Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3149

AN AI-DRIVEN SMART MONITORING SYSTEM FOR ECOLOGICAL FLOW IN SMALL HYDROPOWER STATIONS

HuangXin Deng
Department of Engineering, University of Cambridge, Cambridge, United Kingdom.

Abstract: Ecological flow refers to the minimum water release required to sustain downstream ecosystems. Monitoring compliance at small hydropower stations is usually challenging because outlets may vary in design, live cameras are often misaligned, or monitoring reports can be incomplete or falsified. This paper introduces a novel AI-driven monitoring system that overcomes these challenges by leveraging a multimodal large model, carefully adapted through parameter-efficient fine-tuning. Our approach not only ensures more reliable ecological flow compliance monitoring but also demonstrates how multimodal models can be applied to improve real-world environmental management. The proposed system addresses three critical questions: 1. Are cameras correctly installed? 2. Is water flowing? and 3. Does the release comply with regulatory standards? Despite being trained on a limited labelled dataset, the model achieves over 90% precision across diverse outlet types with narrow confidence intervals. The system is now deployed at nearly 5,000 stations, minimising dependence on manual inspections while enabling continuous, transparent, and scalable monitoring of ecological flow.
Keywords: Ecological flow; Hydropower monitoring; Computer vision; Multimodal large models

1 INTRODUCTION

Small hydropower stations are vital for renewable energy production, but can adversely affect river ecosystems if ecological flow standards are not maintained. Ecological flow, defined as the water volume required to sustain aquatic life and river health, is critical for preserving biodiversity and ecosystem services [1]. Conventional monitoring is still widely used because field inspections and self-reported discharge records provide a straightforward and low-cost way for regulators to collect compliance data [2]. However, these methods are limited: inspections are labour-intensive and intermittent, while self-reports are vulnerable to misreporting or falsification. Sensor-based hydrological estimation can improve continuity, but requires costly deployment and maintenance across thousands of distributed sites. Recent surveys point out these systemic challenges and emphasise the need for scalable, evidence-based monitoring solutions [3]. These limitations motivate AI-driven methods for automated and scalable outlet monitoring.

Although AI has gained significant attention and popularity, existing approaches remain challenging to implement in real-world scenarios for ecological flow management. For example, many methods rely on lightweight vision models or handcrafted rules that presume stable video feeds, carefully defined regions of interest, and controlled lighting [4-7]. In practice, however, hydropower stations exhibit considerable variation [1, 8]. Outlet geometry and installation settings depend on local conditions, network bandwidth is often constrained, and illumination, as well as seasonal factors, fluctuate widely. Consequently, monitoring pipelines that rely heavily on ROI design [9], motion tracking, or optical flow tend to underperform in practice and impose substantial maintenance burdens [6, 7].

In Guangdong Province, China, nearly 10,000 small hydropower stations are in operation, with discharge facilities ranging from gates and pipes to open channels and weirs [10]. These outlets exhibit considerable structural and visual diversity, with few consistent features, which creates significant challenges for automated visual monitoring. This operational context motivates the following research question:

RQ1: How can we learn a robust recogniser that generalises across highly diverse outlet structures and scenes where traditional lightweight models fail to transfer?

Beyond merely detecting the presence of water, compliance further depends on whether the water is actually moving. However, visual cues of flow vary considerably with ground surface, turbulence, and proximity to rivers. Given that bandwidth limits often preclude video transmission, existing systems usually rely on motion-based techniques, such as carefully tuned regions of interest or optical flow. These methods, however, are brittle and demand substantial manual configuration. This motivates a second research question:

RQ2: How can we infer active flow from single images, without relying on video streams or handcrafted motion heuristics, in order to reduce engineering costs and improve scalability?

Absolute discharge cannot be reliably measured from still images; instead, monitoring practices rely on visual evidence against reference conditions. Traditional lightweight models attempt to approximate this by defining sensitive zones or employing heuristic flow-speed proxies, but these methods are brittle and prone to errors [4, 6, 7]. Prior ecological monitoring studies have relied on handcrafted features or small convolutional neural network (CNN) classifiers, which fail to generalise across diverse outlets and environments [1, 8]. This motivates a third research question:

RQ3: In the absence of reliable per-frame metering, how can compliance be evaluated from visual evidence alone, without unstable heuristics or rigid sensitive-zone design?

Figure 1 summarises these research questions alongside their task definitions, input–output pairs, and corresponding metrics.

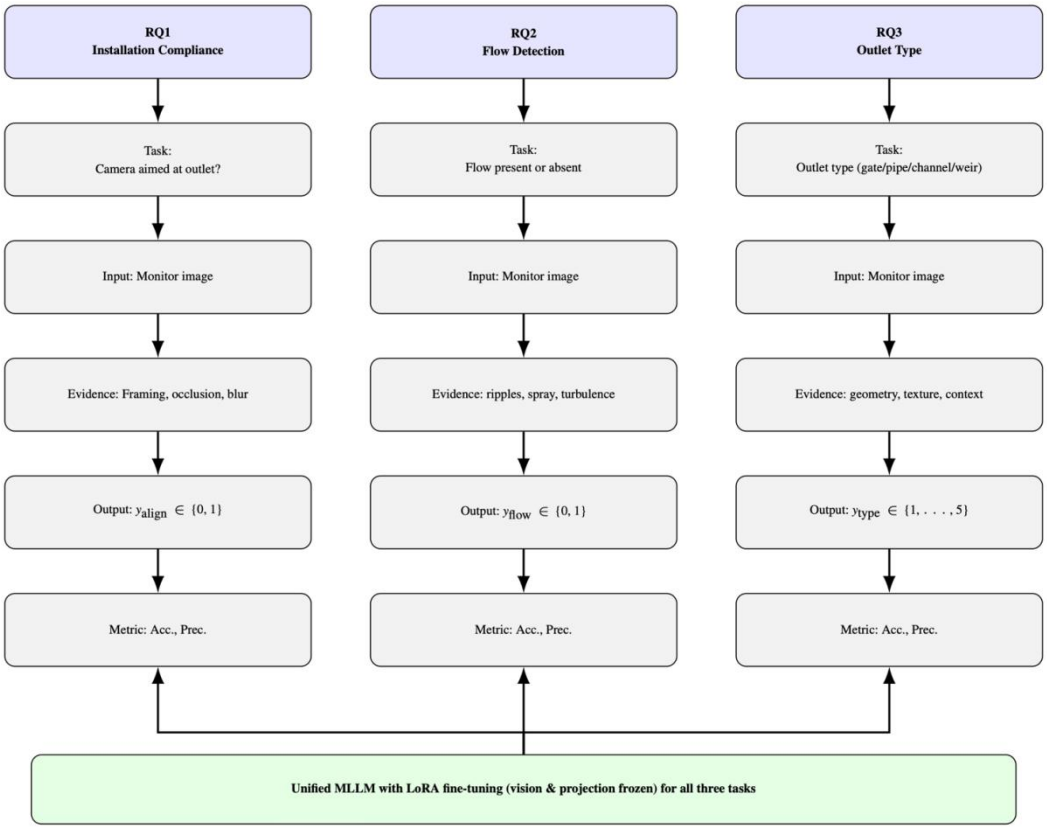

Figure 1 Summary of the Three Research Questions (RQs) and Their Mapping to Tasks, Inputs, Outputs, and Evaluation Metrics. A Single LoRA-adapted Multimodal Model Addresses All Three

Recent advances in multimodal large models (MLMs) offer a more powerful foundation for tackling these challenges. These models integrate visual perception with language reasoning, allowing more accurate identification of subtle evidence such as ripples, turbulence, or occlusion while also supporting transparent decision-making [3]. Representative examples include CLIP [11], which aligns images with natural language supervision for open-vocabulary recognition; BLIP-2 [12], which bridges vision encoders and large language models through lightweight adapters, and LLaVA [13], which enables visual instruction following. These systems have demonstrated strong performance in domains such as medical imaging, industrial inspection, and remote sensing.

For ecological monitoring, multimodal large models (MLMs) are particularly valuable because they can directly associate observable outlet features with textual compliance criteria. By integrating visual indicators of water presence with contextual reasoning, MLMs enable evidence-based assessments of hydropower operations. Emerging studies demonstrate that applying multimodal reasoning to environmental management can enhance transparency, scalability, and trust in automated monitoring pipelines [2, 8].

However, to fully leverage these capabilities, model fine-tuning is essential for achieving optimal performance. In ecological monitoring, assembling large-scale annotated datasets is inherently challenging: water flow conditions vary widely, expert labelling is expensive, and many stations suffer from inconsistent data quality. Parameter-efficient fine-tuning (PEFT) methods, such as LoRA [14, 15], address these limitations by enabling the adaptation of large models using limited labelled data and modest computational resources. This efficiency is crucial for developing scalable and practical monitoring systems that can be deployed in real-world environmental settings.

This paper presents the first application of a vision–language model to ecological flow monitoring. Unlike prior approaches that rely on handcrafted heuristics or narrowly trained classifiers, our system operationalises compliance supervision under real-world constraints of data scarcity, bandwidth limits, and hardware restrictions. Utilising a large multimodal backbone adapted with PEFT, the proposed system is designed to address the three core tasks of ecological flow regulation. In summary, we make the following contributions:

• We introduce a novel multimodal monitoring framework that, for the first time, unifies camera-view validation, single-image flow detection, and reference-based compliance assessment within a single system, improving the comprehensiveness and scalability of ecological flow monitoring.

• We adapt a large multimodal model to the ecological-flow domain via parameter-efficient fine-tuning, demonstrating that such methods can achieve strong generalisation from scarce annotations while covering the wide structural and environmental variability of real-world stations.

• The framework has been deployed at scale across nearly 5,000 hydropower stations, achieving over 90% accuracy while markedly reducing reliance on manual inspections. Importantly, it operates with modest computational resources and lower bandwidth requirements, making large-scale, continuous monitoring both practical and sustainable.

2 LITERATURE REVIEW

2.1 Ecological Flow Monitoring

Ecological flow refers to the minimum water release required to sustain downstream ecosystems and preserve river health. It ensures that aquatic life, sediment transport, and ecological services continue despite the presence of hydropower stations [1, 16, 17]. Failing to maintain ecological flow can reduce biodiversity, degrade habitats, and destabilise river systems [18, 19]. This has led to increasing policy attention, as ecological flow compliance is now mandated in many regions [2, 8]. Monitoring ecological flow in practice remains challenging. Traditional approaches often rely on manual inspections and hydrological models [20, 21]. Manual inspections are costly and cannot cover the large number of small hydropower stations distributed across a region [1]. Hydrological models depend on continuous sensor readings, which are expensive to install and maintain, and are prone to falsified reporting in practice [2, 8]. As a result, regulators frequently face gaps in data quality and timeliness [3].

Recent research has explored data-driven and vision-based approaches. Lightweight CNNs and statistical methods attempt to estimate flow using proxy indicators [5, 9], but they struggle to generalise across diverse outlet types and environmental conditions [3]. Video-based methods such as optical flow estimation or motion recognition networks can detect water movement more reliably [4, 6 ,7, 22], yet they require continuous video input and substantial computational resources [2, 23]. These requirements limit scalability across thousands of stations. The lack of scalable and trustworthy solutions emphasises the need for new methods that can operate from still images and support large-scale regulatory monitoring [3, 8].

2.2 Video and Image-Based Flow Detection

Traditional flow monitoring methods often rely on video analysis. Optical flow models such as FlowNet and RAFT estimate pixel-level motion between frames [6, 7], while action recognition networks such as SlowFast learn temporal patterns across video clips [4]. These approaches work well in controlled environments but require continuous video, stable cameras, and consistent lighting. They also involve complex setup steps, such as defining sensitive zones or analysing motion masks, which increase cost and reduce adaptability. In practice, many small hydropower stations can only upload still frames at intervals due to limited bandwidth, making video-based methods infeasible for large-scale monitoring.

Image-based approaches offer a simpler alternative. Quality metrics such as Laplacian variance or no-reference scores like NIQE can filter blurred or occluded frames [5]. For compliance checks, feature-based similarity methods, including LPIPS [9], CLIP embeddings [11], and DINOv2 [24], compare test images against reference exemplars that represent compliant flow. These techniques allow static monitoring without the need for continuous video. However, they remain sensitive to environmental variation, such as lighting changes or weather effects, and may produce false classifications if used in isolation.

2.3 Multimodal Models and Parameter-Efficient Fine-Tuning

Multimodal large models (MLMs) have become central to vision-language research, combining visual encoders with transformer-based language decoders. Recent work shows that such models can align image and text features effectively and perform a wide range of tasks without task-specific architectures [3]. InternVL and similar systems integrate strong vision encoders with pretrained language backbones, enabling fine-grained visual grounding and general reasoning [2].

Fine-tuning these large models directly is often impractical due to their scale. Parameter-efficient fine-tuning (PEFT) methods address this challenge. LoRA introduced low-rank adapters that significantly reduce training cost while preserving performance [14]. Extensions such as DoRA, PiSSA, and AdaLoRA refine LoRA's efficiency by improving initialisation or dynamically allocating parameters [25-27]. QLoRA combines low-rank adaptation with quantised backbones, further lowering memory requirements [28]. Surveys also show the growing role of PEFT in adapting multimodal models to domain-specific applications [15, 29].

3 BACKGROUND

Field deployment of ecological flow monitoring systems faces several challenges. Outlets differ in structure, ranging from gates and pipes to open channels and weirs, making it difficult to apply uniform rules. Camera installations are often misaligned or obstructed, and poor lighting or image quality further reduces reliability. Moreover, reliable discharge measurements are rarely available, particularly for small hydropower stations distributed across wide regions. These conditions limit the effectiveness of traditional rule-based or sensor-driven methods, which are costly to maintain and prone to errors at scale. To address these issues, we formalise the task as a multi-task classification problem with three components: outlet type recognition, camera installation compliance, and flow detection. This framing provides a foundation for developing AI-based monitoring systems that are both scalable and robust under real-world constraints.

3.1 Installation Compliance Check

Reliable ecological flow monitoring requires that cameras be properly aligned with the target outlet. In real deployments, however, misalignment and degraded image quality are common, introducing significant uncertainty into compliance assessment. Since outlet types and camera positions vary widely, rule-based or template-based AI models do not generalise well. The system flags frames as invalid if the outlet is missing, off-frame, or the image is too blurry or overexposed. Figure 2 (A) illustrates three representative cases: a misaligned view (left), a correctly aligned view (middle), and a blurry frame rejected by the quality filter (right).

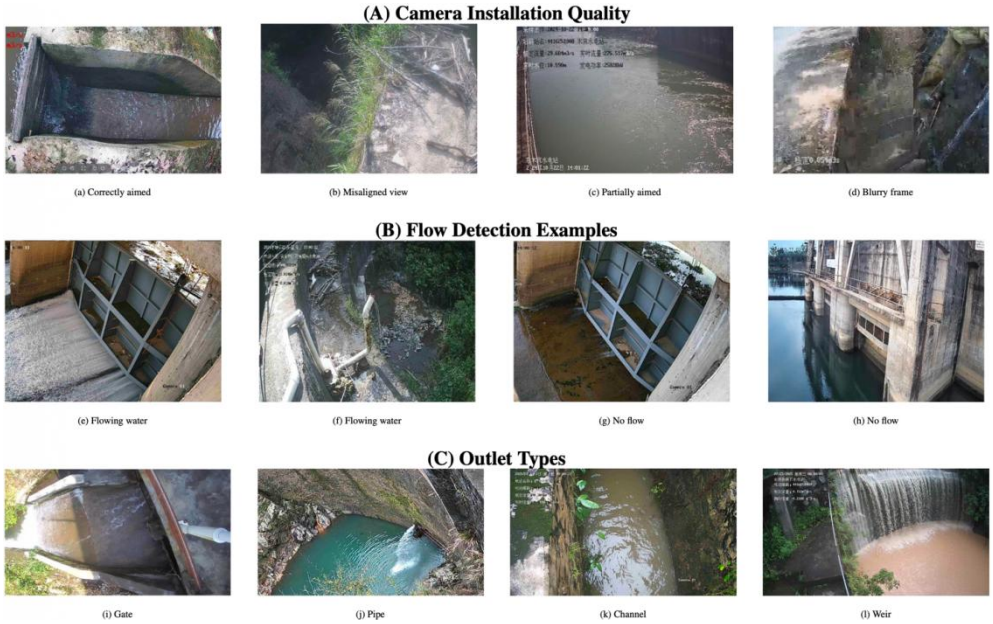

Figure 2 Examples from the Dataset Including Installation Quality (A), Flow Detection (B), and Outlet Types (C)

Camera alignment is especially critical in regulatory monitoring, as misaligned views invalidate downstream flow detection and compliance assessment. In practice, cameras at remote hydropower stations are frequently affected by environmental factors such as lighting changes, fog, lens contamination, or physical displacement due to weather events. These issues exacerbate the diversity introduced by non-standardised installation practices, making manual inspection costly and rule-based approaches impractical. A robust installation compliance check is therefore a prerequisite for reliable monitoring.

3.2 Flow Release Detection

Determining whether water is actively released is central to compliance monitoring, yet it remains one of the most challenging tasks. Simply detecting the presence of water is insufficient, as many outlets contain stagnant pools or discharge directly into rivers where water may be visible but not moving. Traditional video-based methods, such as optical flow or frame differencing, are impractical under bandwidth constraints and fragile in diverse environmental settings. Moreover, lighting, weather, and background conditions can easily obscure or mimic motion cues. These challenges motivate the need for models that can infer active flow robustly from still images by recognising subtle visual patterns indicative of motion, such as ripples, spray, and surface turbulence.

3.3 Outlet Type Classification

Before assessing alignment or flow, the system must first determine the structural type of the discharge outlet. Each hydropower station may use a different outlet design, including gates, pipes, channels, and weirs, while some sites are difficult to categorise due to occlusion or atypical structures. Accurately recognising the outlet type helps downstream modules interpret visual cues correctly—for instance, surface turbulence appears differently in a gate outlet than in a weir. The multimodal model is fine-tuned to classify outlet types into five categories: gate, pipe, channel, weir, and unknown. This classification step ensures that the subsequent tasks of camera alignment and flow detection can adapt to the outlet's geometric characteristics. The model learns structural patterns such as concrete frames, discharge openings, and water guidance elements that distinguish one outlet type from another. Representative examples are shown in Figure 2(C), illustrating the visual diversity across categories.

3.4 Problem Formulation

The ecological flow monitoring task can be formulated as a multi-task classification problem, where the objective is to ensure that each monitoring image provides a valid view of the outlet and contains sufficient visual cues for compliance

assessment. Each site periodically produces a daytime still image $x \in R^{H\times W\times 3}$, where H and W denote the image height and width in pixels, and the final dimension of 3 corresponds to the RGB colour channels:

$$y = \{y_{type}, y_{align}, y_{flow}\} \quad (1)$$

The labels are defined as follows:

• $y_{type} \in \{1, \dots, 5\}$: outlet category (1 = gate, 2 = weir, 3 = pipe, 4 = channel, 5 = unknown),

• $y_{type} \in \{1, \dots, 5\}$: whether the camera is correctly aimed at the outlet,

• $y_{flow} \in \{0, 1\}$: whether active water release is visible.

Absolute discharge estimation is excluded since single frames do not provide sufficient evidence for reliable regression.

3.4.1 Model predictions

The backbone model $f\theta$, with parameters θ, produces predictions for each input image:

$$f_\theta(x) = \hat{y} = (\hat{y}_{type}, \hat{y}_{align}, \hat{y}_{flow}) \quad (2)$$

Here, $\hat{y}$ denotes the model's predicted probability distribution, while the corresponding ground-truth label y is used for training.

3.4.2 Training objective

Given a dataset with N samples, the goal is to learn θ such that the discrepancy between prediction $\hat{y}$ and ground truth y is minimised across all tasks. This is expressed through the combined loss:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\ell_{CE}(y^i_{type}, \hat{y}^i_{type}) + \ell_{BCE}(y^i_{align}, \hat{y}^i_{align}) + \ell_{BCE}(y^i_{flow}, \hat{y}^i_{flow})\right] \quad (3)$$

where:

• l_{CE} measures the difference between the true outlet type and the predicted categorical distribution,

• l_{BCE} measures the difference between the binary ground truth and predicted probabilities for alignment and flow.

This objective aggregates the discrepancies between the predictions and ground truth across outlet classification, alignment detection, and flow detection, enabling the model to jointly improve all tasks.

4 FRAMEWORK

In this section, we describe the adaptation strategy developed to make large multimodal models practical for ecological flow monitoring. Fully fine-tuning models with billions of parameters is impractical given the limited computational resources and the scarcity of domain-specific training data compared to general-purpose benchmarks. To overcome these challenges, we adopt a parameter-efficient fine-tuning approach based on Low-Rank Adaptation.

4.1 PEFT on MLMs

4.1.1 Low-Rank Adaptation (LoRA)

LoRA adapts large transformer models by inserting trainable low-rank matrices into existing linear layers while keeping most pretrained weights frozen [14]. This design captures task-specific knowledge within a compact subspace, substantially reducing redundancy in the parameter space. LoRA achieves performance comparable to full fine-tuning while updating less than 1% of parameters [14]. Its efficiency makes it particularly suitable for ecological flow monitoring, where annotated data are limited and computational resources are constrained. Prior studies have demonstrated LoRA's versatility across NLP, vision, and multimodal tasks such as visual grounding, instruction tuning, and cross-modal alignment [13, 29-32]. Extensions including DoRA, PiSSA, AdaLoRA, and QLoRA further highlight the adaptability of low-rank methods under resource-limited conditions [25-28]. Formally, for a general model with parameter weight matrix $W \in R^{d\times k}$, where d is the input dimension and k is the output dimension, LoRA introduces a low-rank update:

$$W' = W + \Delta W, \ \Delta W = AB \quad (4)$$

where $A \in R^{d\times r}$ and $B \in R^{r\times k}$, with $r \ll \min(d, k)$ defined as the LoRA rank. The pretrained weights W remain frozen, and only (A, B) are optimised with a small r, reducing trainable parameters by more than 95% while preserving multimodal capacity [15, 29].

4.1.2 Multimodal large model (MLM) framework

A multimodal large model (MLM) integrates visual and textual information through a vision encoder and a language model connected by a projection or fusion module. Formally, given an image x_v and an optional textual input x_t, the vision encoder E_v extracts a sequence of visual features:

$$h_v = E_v(x_v) \in R^{n\_v\times d\_v} \quad (5)$$

where n_v is the number of visual tokens and d_v is the embedding dimension. These visual embeddings are mapped into the language embedding space via a projection function $P : \mathbb{R}^{dv} \rightarrow \mathbb{R}^{dl}$:

$$z_v = P(h_v) \quad (6)$$

The language model E_l then consumes the fused input $[z_v; E_{(t)(x_t)}]$, where E_t is the text tokeniser or embedding layer, to perform multimodal reasoning and produce output predictions:

$$\hat{y} = f_\theta(x_v, x_t) = E_l([z_v; E_t(x_t)]) \quad (7)$$

This architecture allows the model to align visual cues with linguistic semantics, enabling downstream tasks such as outlet classification, alignment verification, and flow detection.

4.1.3 Application of LoRA to MLMs

In multimodal architectures, LoRA can be flexibly applied to different components. Typically, language-side adapters are used to enhance reasoning and cross-modal alignment, while the vision encoder remains frozen to preserve general feature representations [13, 29]. In our setup, the vision encoder E_v takes an image input x_v and produces a visual feature embedding:

$$h_v = E_v(x_v) \in R^{d\_v} \tag{8}$$

These embeddings are projected into the language model space and fed into the language encoder-decoder E_l, which takes both the visual features h_v and a text query x_t to generate predictions:

$$\hat{y} = \left(\hat{y}_{type}, \hat{y}_{align}, \hat{y}_{flow}\right) = f_\theta(x_v, x_t) \tag{9}$$

Since fully fine-tuning all parameters θ is infeasible, we apply LoRA only to the linear layers within E_l, keeping E_v fixed. The fine-tuned model f_θ, is then optimised with a combined classification objective:

$$L(\theta') = \frac{1}{N}\sum_{i=1}^{N}\left[\ell_{CE}\left(y^i_{type}, \hat{y}^i_{type}\right) + \ell_{BCE}\left(y^i_{align}, \hat{y}^i_{align}\right) + \ell_{BCE}\left(y^i_{flow}, \hat{y}^i_{flow}\right)\right] \tag{10}$$

where N is the number of training samples, l_{CE} is the cross-entropy loss for multi-class outlet type classification, and l_{BCE} denotes binary cross-entropy losses for camera alignment and flow detection tasks.

4.2 Model Architecture

Building on these techniques, we adopt a modular fine-tuning architecture that extends beyond LLM-only LoRA and enables flexible adaptation across different components of the multimodal pipeline. As illustrated in Figure 3, the design decomposes parameter-efficient adaptation into complementary modules that jointly enhance reasoning, perception, and cross-modal alignment:

1. LLM-side LoRA focuses on adapting the high-level reasoning layers of the language model [14]. This module updates a small subset of projection and attention weights to capture domain-specific semantics in ecological text and metadata, improving logical consistency and interpretability in model outputs.
2. Vision-side LoRA improves the visual backbone to account for domain shifts in outlet imagery such as varying lighting, vegetation cover, or seasonal flow appearance [13]. By introducing low-rank adapters into vision transformers or convolutional blocks, the model learns to emphasise hydrologically relevant features (e.g., spillway patterns, water reflectance) without losing pretrained general visual capacity.
3. Cross-modal adaptation employs lightweight intermediate alignment modules such as IA^3 or low-rank bridges to improve the interaction between visual and textual representations [33]. These modules operate in the fusion space, aligning projected vision embeddings z_v with language embeddings $E_t(x_t)$ to ensure coherent grounding between visual cues and regulatory descriptions (e.g., “outlet blocked”, “flow present”).
4. QLoRA backbones enable low-bit quantised fine-tuning of large language components [28], reducing GPU memory requirements while maintaining accuracy. This makes large multimodal models trainable on modest hardware, which is essential for continuous, on-site ecological monitoring systems.

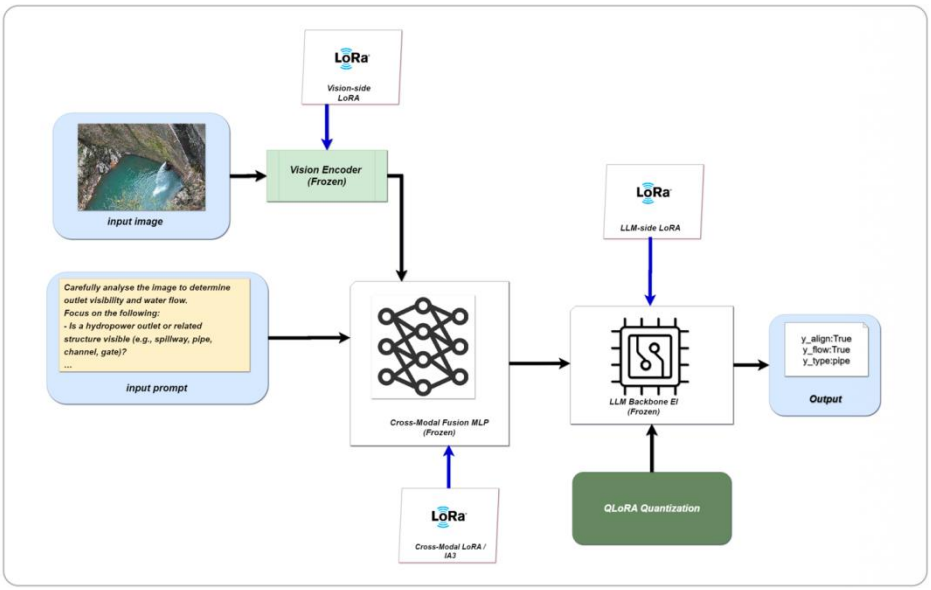

Figure 3 Modular Parameter-Efficient Fine-Tuning Architecture Integrating LoRA and QLoRA Across Vision, Fusion, and Language Components. Vision-Side LoRA Refines Perception; Cross-Modal LoRA or IA^3 Improves Multimodal Alignment; LLM-Side LoRA Adapts Reasoning; and QLoRA Backbones Reduce Memory Overhead for Scalable Deployment

This modular design provides a unified framework for hierarchical adaptation across modalities: vision-side modules capture domain-specific perception, language-side modules refine reasoning, and cross-modal bridges ensure semantic

alignment. Overall, these modules deliver robust and efficient adaptation to real-world monitoring conditions while keeping fine-tuning computationally lightweight. Such flexibility is particularly valuable for infrastructure and environmental compliance applications, where data heterogeneity and operational constraints are common [29]. Table 1 summarises the components updated during LoRA-based fine-tuning. This configuration reflects a balance between computational efficiency and methodological stability. Fully fine-tuning a large multimodal model would exceed available hardware resources and risk overfitting, given the limited size of ecological datasets. Alternative PEFT strategies, such as prompt tuning or adapter layers, often yield weaker performance on multimodal reasoning tasks, while applying LoRA to vision backbones generally requires large-scale visual data to avoid degrading pretrained representations [15, 29, 33]. By constraining LoRA updates to the language-side linear layers, the model preserves pretrained multimodal knowledge while achieving efficient adaptation to the ecological flow monitoring domain.

Table 1 Configuration of Trainable and Frozen Components During LoRA-Based Fine-Tuning

Module	Component Type	Adaptation Method	Trainable Status
Vision Encoder (*Ev*)	Visual Backbone	None (Frozen)	Frozen
Vision Adapter	Vision Enhancement Block	LoRA ($r = 8$)	Trainable
Cross-Modal MLP	Fusion / Alignment Layer	IA3 (scaling adaptation)	Trainable
Language Model (*El*)	Transformer Decoder	LoRA ($r = 16$)	Trainable
Quantised Backbone	LLM Base Parameters (4-bit)	QLoRA quantisation	Frozen during training
Tokeniser & Embedding	Input Layer	-	Frozen

4.3 System Overview

The proposed ecological flow monitoring system adopts a modular, four-layer architecture that bridges perception, data management, and business-level decision support, as illustrated in Figure 4. Each layer encapsulates distinct functionality while maintaining seamless data flow from field sensors to regulatory dashboards.

4.3.1 Perception layer

This layer collects real-time visual data from multiple camera installations positioned at key ecological outlets such as channels, gates, weirs, and pipes. A fine-tuned InternVL-26B multimodal model serves as the perceptual core, performing automatic camera calibration, outlet classification, and flow-state detection. By combining visual understanding with multimodal reasoning, the system can assess compliance with ecological discharge requirements in real time.

4.3.2 Data storage and visualisation layer

Processed visual data and metadata are transmitted to a centralised supervision platform that interfaces with the Guangdong Provincial Water Resources Video Platform. This layer provides secure data archiving, retrieval, and dashboard-based visualisation for both regulators and hydropower operators, ensuring transparency and traceability of monitoring results.

4.3.3 Data parsing layer

This intermediate layer bridges perception and business logic by handling data aggregation, frame extraction, video decoding, and analytical model updates. It forms the operational backbone of the Multimodal Image Analysis System, supporting continuous fine-tuning and result reporting for ecological monitoring workflows.

4.3.4 Business logic layer

At the top level, this layer encodes the regulatory and operational logic governing ecological flow compliance. It manages alert generation, system monitoring, reporting protocols, and user interfaces for data presentation and remote maintenance. Through this layer, the system translates model outputs into actionable insights aligned with policy and compliance frameworks.

The system processes both visible-light frames and laser-derived data, guaranteeing reliability under diverse conditions. Running on hardware with limited resources (e.g., only a few GPUs), it supports both batch and real-time inference, and this layered design provides an automated pipeline from raw data ingestion to compliance visualisation, allowing scalable and transparent ecological flow monitoring.

Figure 4 Simplified System Workflow from Business Logic to Perception

5 EXPERIMENT

5.1 Dataset Collection

The dataset used in this study was collected from 4,689 small hydropower stations distributed across the province. Each station provided a set of images captured from a fixed monitoring point between 2024 and 2025, primarily under natural daytime conditions (08:00–18:00). Two images were initially selected from each station as annotation candidates, yielding

9,378 filtered samples. From these, we manually labelled 1,000 clear and representative images for model training and evaluation. To ensure high-quality supervision, all images were annotated by domain experts from the provincial hydropower authority. Disagreements were resolved through majority voting.

To avoid information leakage, the dataset was split strictly by station-level rather than by image-level. Images collected from the same hydropower station never appeared in both the training and validation sets, ensuring that model evaluation reflected true cross-site generalisation rather than memorisation of fixed camera viewpoints or outlet layouts. To account for the substantial structural variability across stations, reflecting differences in outlet geometry, construction materials, flow-release mechanisms, and installation practices, the dataset was categorised into five outlet types: gates, pipes, open channels, weirs, and unknown. This categorisation was necessary because flow appearance, turbulence patterns, and camera alignment cues differ across outlet designs, and treating them as a single homogeneous class would obscure meaningful visual heterogeneity.

For each outlet type, we included 200 positive and 50 negative samples. This balanced design ensured that the model learned both compliant and non-compliant scenarios for each outlet structure, which is essential for reliable downstream compliance assessment. In total, the training set contained 1,000 labelled images, and an additional 250 samples from unseen stations were held out as the test set. All images were pre-processed to a fixed resolution of 448 × 448 pixels to normalise input scale across diverse camera models and installation distances. Each sample was annotated for three supervised learning tasks:

1. Installation compliance – evaluating whether the camera position and angle meet installation standards;
2. Flow release detection – determining whether water flow is present;
3. Flow compliance assessment – judging flow conformity based on visual similarity to reference conditions.

Each task included approximately 3,000 labelled samples, and Table 2 summarises the key dataset statistics. The dataset was stored in a structured format, where each entry corresponded to a single image and its associated prompt–response pair. Each image was paired with a carefully designed conversational prompt to facilitate multimodal instruction tuning. All images were standardised and processed, with thumbnail support for efficient data loading. To ensure consistency and reliability, all files were encoded in the same format, and duplicate samples across dataset splits were systematically removed.

An overview of the data preparation and annotation workflow is shown in Figure 5. The pipeline begins with raw images collected from hydropower stations, which are then paired with task-specific textual prompts. These image–prompt pairs were used to train and evaluate a fine-tuned vision–language model across the three defined tasks. As shown in the figure, the model received both the image and prompt as input and produced structured question–answer outputs, including outlet type identification, camera installation assessment, and flow detection.

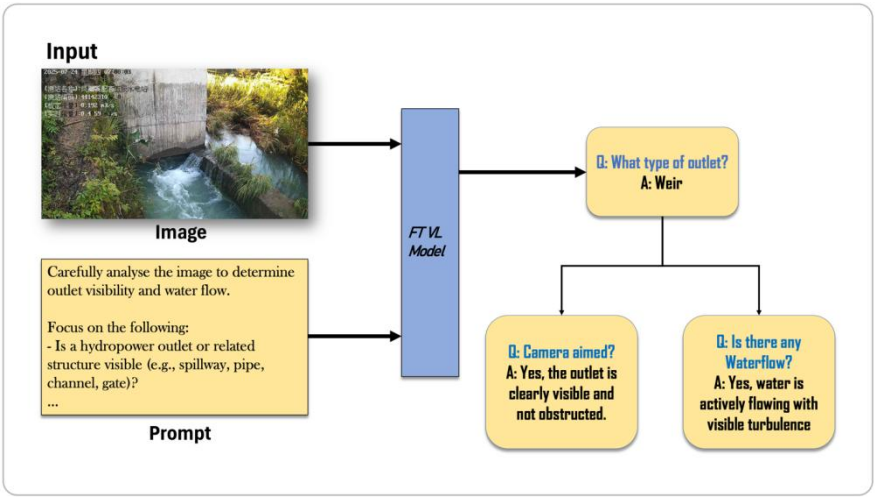

Figure 5 Dataset Pipeline for Ecological Flow Monitoring. Images are Collected from Hydropower Stations, Categorised into Installation, Flow, Annotated with (Prompt, Response) Pairs, and Pre-Processed (Resized to 448 × 448, Thumbnail Generation, and Duplicate Removal) Before Being Used for Fine-Tuning

Table 2 Dataset Statistics

Statistic	Value
Total number of raw images	46,890
Number of hydropower stations	4,689
Filtered candidate set	9,378 images
Labelled training images	1,000 (750 train / 250 test)
Outlet types	Gate, Pipe, Channel, Weir, Unknown
Samples per outlet type	200 positive + 50 negative
Image resolution (after preprocessing)	448 × 448 pixels

Statistic	Value
Capture period	2024 – 2025
Capture time (daily)	08:00 – 18:00
Lighting conditions	Daytime
Seasonal coverage	Spring, Summer, Autumn
Labelled tasks per image	Installation, Flow, Type

5.2 Experimental Setup

Monitoring ecological flow in small hydropower stations involves two main challenges: complex outlet structures and limited annotated data. To address these, we adopt a parameter-efficient multimodal learning strategy, leveraging a 26B-parameter backbone pre-trained on large-scale vision–language data while enabling domain-specific adaptation with minimal compute overhead.

Training is conducted on four NVIDIA A6000 GPUs (48 GB each) with BF16 precision and ZeRO Stage 3 optimiser sharding via DeepSpeed. The effective batch size is 16 (two samples per GPU with gradient accumulation over four steps). Models are trained for 30 epochs with cosine learning-rate decay (initial rate 2×10^{-5}) and a 3% warm-up ratio. All images are standardised to 448 × 448 pixels, with dynamic resizing to improve robustness across varying outlet geometries.

5.3 Benchmark Models

We use InternVL2.5-26B as the foundation model for its strong visual–language alignment and cross-domain generalisation [33]. Parameter-efficient LoRA adapters are inserted into both the vision encoder and cross-modal fusion layers, enabling domain specialisation without modifying the full backbone. To provide comprehensive comparisons, we evaluate classical and deep learning baselines under identical pre-processing and data splits:

5.3.1 Classical models

Trained on 128 × 128 image embeddings:

• Logistic Regression: tuned over L2 regularisation.

• Support Vector Machine (RBF): tuned over kernel bandwidth and penalty coefficient.

• XGBoost: optimised over tree depth, number of estimators, and learning rate.

5.3.2 Deep learning models

• DNN: three fully connected layers with ReLU activations, trained on 128 × 128 inputs using Adam (2×10^{-4}, cosine decay, batch size 16) for 30 epochs.

• CNN: three convolutional layers followed by two fully connected layers with ReLU and max-pooling, trained on 224 × 224 images under the same optimiser and schedule.

5.3.3 Multimodal model

• InternVL2.5-26B (LoRA fine-tuned): adapted to the ecological-flow domain via parameter-efficient LoRA tuning.

5.4 Evaluation Protocol

All models are evaluated on a validation set comprising 250 images, with 50 samples for each outlet category. They follow identical pre-processing and normalisation procedures to ensure consistency and fairness across experiments. For equitable evaluation, both the CNN and DNN are trained for 30 epochs using the Adam optimiser with cosine learning rate decay, an initial learning rate of 2×10^{-4}, and a batch size of 16. The classical models are trained on image embeddings extracted from the same dataset, using identical training and validation splits. This setup ensures a balanced representation across outlet types and consistent experimental conditions, as described in Section 5.1.

6 RESULTS & DISCUSSION

This section presents a comparative evaluation across classical, deep, and multimodal architectures. The results demonstrate a consistent performance progression from traditional methods to neural models, and finally to parameter-efficient multimodal adaptation.

6.1 Quantitative Results

Table 3 summarises model performance across alignment, flow detection, and outlet-type prediction for benchmark models. Notably, classical models achieve moderate accuracy (68–76%), which is likely limited by their reliance on fixed global image embeddings. Among them, XGBoost benefits from nonlinear interactions, consistently outperforming linear models.

Table 3 Performance Comparison of Benchmark Models on Alignment, Flow, and Outlet-Type Recognition (95% Confidence Intervals; 250 Validation Samples)

Model	Align Accuracy (%)	Align Precision (%)	Flow Accuracy (%)	Flow Precision (%)	Type Accuracy (%)	Type Precision (%)
Logistic Regression	68.7 ± 2.5	68.2 ± 2.7	69.4 ± 2.4	68.9 ± 2.6	42.3 ± 3.4	41.5 ± 3.5
SVM (RBF Kernel)	72.0 ± 2.3	71.5 ± 2.5	73.1 ± 2.2	72.4 ± 2.4	55.6 ± 3.0	54.9 ± 3.1
XGBoost	76.4 ± 2.0	75.8 ± 2.1	77.5 ± 1.9	76.9 ± 2.0	61.8 ± 2.8	60.9 ± 2.9
DNN (3 FC)	80.9 ± 1.8	80.2 ± 1.9	82.1 ± 1.7	81.4 ± 1.8	70.2 ± 2.5	69.4 ± 2.6
CNN (3 Conv + 2 FC)	86.5 ± 1.4	85.9 ± 1.5	87.8 ± 1.3	87.1 ± 1.4	71.4 ± 2.3	70.8 ± 2.4
Intern-VL26B (before fine-tuning)	68.4 ± 2.6	68.0 ± 2.6	69.0 ± 2.5	69.0 ± 2.6	68.8 ± 2.6	68.4 ± 2.6
Intern-VL26B (after fine-tuning)	**92.0 ± 3.0**	**91.8 ± 3.1**	**92.4 ± 3.0**	**92.8 ± 3.0**	**90.3 ± 3.0**	**89.8 ± 3.2**

Deep learning baselines demonstrate stronger spatial reasoning capabilities: the DNN surpasses 80% accuracy, while the CNN reaches 86–88%, showing the advantage of hierarchical feature extraction for complex outlet structures.
The zero-shot InternVL2.5-26B performs comparably to CNNs on alignment and flow detection (68–69%), indicating that pre-trained visual–language priors alone are insufficient for domain-specific ecological-flow tasks. After LoRA-based fine-tuning, the model exhibits a substantial improvement, achieving over 92% accuracy and precision across the evaluated tasks. Confidence intervals are narrower post-tuning, reflecting more stable convergence and enhanced model calibration.

6.2 Per-Category Analysis

Table 4 presents detailed validation results for each category before and after LoRA-based fine-tuning. Overall, the fine-tuned InternVL2.5-26B model achieves substantial improvements across all metrics, with an average increase of approximately 23–24 percentage points in both accuracy and precision. This demonstrates the effectiveness of parameter-efficient adaptation in leveraging limited domain-specific data while preserving the pre-trained knowledge of the large backbone.

Table 4 Comparison of Validation Performance (%) Before and After LoRA-Based Fine-Tuning (95% Confidence Intervals; 250 Samples, 50 Per Category)

Category	Align Acc.		Flow Acc.		Align Prec.		Flow Prec.		Type Acc.		Type Prec.	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Gate	68.5± 2.6	92.3± 3.1	69.2± 2.5	93.0± 3.0	68.1± 2.7	91.7± 3.2	69.0± 2.6	93.2± 3.1	69.0± 2.6	92.4± 3.0	68.5± 2.7	92.0± 3.0
Pipe	67.8± 2.7	90.1± 3.3	68.4± 2.6	92.0± 3.1	67.0± 2.8	90.8± 3.3	68.1± 2.7	92.1± 3.2	68.2± 2.7	89.8± 3.3	67.7± 2.8	88.4± 3.4
Channel	69.4± 2.5	94.1± 2.9	70.1± 2.4	93.4± 3.0	69.0± 2.5	93.2± 3.0	70.3± 2.4	94.3± 2.9	70.0± 2.5	90.7± 3.0	69.5± 2.5	90.3± 3.1
Weir	68.8± 2.6	91.2± 3.2	69.3± 2.5	92.4± 3.1	68.5± 2.6	91.8± 3.2	69.2± 2.5	93.0± 3.1	69.0± 2.6	90.1± 3.2	68.7± 2.6	90.0± 3.2
Unknown	67.2± 2.8	90.3± 3.3	67.8± 2.7	91.1± 3.2	66.9± 2.8	90.5± 3.3	67.6± 2.7	91.4± 3.2	68.0± 2.8	88.6± 3.3	67.4± 2.8	88.2± 3.4
Overall	**68.4± 2.6**	**92.0± 3.0**	**69.0± 2.5**	**92.4± 3.0**	**68.0± 2.6**	**91.8± 3.1**	**69.0± 2.6**	**92.8± 3.0**	**68.8± 2.6**	**90.3± 3.0**	**68.4± 2.6**	**89.8± 3.2**

A closer inspection of the per-category performance reveals several notable patterns:
• Pipe and Unknown categories, which exhibit higher structural variability and ambiguous visual cues, benefit the most from LoRA tuning. The alignment accuracy for Pipe improves from 67.8% to 90.1%, and for Unknown from 67.2% to 90.3%, highlighting the model's enhanced ability to disambiguate complex structures after domain adaptation.
• Channel shows the highest absolute post-tuning performance (94.1% alignment accuracy), suggesting that the pre-trained visual–language priors are particularly effective for geometrically regular and well-represented categories.
• Weir and Gate categories also demonstrate notable improvements, with reduced confidence intervals indicating more stable predictions and better model calibration.
The reduction in confidence interval widths across all categories reflects increased prediction stability and robustness after fine-tuning. Collectively, these results indicate that LoRA-based adaptation not only improves overall accuracy and precision but also enhances model reliability across structurally diverse outlet types, making the system suitable for practical ecological-flow monitoring scenarios.

6.3 Ablation Study

To examine how decoding hyperparameters influence inference stability, we perform an ablation study on the temperature value used during generation. Temperature controls sampling randomness, where lower values produce more deterministic predictions and higher values introduce greater variability. Since ecological-flow assessment is a deterministic visual classification task, excessively high temperatures may degrade reliability.

We evaluate temperature settings of {0.1, 0.3, 0.5, 0.7, 1.0} while fixing all other decoding parameters to their default values (top-k = 40, top-p = 0.95, frequency/presence penalties = 0). Experiments are conducted on the alignment detection task within the Channel-type outlet subset, a challenging scenario characterised by surface reflections and turbulent patterns. Results are shown in Figure 6.

As illustrated in Figure 6, accuracy peaks at T = 0.3. Lower temperatures (e.g., 0.1) yield overly conservative logits and occasional misclassifications under visually ambiguous conditions, whereas higher temperatures introduce unnecessary stochasticity that reduces prediction consistency. The same trend is observed across other outlet types, with a variation of less than 1%.

Other decoding parameters (top-k, top-p, and penalty coefficients) have negligible influence (less than 1% difference across all metrics), as they primarily affect linguistic diversity rather than visual recognition. Thus, we fix these parameters in all experiments and use temperature as the sole varying factor.

Overall, the ablation confirms that the fine-tuned multimodal model remains robust across a wide range of decoding settings. A moderate temperature, such as T = 0.3, offers the best balance between determinism and stability, and is adopted in all subsequent evaluations.

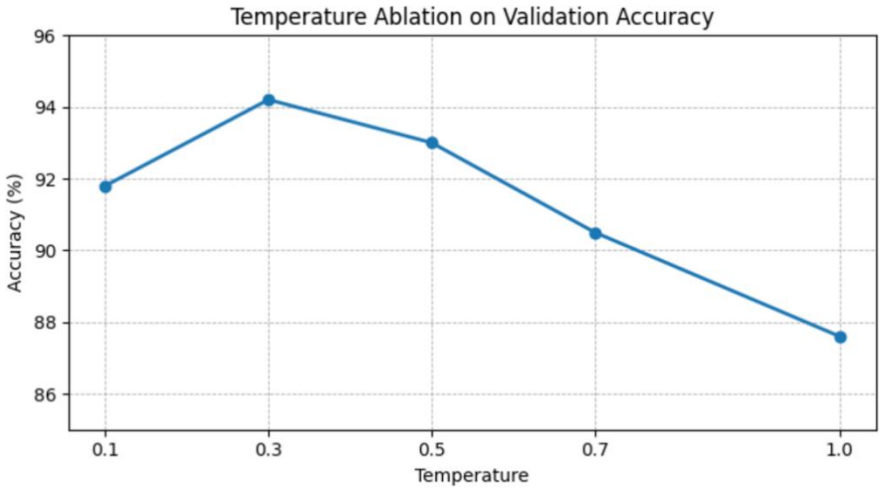

Figure 6 Effect of Inference Temperature on Alignment Accuracy

6.4 Operational Considerations

Following deployment across approximately 5,000 hydropower stations, the fine-tuned model demonstrates reliable single-frame evaluation of installation alignment and ecological-flow conformity. Unlike video-based methods, the proposed framework operates with minimal computational and bandwidth requirements, offering a practical and scalable solution for real-world hydropower monitoring.

Overall, LoRA fine-tuning yields a gain of over 20% compared to both CNN and zero-shot baselines, confirming its effectiveness in domain-specific multimodal learning. Following large-scale deployment across approximately 5,000 hydropower stations, the fine-tuned model autonomously inspects installation compliance and flow conformity, reliably identifying non-compliant cases and anomalies. Unlike video-based or transformer-only systems, the proposed single-frame inference framework operates with minimal compute and bandwidth requirements, making it highly suitable for continuous, real-world ecological monitoring at scale.

7 CONCLUSION

We introduced a scalable AI system for monitoring ecological flow releases in small hydropower stations. The proposed framework adapts the InternVL-26B vision–language model through LoRA-based parameter-efficient fine-tuning, enabling effective learning from a limited dataset of 1,000 labelled samples. The resulting model achieves over 90% accuracy and precision across all outlet categories, meeting practical regulatory requirements while maintaining statistical reliability. Compared with traditional CNN-based or video-dependent approaches, the system operates efficiently on single-frame imagery and requires substantially less computation, supporting large-scale deployment across diverse hydropower facilities. Deployment across nearly 5,000 stations further demonstrates the feasibility of multimodal models in real-world environmental monitoring and regulatory workflows. Despite these strengths, several limitations remain. The current dataset displays seasonal imbalance and is dominated by daytime conditions; model performance under low-light, nocturnal, or extreme weather scenarios is not yet fully evaluated. The system infers compliance based on visual

patterns rather than quantitative discharge estimation, which, although sufficient for enforcement, limits hydrological interpretability. In addition, reliability may degrade under persistent hardware faults or suboptimal camera placement, highlighting the need for improved fault-tolerance and quality-control mechanisms.

Future work will focus on expanding the dataset to include greater seasonal, geographic, and environmental diversity, as well as strengthening robustness under adverse conditions. Lightweight deployment techniques—such as quantised LoRA, structured pruning, and knowledge distillation—will be explored to facilitate efficient on-device inference. Integrating temporal cues from multi-frame sequences or incorporating complementary sensing modalities (e.g., infrared imagery or in situ flow sensors) represents another promising direction. From a policy perspective, the findings illustrate the potential of large multimodal models to support transparent, scalable, and automated ecological compliance monitoring, offering a foundation for next-generation environmental governance systems.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Yu Z, Zhang J, Zhao J, et al. A new method for calculating the downstream ecological flow of diversion-type small hydropower stations. Ecological Indicators, 2021, 125: 107530.

[2] Liu W, Chen J, Wang H, et al. Perspectives on Advancing Multimodal Learning in Environmental Science and Engineering Studies. Environmental Science & Technology, 2024. DOI: 10.1021/acs.est.4c03088.

[3] Yin S, Fu C, Zhao S, et al. A Survey on Multimodal Large Language Models. National Science Review, 2024, 11(12): nwae403. DOI: 10.1093/nsr/nwae403.

[4] Feichtenhofer C, Fan H, Malik J, et al. SlowFast Networks for Video Recognition. arXiv preprint, 2019. DOI: 10.48550/arXiv.1812.03982.

[5] Mittal A, Soundararajan R, Bovik AC. Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters, 2012, 20(3): 209-212.

[6] Dosovitskiy A, Fischer P, Ilg E, et al. FlowNet: learning optical flow with convolutional networks. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015: 2758-2766.

[7] Teed Z, Deng J. RAFT: recurrent all-pairs field transforms for optical flow. Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, 2020, 12347. DOI: 10.1007/978-3-030-58536-5_24.

[8] Elgendy H, Sharshar A, Aboeitta A, et al. GeoLLaVA: Efficient Fine-Tuned Vision-Language Models for Temporal Change Detection in Remote Sensing. arXiv preprint, 2024. DOI: 10.48550/arXiv.2410.19552.

[9] Zhang R, Isola P, Efros AA, et al. The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 586-595. DOI: 10.1109/CVPR.2018.00068.

[10] GPGGWR Commission. Small hydropower stations in Guangdong Province. GPGGWR Commission, 2018.

[11] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. arXiv preprint, 2021. DOI: 10.48550/arXiv.2103.00020.

[12] Li J, Li D, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. ICML'23: Proceedings of the 40th International Conference on Machine Learning, 2023, 814: 19730-19742.

[13] Liu H, Li C, Wu Q, et al. Visual instruction tuning. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023, 1516: 34892-34916.

[14] Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv preprint, 2021. DOI: 10.48550/arXiv.2106.09685.

[15] Han Z, Gao C, Liu J, et al. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. arXiv preprint, 2024. DOI: 10.48550/arXiv.2403.14608.

[16] Gao L, Wang J, Zhou M. Ecological flow management and hydropower trade-offs: A review. Renewable and Sustainable Energy Reviews, 2023, 176: 113162.

[17] Shen D, Chen L, Zhao W. Hydropower development and ecological flow requirements in China. Ecological Indicators, 2020, 117: 106603.

[18] Li Q, Huang M, Zhang L. Ecological flow policies and river biodiversity protection in China. Water Policy, 2019, 21(4): 791-805.

[19] Chen Y, Wu G, Liu Y. River health assessment and ecological flow requirements: Progress and prospects. Journal of Hydrology, 2021, 599: 126381.

[20] Xu P, Zhang L. Hydrological monitoring methods and challenges in ecological flow management. Hydrological Sciences Journal, 2021, 66(5): 747-758.

[21] Zhou W, Fang Z, Liu C. Monitoring ecological flow in small rivers: Approaches and limitations. Ecological Indicators, 2020, 110: 105948.

[22] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015: 4489-4497. DOI: 10.1109/ICCV.2015.510.

[23] Chen Z, Yang H, Liu W. Video-based monitoring of river flow using deep learning. Water, 2021, 13(21): 3075.
[24] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint, 2023. DOI: 10.48550/arXiv.2304.07193.
[25] Liu S Y, Wang C Y, Yin H, et al. DoRA: Weight-Decomposed Low-Rank Adaptation. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR, 2024, 235: 32100-32121.
[26] Meng F, Wang Z, Zhang M. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems, 2024, 3846: 121038-121072.
[27] Zhang Q, Chen M, Bukharin A, et al. AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. International Conference on Learning Representations (ICLR), 2023.
[28] Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: efficient finetuning of quantized LLMs. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023, 441: 10088-10115.
[29] Zhang R, Gui L, Sun Z, et al. Direct Preference Optimization of Video Large Multimodal Models from Language Model Reward. arXiv preprint, 2024. DOI: 10.48550/arXiv.2404.01258.
[30] Li K, He Y, Wang Y, et al. VideoChat: Chat-Centric Video Understanding. arXiv preprint, 2023. DOI: 10.48550/arXiv.2305.06355.
[31] Chen Z, Qi Z, Cso X, et al. Class-level Structural Relation Modelling and Smoothing for Visual Representation Learning. arXiv preprint, 2023. DOI: 10.48550/arXiv.2308.04142.
[32] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP. arXiv preprint, 2019. DOI: 10.48550/arXiv.1902.00751.
[33] Chen Z, Wang W, Cao Y, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint, 2024. DOI: 10.48550/arXiv.2412.05271.