratizing and preserving expert knowledge. The system formalizes tacit expertise into updateable knowledge graphs. This process safeguards critical scientific insights that are at risk of being lost. Ultimately, the project empowers local communities with professional-grade monitoring tools,

3 Strategies

To overcome the operational and knowledge barriers identified above, we propose **Knowledge-Adaptive Edge Expert Agents (KADEX)** as a strategic pivot from **model adaptation** to **knowledge adaptation**. Today’s wild on-site counting pipelines are still dominated by pure vision detectors (e.g., YOLO [Wang *et al.*, 2024] and RT-DETR [Zhao *et al.*, 2024]), while recent attempts to replace them with VLMs remain constrained by the same two field realities: limited edge compute only sustains small-parameter models, and end-to-end VLM retraining/fine-tuning amplifies data, energy, and bandwidth costs. KADEX is motivated by a key observation: for small parameter VLMs, the dominant bottleneck is often missing expert logic, and this logic does not need to be permanently embedded in model weights. Instead, it can be externalized, updated, and retrieved on demand as a lightweight knowledge layer.

3.1 Knowledge Adaptation to Compensate VLMs

Figure 3 illustrates the opportunity behind KADEX. Even when a small parameter VLM (Qwen3-VL-8B [Bai *et al.*, 2025]) fails at fine-grained species identification, its visual encoder can still extract discriminative field features (e.g.,

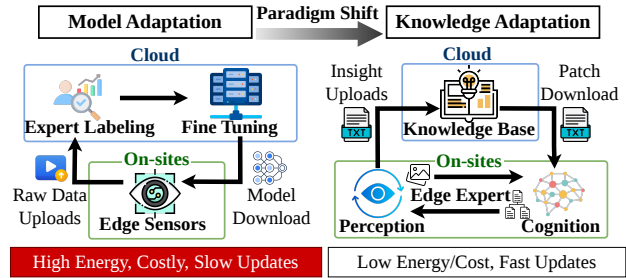


Figure 4: Strategic paradigm shift: Instead of retraining heavy models in the cloud (high energy/cost), KADEX updates a lightweight external knowledge base on the edge (low energy/cost).

spots, body color, shape). The failure arises when these cues are mapped to labels without the exclusionary and contextual rules that experts apply (e.g., “Sockeye do not have back/tail spots”). KADEX decouples expert logic from model parameters to avoid expensive retraining. It achieves this by using retrieval-augmented generation (RAG) [Lewis *et al.*, 2020; Li *et al.*, 2024; Yu *et al.*, 2024] to inject local rules and constraints into a lightweight knowledge base. This enables rapid, low-bandwidth iteration of domain knowledge while keeping on-device inference feasible.

3.2 Decoupled Perception and a Transferable Knowledge Layer

Operationally, KADEX implements this pivot by decoupling perception from cognition. A stable, lightweight perception module (e.g. CLIP [Radford *et al.*, 2021]) extracts visual evidence, while a dynamic retrieval-based knowledge base provides evolving expert constraints. Figure 4 shows that adaptation is driven by updating the external knowledge layer, rather than repeatedly retraining and redeploying model weights. This design keeps iteration feasible under sparse connectivity and tight energy budgets.

More broadly, KADEX targets a recurring challenge in ecological recognition: visual cues alone are often insufficient for taxonomic decisions, even with strong encoders. Figure 5 illustrates this “lookalike” regime, where visually similar entities require scientifically precise distinctions. KADEX couples visual evidence with explicit, updateable constraints and spatiotemporal priors, enabling transfer beyond salmon to other biodiversity and environmental sensing tasks where domain logic is essential.

Two simple cases illustrate this logic beyond salmon. For visual ambiguity, *Chromis viridis* and *C. atripectoralis* are near-identical blue-green reef fishes in the Indo-Pacific. Experts separate them by the black pectoral-axil mark. KADEX represents this as an attribute-level rule: an activated “black pectoral-axil spot” node supports *C. atripectoralis* and conflicts with *C. viridis*. For spatial and part-based ambiguity, consider a camera trap near Vancouver, BC, Canada. A partially occluded small felid with a short tail may match both *Lynx rufus* and *Lynx canadensis* from visual evidence alone. This case is useful because *L. canadensis* is rarely expected in the Lower Mainland, while *L. rufus* is locally plausible in southern coastal B.C. KADEX activates context nodes and locally relevant species communities in the L-SKG. This can

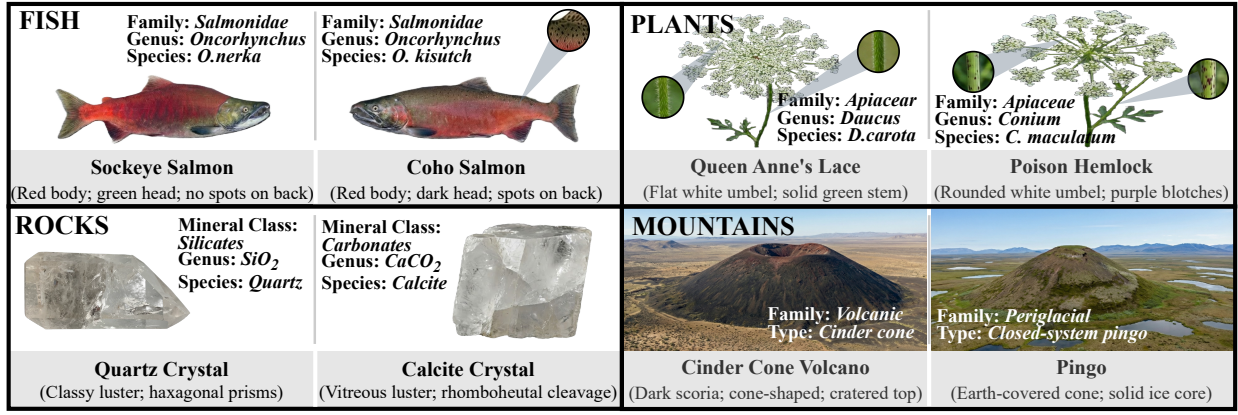


Figure 5: Visual ambiguity vs. taxonomic precision: Distinguishing visually similar but scientifically distinct entities extends beyond salmon to plants, minerals, and landforms.

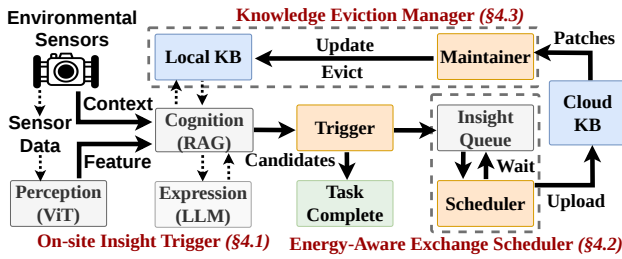


Figure 6: Overview of KADEX.

favor *Lynx rufus* and produce an auditable explanation. In both cases, expert notes become reusable structured patches rather than new model weights.

4 Methods

To resolve the operational conflict between strict resource limits and the need for continuous adaptation, we propose KADEX. This architecture shifts the adaptation paradigm from heavy model retraining to lightweight knowledge management. KADEX separates perception from cognition. It employs a fixed vision encoder to extract features and a Local Structured Knowledge Graph (L-SKG) to interpret them via compact updates. However, this design introduces three specific implementation challenges in off-grid scenarios: (i) identifying ambiguous observations that require expert review; (ii) scheduling data exchange under fluctuating energy budgets; and (iii) maintaining a compact knowledge base within finite storage.

As illustrated in Figure 6 and Algorithm 1, KADEX addresses these through three synchronized modules. The On-site Insight Trigger (§4.1) functions as a cognitive gatekeeper. It utilizes structural entropy to filter routine data and queue only high-value insights. The Energy-Aware Exchange Scheduler (§4.2) optimizes satellite uplink usage. This module dynamically weighs informational urgency against real-time battery states. Finally, the Knowledge Eviction Manager (§4.3) enforces storage constraints. It applies a community-level policy to retire outdated semantic regions, ensuring the local knowledge base remains relevant and coherent.

Algorithm 1 KADEX High-Level Operational Lifecycle

Require: Video stream $\{I_t\}$, L-SKG $\mathcal{G}_{\text{local}}$, Battery state $B(t)$
Ensure: Continuously adapted knowledge $\mathcal{G}_{\text{local}}$ and insight uploads

```

1:  $Q(t) \leftarrow \emptyset$ ; SystemActive  $\leftarrow$  true;
2: while SystemActive do
3:    $I_t \leftarrow$  CaptureFrame();  $B(t) \leftarrow$  UpdateBatteryState();
4:   // Phase 1: On-site Insight Trigger (Cognitive Gatekeeper)
5:    $G_{\text{inf}} \leftarrow$  PerceiveAndReason( $I_t, \mathcal{G}_{\text{local}}$ );
6:   if IsHighEntropy( $G_{\text{inf}}, \tau_{\text{trigger}}$ ) then
7:     // Capture ambiguous frames for expert review
8:      $Q(t).enqueue(\text{PackInsight}(I_t))$ ;
9:   else
10:    ExecuteLocalTask( $G_{\text{inf}}$ ); // Routine inference
11:   end if
12:   // Phase 2: Energy-Aware Exchange Scheduler
13:    $E_{\text{budget}} \leftarrow$  ComputeSafeBudget( $B(t), B_{\text{safe}}$ );
14:   if  $E_{\text{budget}} > 0$  then
15:     Sort  $Q(t)$  by priority (Entropy descending);
16:     TransmitInsights(Uplink,  $Q(t), E_{\text{budget}}$ );
17:   end if
18:   // Phase 3: Community-Level Knowledge Manager
19:   if received patches  $\{\text{Patch}(c)\}$  from Cloud then
20:      $\mathcal{G}_{\text{local}} \leftarrow$  IntegratePatches( $\mathcal{G}_{\text{local}}, \{\text{Patch}(c)\}$ );
21:     // Maintain structural coherence under storage limits
22:     EnforceEvictionPolicy( $\mathcal{G}_{\text{local}}, \text{Cap}$ );
23:   end if
24:    $t \leftarrow t + 1$ ;
25: end while

```

4.1 On-site Insight Trigger

The On-site Insight Trigger determines whether a new frame reflects routine conditions or reveals patterns that the Local Structured Knowledge Graph (L-SKG) cannot explain. Thresholds based only on visual similarity are unreliable because they ignore ecological structure and explicit logical constraints. To address this limitation, we measure the semantic fit between the observation and the L-SKG using Structural Graph Entropy, which reflects whether the observation supports a single coherent interpretation or forces the system into uncertainty or contradiction.

The L-SKG is a compact graph stored on the edge device, defined as $G = (V, E)$. The node set V contains en-

tity nodes V_{ent} , attribute nodes V_{att} , and context nodes V_{ctx} . Entity nodes represent species such as chinook or sockeye. Attribute nodes encode observable morphological traits such as dorsal spots or a hooked lower jaw. Context nodes describe site and environmental factors such as site identity, GPS region, water temperature, seasonal stage, or diel lighting. The edge set contains only two semantic types: (1) supporting edges, which link attributes to entities with positive weights ($w(a, e) > 0$), and (2) conflict edges, which specify mutually exclusive relationships grounded in expert knowledge.

When the perception module (e.g. CLIP [Radford *et al.*, 2021]) processes an image I , it outputs feature embeddings $V_f = \{f_1, \dots, f_m\}$. Each feature activates a subset of attribute nodes, yielding the set A_{feat} . Sensor readings activate a set of context nodes A_{ctx} . The combined activation set is $A = A_{feat} \cup A_{ctx}$. From this set, the trigger constructs an **inference subgraph** G_i that focuses only on knowledge relevant to the current frame. It first identifies the neighboring entity nodes $V_{ent}^* = N(A) \cap V_{ent}$. The node set and edge set of the inference subgraph are defined as follows:

$$G_i = (V_i, E_i), V_i = A \cup V_{ent}^*, E_i = \{(u, v) \in E : u, v \in V_i\}.$$

This construction isolates the specific portion of the knowledge graph needed to interpret the current observation and excludes species or contextual knowledge unrelated to the active attributes and conditions.

Within this subgraph, the trigger computes a **semantic support score** for each candidate entity. It first checks for explicit contradictions: if any activated attribute $a \in A_{feat}$ connects to an entity e through a conflict edge, then e is incompatible with the observation and is assigned $S(e) = 0$. For all remaining entities, the support score aggregates evidence from supporting edges:

$$S(e) = \sum_{a \in A_{feat}} w(a, e). \quad (1)$$

Attributes without supporting edges contribute zero. This score increases when the observed traits match the species' typical descriptors in the L-SKG and drops to zero when a hard conflict is detected.

Only entities with positive support form the valid set $E_{valid} = \{e \in V_{ent}^* \mid S(e) > 0\}$. The trigger then normalizes support scores into a probability distribution:

$$P(e \mid V_f, C_x) = \frac{\exp(S(e))}{\sum_{e' \in E_{valid}} \exp(S(e'))}, \quad (2)$$

and assigns zero probability to all excluded entities. This distribution captures both the strength of positive evidence and the effect of explicit constraints that remove otherwise plausible classes.

The **Structural Graph Entropy** is the Shannon entropy of this distribution:

$$H(G_i) = - \sum_{e \in E_{valid}} P(e \mid V_f, C_x) \log P(e \mid V_f, C_x). \quad (3)$$

Entropy is low when one entity dominates, indicating that the L-SKG provides a stable, coherent explanation. Entropy rises when several explanations compete or when conflict

edges eliminate what would otherwise be the leading entity. Thus, the entropy reflects both ambiguity and structural tension without requiring a separate conflict penalty.

The trigger compares the entropy to a threshold $\tau_{trigger}$. Frames with entropy below the threshold are routine and update only local statistics. Frames with entropy above the threshold are marked as high-value because they indicate missing knowledge or a violation of existing structure.

4.2 Energy-Aware Exchange Scheduler

The existing system relied on fixed rules for uplink control, i.e., preset 2-hour upload windows per day. These rules fail when the battery level and the rate of insight generation fluctuate. During long low-light periods, a fixed schedule may continue transmitting and drain the battery below a safe margin. Conversely, during bursts of high-entropy insights, a fixed cap cannot prioritize the most critical accumulated information. To address this, we model uplink scheduling as an energy-aware optimization problem and derive a closed-loop suitable for edge deployment.

As described in Sec 4.1, the On-site Insight Trigger filters raw frames and generates an Insight Packet only when the structural graph entropy $H(G_{inf})$ is high. At each time slot t , the device maintains a queue $\mathcal{Q}(t)$ of such packets. Unlike simple FIFO buffers, the scheduler must prioritize observations that offer the highest information gain. Therefore, we sort $\mathcal{Q}(t)$ by entropy $H(p)$ in descending order. Each packet $p \in \mathcal{Q}(t)$ stores its entropy, a fused feature vector z_p , and a timestamp. With an estimate of the energy per bit, the cost of uploading a packet is approximated by a constant e_{pkt} .

Let $B(t)$ denote the battery state and B_{safe} a safety threshold that preserves energy for essential sensing and local inference. The scheduler assigns a dynamic communication budget:

$$E_{budget}(t) = \max\{0, B(t) - B_{safe}\}, \quad (4)$$

which grows when the battery is healthy and shrinks to zero during extended low-light periods, automatically pausing uplink activity.

Given this budget, the scheduler solves a per-slot optimization problem. Let the sorted insight queue at time t be

$$\mathcal{Q}(t) = \{p_1(t), p_2(t), \dots, p_{n(t)}(t)\},$$

where $H(p_1) \geq H(p_2) \geq \dots \geq H(p_{n(t)})$. The decision variable $k(t) \in \{0, 1, \dots, n(t)\}$ specifies how many packets to upload from the head of the priority queue. The objective is to transmit the maximum total entropy allowed by the current energy constraint:

$$\max_{k(t)} \sum_{j=1}^{k(t)} H(p_j(t)) \quad \text{s.t.} \quad k(t) e_{pkt} \leq E_{budget}(t). \quad (5)$$

Since the packets are already sorted by value (entropy), the optimal strategy is simply to transmit as many top-priority packets as the budget allows:

$$k^*(t) = \min\left(n(t), \left\lfloor \frac{E_{budget}(t)}{e_{pkt}} \right\rfloor\right). \quad (6)$$

This prioritization keeps routine monitoring active under

low energy. It also moves the most useful novel cases to the front of the upload queue when communication is available.

Once uploaded, high-entropy packets drive the expert-in-the-loop workflow. This is not full stream annotation. Experts review only cases that remain unclear after cloud retrieval. If the cloud master graph can explain the packet, it returns a knowledge patch to the edge. Otherwise, an expert gives a short text or audio note. A knowledge graph-based RAG pipeline [Edge *et al.*, 2024] converts the note into structured nodes and edges. The cloud master graph is updated first. The edge receives only the relevant patch.

4.3 Knowledge Eviction Manager

After uplink transmission, the cloud processes the uploaded Insight Packets and applies them to the cloud master graph $\mathcal{G}_{\text{global}}$. Based on this cloud view, the cloud generates device-specific knowledge patches. These patches are naturally organized by semantic communities. A community may correspond to a species (e.g., *chinook*), a species and season combination (e.g., *chinook-spawning*), a species and sex category, or even finer-grained variants (e.g., *jack chinook*). Each patch contains the subgraph of that community together with its local updates, including new nodes, edges, and associated weights. When delivered to the device, these patches incrementally update the L-SKG.

Conventional LRU policies operate at the level of individual items, which here correspond to individual nodes or edges. Such fine-grained eviction is unsuitable for our inference procedure in Section 3.2. The inference subgraph G_{inf} is constructed by expanding around a set of activated nodes and collecting their neighbors. Evicting isolated nodes or edges breaks the neighborhood structure inside a semantic community, leaving incomplete local contexts and inconsistent constraint sets. To avoid this fragmentation, we use a *community-level* LRU policy. All loading, updating, and eviction occur at the granularity of semantic communities, preserving the structural coherence needed for local inference.

The device stores its L-SKG as a union of community subgraphs. Let $\mathcal{C}_{\text{local}}$ be the current set of cached communities, and let each $c \in \mathcal{C}_{\text{local}}$ have an associated subgraph $\mathcal{G}_c^{\text{local}}$ and a timestamp $\text{LastUsed}(c)$. This timestamp is refreshed whenever the community is accessed during inference or updated by an incoming patch. The local storage is bounded by a capacity Cap , measured in nodes, edges, or total bytes. When new patches arrive, they may cause the L-SKG to exceed this capacity. In that case, the system performs LRU eviction at the community level: communities with the earliest LastUsed values are removed entirely until the capacity constraint is satisfied.

This community-level LRU strategy ensures that the device maintains coherent semantic regions of the L-SKG. By evicting entire communities rather than isolated items, the method avoids breaking neighborhood structure and keeps the local graph aligned with the inference process in Sec 4.1. Under limited storage, the device automatically retains the communities that have been most relevant to recent observations, providing a compact and stable prior for subsequent entropy-based inference.

5 Case Study and Implementation Plan

Our implementation plan follows a staged pathway that couples system hardening with knowledge curation, moving from an expert-built cloud knowledge base to an expert-in-the-loop field workflow, and finally to expert-bounded autonomous operation. We start from a practical observation in this domain: the limiting factor for small on-device models is often not feature extraction, but the absence of explicit expert logic. KADEX therefore treats knowledge as a first-class artifact that is constructed, audited, and updated through lightweight patches, while keeping on-site computation and communication bounded.

Phase 1: Cloud knowledge base bootstrapping from expert narratives. We first bootstrap a master knowledge base in the cloud from detailed species explanations provided by domain experts. These materials are recorded as structured audio narratives. They cover diagnostic traits, exclusionary rules, life-stage transitions, and site/season-specific priors. We then transcribe them into text and convert them into a structured knowledge graph schema. This process turns rich expertise into explicit entities and constraints. Examples include trait-species relations, negative rules, and spatiotemporal applicability. The result is an auditable starting point that does not depend on data-intensive retraining.

Phase 2: Expert-in-the-loop deployment and iterative knowledge patching. We then deploy KADEX at partner-designated sites under real constraints. These include duty-cycled satellite connectivity and solar/battery energy budgets. During this stage, KADEX runs with an expert-in-the-loop workflow. The **On-site Insight Trigger** surfaces only high-uncertainty or novel cases as compact insight packets. Local domain experts review these packets and correct the reasoning when needed. They submit updates as text or short audio notes. These corrections become knowledge patches. The patches update the cloud master graph and are selectively pushed back to edge devices. This allows the fleet to adapt without uploading raw video or retraining model weights. Experts review only unresolved high-entropy packets, not the full stream. One active site may need tens to low hundreds of expert hours in early Phase 2, then periodic auditing.

Phase 3: Expert-bounded autonomy with preserved digital heritage. Once the agent’s behavior stabilizes, KADEX can operate autonomously for low-entropy routine observations. This autonomy remains bounded by expert authority. It conserves scarce expert time for high-entropy observations, unknown species or life stages, site-specific ecological change, sensor failures, and management-relevant decisions. The knowledge base then becomes both durable *digital heritage* and an interface for expert validation. Expert logic is preserved in an explicit form. New patches still require expert validation, revision, or abstention when the system lacks a reliable explanation.

Case study and resources. We will use existing multi-site historical data to replay streams and test the full KADEX life-cycle, then transition to the partner-provided field pilot for iterative patching and hardening. The project already has core Phase 1 infrastructure, including edge compute nodes

(Jetson-class devices¹), solar harvesting and battery storage, and historical observations. The main Phase 2 requirement is sustained access to domain experts who validate surfaced in-sight packets and issue structured knowledge updates through text or short audio notes.

6 Evaluation Criteria

To ensure the proposed solution meets the rigorous demands of scientific monitoring and off-grid operation, we establish a multi-dimensional evaluation framework. This protocol encompasses ecological precision, operational sustainability, and communication efficiency, grounded in the constraints identified during our preliminary pilot analysis.

6.1 Taxonomic Precision and Reliability

The primary success metric is the system’s ability to replicate expert-level classification in the wild. We evaluate performance using F1-score [Goodfellow, 2016] and mean Average Precision (mAP) on expert-annotated longitudinal datasets. To validate the efficacy of our approach, we benchmark against three categories of baselines. First, we compare results with standard real-time object detectors, specifically YOLO [Kay *et al.*, 2022] and RT-DETR [Zhao *et al.*, 2024]. Second, we evaluate performance against the domain-specific STSVT system [Xu *et al.*, 2024]. Third, we assess our method against advanced multimodal foundation models, including VLM (e.g. Qwen3-VL [Bai *et al.*, 2025]) and recent fusion approaches [Xu *et al.*, 2025]. These comparisons focus on fine-grained classification accuracy for visually similar species and the system’s ability to reject non-target “confuser” species. Initial experiments suggest that integrating expert logic helps mitigate visual hallucinations common in VLMs. This approach shows promise for distinguishing visually similar species.

6.2 Survival Under Intermittent Energy

For off-grid viability, we evaluate system resilience using loss-of-load probability (LOLP) [Allan and others, 2013]. This standard metric quantifies the likelihood of the Battery State of Charge (SOC) dropping below the critical safety threshold (30%) over a long-term deployment. We compare our adaptive policy against the fixed duty-cycle baseline [Xu *et al.*, 2024], which enforces a static two-hour daily transmission window. The evaluation measures the reduction in LOLP under constrained energy profiles, such as prolonged overcast periods. Our target is to achieve an LOLP of zero (indicating continuous availability) in scenarios where the baseline experiences forced shutdowns. Initial simulations using historical solar irradiance data indicate that our energy-aware scheduler successfully maintains the battery state of charge above the safety margin during extended low-light intervals.

7 Expected Results and Impacts

This project aims to make AI monitoring practical in remote ecosystems, while preserving expert knowledge in a form that can be reused and updated over time.

¹<https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/>

Expected technical outcomes. KADEX replaces repeated cloud retraining with lightweight knowledge updates. We expect this to reduce energy use and data transfer compared to cloud-centric model adaptation, which helps maintain stable operation under solar and battery limits. We also expect more reliable and accurate fine-grained identification in the field, as each decision is constrained by an explicit expert knowledge graph rather than being hidden in model weights. Evaluation will follow a staged process: offline replay on multi-site historical data, then expert-in-the-loop validation in a partner-designated pilot.

Expected societal and scientific impacts. KADEX promotes knowledge sustainability by converting tacit, local expertise into explicit and reusable rules and graph structures. This reduces knowledge loss when experts retire or teams change. KADEX also makes system updates easier for practitioners. Instead of retraining models, they can improve the system by adding or revising rules, priors, and context descriptions. Finally, KADEX lowers the barrier to long-term monitoring in under-resourced settings by reducing dependence on continuous connectivity and large data uploads. The same approach can extend beyond salmon to other biodiversity tasks where lookalike species and changing conditions make vision-only models unreliable. Because expert time is scarce, KADEX shifts effort from full-stream annotation to graph validation, anomaly review, and management decisions. As physical AI matures [NVIDIA, 2026], robots may collect richer field evidence. Experts will still be needed for knowledge-base design, sampling, site context, and cases beyond sensor and physical AI coverage.

8 Practical Challenges

Deploying knowledge-adaptive agents in infrastructure-constrained wild environments remains uncertain. In particular, the system’s initial reasoning depends on bootstrapped expert knowledge, and “unknown” species (unknown unknowns) challenge closed-world graphs. KADEX therefore treats high-entropy events as review triggers. If a case is unknown only to the L-SKG, the cloud master graph sends a knowledge patch back to the edge. If the case is also unknown to the cloud master graph, experts review a compact packet and update the graph with text or audio notes. If experts cannot assign a reliable species label, the system stores the case as an anomaly with time, location, and visual evidence. It does not force a classification. During long outages, low-entropy cases still run locally, while high-entropy packets wait until energy and connectivity allow upload. This workflow lets experts patch the graph over time and turns field intuition into reusable digital heritage.

In addition, hardware constraints impose physical trade-offs. Finite edge storage requires eviction of inactive knowledge communities, which may remove local context for rare or seasonal species. The cloud master graph limits this risk. It grows only when verified and reusable patches are available. Unresolved anomalies are not added as low-confidence species nodes, which helps limit graph noise. When an evicted species reappears, the edge uploads a high-entropy packet and receives the relevant cloud subgraph. Before patching, the system may abstain or produce a false negative.

Ethics Statement

This research follows a “Responsibility-by-Design” framework that prioritizes ethical standards, local rights, and Indigenous Data Sovereignty. Local partners retain ownership of raw data and control how sensitive information is shared, which helps prevent exploitation and preserve local authority. For privacy, KADEX processes data on the edge, filters images of human activities locally, and transmits only ecological data. For accountability, KADEX uses an explicit knowledge graph rather than an opaque “black-box” model, allowing experts to audit decision logic and correct biased or unsafe patches. Human oversight is treated as meaningful authority, not symbolic approval. Local experts and Indigenous partners must be able to inspect evidence paths, reject or freeze unsafe knowledge patches, and define which decisions require human or community review.

Acknowledgments

This work was supported in part by an NSERC Discovery Grant, a British Columbia Salmon Recovery and Innovation Fund (BCSRIF 2022 401), a Mitacs Accelerate Cluster Grant, and support from experiment.com. We thank the Heiltsuk, Haida, Kitasoo Xai'xais, Taku River Tlingit, and Gitga'at First Nations for their trust, guidance, and collaboration. We also thank the Skeena Fishery Commission and the Gitanyow Fisheries Authority for their partnership in field monitoring and salmon stewardship. The corresponding author for this work is Jiangchuan Liu.

References

- [Allan and others, 2013] Ronald N Allan et al. *Reliability evaluation of power systems*. Springer Science & Business Media, 2013.
- [Atlas et al., 2021] William I Atlas, Natalie C Ban, Jonathan W Moore, Adrian M Tuohy, Spencer Greening, Andrea J Reid, Nicole Morven, Elroy White, William G Housty, Jess A Housty, et al. Indigenous systems of management for culturally and ecologically resilient pacific salmon (*oncorhynchus* spp.) fisheries. *BioScience*, 2021.
- [Atlas et al., 2023] William I Atlas, Sami Ma, Yi Ching Chou, Katrina Connors, Daniel Scurfield, Brandon Nam, Xiaoqiang Ma, Mark Cleveland, Janvier Doire, Jonathan W Moore, et al. Wild salmon enumeration and monitoring using deep learning empowered detection and tracking. *Frontiers in Marine Science*, 2023.
- [Axford et al., 2024] Daniel Axford, Ferdous Sohel, Mathew A Vanderklift, and Amanda J Hodgson. Collectively advancing deep learning for animal detection in drone imagery: Successes, challenges, and research gaps. *Ecological Informatics*, 2024.
- [Bai et al., 2025] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Ke-qin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- [Di Lorenzo and Mantua, 2016] Emanuele Di Lorenzo and Nathan Mantua. Multi-year persistence of the 2014/15 north pacific marine heatwave. *Nature Climate Change*, 2016.
- [Díaz Musmanni, 2023] Gabriela Díaz Musmanni. Protecting wildlife in a changing climate: Four powerful adaptation strategies. <https://gca.org/protecting-wildlife-in-a-changing-climate-four-powerful-adaptation-strategies/>, 2023.
- [Dorner et al., 2018] Brigitte Dorner, Matthew J Catalano, and Randall M Peterman. Spatial and temporal patterns of covariation in productivity of chinook salmon populations of the northeastern pacific ocean. *Canadian Journal of Fisheries and Aquatic Sciences*, 2018.
- [Edge et al., 2024] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.
- [Folkman et al., 2025] Lukas Folkman, Kylie A Pitt, and Bela Stantic. A data-centric framework for combating domain shift in underwater object detection with image enhancement. *Applied Intelligence*, 2025.
- [Frölicher and Laufkötter, 2018] Thomas L Frölicher and Charlotte Laufkötter. Emerging risks from marine heat waves. *Nature communications*, 2018.
- [Goodfellow, 2016] Ian Goodfellow. *Deep learning*. MIT press, 2016.
- [Gordon et al., 2023] Lucia Gordon, Nikhil Behari, Samuel Collier, Elizabeth Bondi-Kelly, Jackson A. Killian, Catherine Ressijac, Peter Boucher, Andrew Davies, and Milind Tambe. Find rhinos without finding rhinos: Active learning with multimodal imagery of south african rhino habitats. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.
- [Hill, 2025] Kelly Hill. Starlink speeds rise, but still fall (mostly) short of fcc standards. *RCR Wireless News*, 2025.
- [Kay et al., 2022] Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The caltech fish counting dataset: a benchmark for multiple-object tracking and counting. In *European Conference on Computer Vision*, 2022.
- [Kay et al., 2024] Justin Kay, Timm Haucke, Suzanne Stathatos, Siqi Deng, Erik Young, Pietro Perona, Sara Beery, and Grant Van Horn. Align and distill: Unifying and improving domain adaptive object detection. *arXiv preprint arXiv:2403.12029*, 2024.
- [Kilduff et al., 2015] D Patrick Kilduff, Emanuele Di Lorenzo, Louis W Botsford, and Steven LH Teo. Changing central pacific el niños reduce stability of north american salmon survival rates. *Proceedings of the National Academy of Sciences*, 2015.

- [Kshitiz *et al.*, 2023] Kshitiz, Sonu Shreshtha, Ramy Mounir, Mayank Vatsa, Richa Singh, Saket Anand, Sudeep Sarkar, and Sevaram Mali Parihar. Long-term monitoring of bird flocks in the wild. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.
- [Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- [Li *et al.*, 2024] Zeyu Li, Qipeng Guo, Yunzhi Yao, et al. Structrag: Boosting knowledge-intensive reasoning of large language models via inference-time hybrid information structurization. In *International Conference on Learning Representations (ICLR)*, 2024.
- [Liu *et al.*, 2024] Weizhen Liu, Jiayu Tan, Guangyu Lan, Ao Li, Dongye Li, Le Zhao, Xiaohui Yuan, and Nanqing Dong. Benchmarking fish dataset and evaluation metric in keypoint detection - towards precise fish morphological assessment in aquaculture breeding. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.
- [Ma *et al.*, 2024] Sami Ma, Yi Ching Chou, Miao Zhang, Hao Fang, Haoyuan Zhao, Jiangchuan Liu, and William I Atlas. Leo satellite network access in the wild: Potentials, experiences, and challenges. *IEEE Network*, 2024.
- [Nature Tech Collective, 2025] Nature Tech Collective. Exploring ground-truth nature tech & the future of biodiversity monitoring. <https://www.naturetechcollective.org/stories/ground-truth-biodiversity-wildlife-monitoring>, 2025.
- [Norouzzadeh *et al.*, 2018] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences (PNAS)*, 2018.
- [NVIDIA, 2026] NVIDIA. Nvidia cosmos: An open platform for physical ai with world foundation models. <https://www.nvidia.com/en-us/ai/cosmos/>, 2026.
- [Paul *et al.*, 2024] Dipanjyoti Paul, Arpita Chowdhury, Xinqi Xiong, Feng-Ju Chang, David Edward Carlyn, Samuel Stevens, Kaiya Provost, Anuj Karpatne, Bryan Carstens, Daniel I Rubenstein, et al. A simple interpretable transformer for fine-grained image classification and analysis. In *International Conference on Learning Representations (ICLR)*, 2024.
- [Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021.
- [Schindler and Hilborn, 2015] Daniel E Schindler and Ray Hilborn. Prediction, precaution, and policy under global change. *Science*, 2015.
- [United Nations, 2015] United Nations. Transforming our world: The 2030 agenda for sustainable development. <https://sdgs.un.org/2030agenda>, 2015.
- [United Nations, 2021] United Nations. Leave no one behind. <https://unsdg.un.org/2030-agenda/universal-values/leave-no-one-behind>, 2021.
- [Wang *et al.*, 2024] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024.
- [Waples *et al.*, 2008] Robin S Waples, George R Pess, and Tim Beechie. Evolutionary history of pacific salmon in dynamic environments. *Evolutionary Applications*, 2008.
- [World Wide Fund for Nature, 2024] World Wide Fund for Nature. Living planet report 2024: A system in peril. Report, WWF, 2024.
- [Wu *et al.*, 2022] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 2022.
- [Xu *et al.*, 2024] Chi Xu, Rongsheng Qian, Hao Fang, Xiaoqiang Ma, William I. Atlas, Jiangchuan Liu, and Mark A. Spoljaric. Salina: Towards sustainable live sonar analytics in wild ecosystems. In *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2024.
- [Xu *et al.*, 2025] Chi Xu, Yili Jin, Sami Ma, Rongsheng Qian, Hao Fang, Jiangchuan Liu, Xue Liu, Edith CH Ngai, William I Atlas, Katrina M Connors, et al. Exploring multimodal foundation ai and expert-in-the-loop for sustainable management of wild salmon fisheries in indigenous rivers. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.
- [Xu *et al.*, 2026] Chi Xu, Jiaxing Li, Mengdi Jin, William I. Atlas, Mark A. Spoljaric, Edith C. H. Ngai, and Jiangchuan Liu. Fused: Toward federated multimodal retrieval across sovereign data domains. In *The ACM Web Conference (WWW)*, 2026.
- [Yang *et al.*, 2024] Zhen Yang, Ming Ding, Tinglin Huang, Yukuo Cen, Junshuai Song, Bin Xu, Yuxiao Dong, and Jie Tang. Does negative sampling matter? a review with insights into its theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024.
- [Yu *et al.*, 2024] Shi Yu, Chaoyue Tang, Bokai Xu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.
- [Zhao *et al.*, 2024] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.